

**Causal connectives and perspective markers in Chinese:
The encoding and processing of subjectivity in discourse**

Published by
LOT
Trans 10
3512 JK Utrecht
The Netherlands

phone: +31 30 253 6111

e-mail: lot@uu.nl
<http://www.lotschool.nl>

Cover illustration: A bridge that connects Netherlands and China.
Photographed by Yibin Gu; edited by Xinyao Jin.

ISBN: 978-94-6093-266-3
NUR 616

Copyright © 2018: Yipu Wei. All rights reserved.

**Causal connectives and perspective markers in Chinese:
The encoding and processing of subjectivity in discourse**

Causale connectieven en perspectiefmarkeerders in het Chinees:
corpus- en verwerkingsstudies naar subjectiviteit in discourse
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen
op woensdag 15 januari 2018
des ochtends te 10.30 uur

door

Yipu Wei

geboren op 13 oktober 1988 te Henan, China

Promotor: Prof.dr. T.J.M. Sanders
Copromotoren: Dr. J. Evers-Vermeul
Dr. W.M. Mak

This thesis was accomplished with financial support from China
Scholarship Council (grant number: 201307720042).

Table of Contents

Chapter 1. Introduction	1
Chapter 2. Collocational analyses at the discourse level using Chinese corpora	15
Chapter 3. The use of perspective markers and connectives in expressing subjectivity: Evidence from collocational analyses	37
Chapter 4. Causal connectives as indicators of source information: Evidence from the visual world paradigm	63
Chapter 5. The role of connectives and perspective markers in the on-line processing of subjective relations	95
Chapter 6. General discussion and conclusion	117
References	129
Appendices	141
Samenvatting in het Nederlands	183
Acknowledgements	195
Curriculum Vitae	199

Chapter 1. Introduction

1.1 Defining subjectivity

People communicate via linguistic elements that are connected in discourse. They have the role of reporter or narrator for all utterances they produce to describe what is happening in the real world or to express what language users think and feel about certain situations. However, the way in which they are present may differ: in example (1), the speaker reports a description of an event, while in the sentences in (2) and (3), the speaker expresses personal opinions or judgements with the linguistic elements *may* and *perhaps*. The speaker can explicitly introduce herself by an expression like *I think*, as illustrated in example (4). In this case, the speaker *I* can be held responsible for the statement of ‘Mary is angry at John’. Hence, in examples (2), (3) and (4), the sentence is expressed from someone’s perspective, which introduces subjectivity. This subjectivity is absent in (1), which is therefore seen as an objective utterance: the speaker is only present as an objective reporter.

- (1) Mary is angry at John.
- (2) Mary may be angry at John.
- (3) Perhaps Mary is angry at John.
- (4) I think Mary is angry at John.

Subjectivity is an encompassing linguistic phenomenon (Finegan, 1995) as well as an important cognitive notion in discourse representations. It is considered to be a gradient phenomenon which concerns ‘the degree of the grounding in the perspective of the speaker’ (Traugott, 1995: 34). An important distinction between the conceptual representations of subjective and objective sentences is whether an intentional mind is involved or not. According to Finegan (1995: 1): ‘subjectivity concerns the involvement of a locutionary agent in a discourse, and the effect of that involvement on the formal shape of discourse – in other words, on the linguistic expression of self’ (cf. Lyons, 1977). In all instances, the speaker is involved as the reporter or narrator of the discourse, but in subjective utterances, the speaker is also involved as the locutionary agent who is responsible for the evaluation, judgement, or reasoning in the sentence. The locutionary agent can be either the speaker of the utterance or a character in the discourse (Sanders, Sanders, & Sweetser, 2009). In examples (2), (3) and (4), the speaker is the locutionary agent who is making the judgement that *Mary is angry at John*. In example (5), the locutionary agent is another character – *Peter*. In (2)-(4) and (5), a high degree of subjectivity is introduced because the thoughts arise from someone’s mind and are not a mere description of reality.

(5) Peter thinks Mary is angry at John.

At the linguistic level, subjectivity is encoded by perspective markers such as *may*, *perhaps*, *I think* and *Peter thinks*, or it can be derived from the propositional content of the sentence that a claim or judgement is involved instead of a description of facts. No such perspective markers are present in example (1), and the author/speaker is just reporting the events he/she observes. In examples (2) to (5), however, the linguistic cues that encode subjectivity introduce the perspective of the narrator or some other character in the discourse. For instance, linguistic cues such as *I think* in (4) and *Peter thinks* in (5) explicitly attribute the claim in the utterance to a source of information; markers such as *may* and *perhaps* in (2) and (3) indicate the presence of a source of information without explicitly mentioning this source. These and other observations in language use lead to the question of what exactly the relation between perspective and subjectivity is.

At the conceptual level, readers/hearers need to represent subjectivity in their mental representations of a discourse, guided by the linguistic cues. Subjectivity, by definition, involves personal opinions/feelings that arise from someone's perspective. Therefore, in the representation of subjectivity, the source to whom the information is attributed is important. In other words, in the representation of subjectivity, perspective taking is important. In the literature, perspective and subjectivity have been treated as one notion (Finegan, 1995), or analyzed separately, focusing on one of the two (Graesser, Olde, & Klettke, 2002; Nuyts, 2001). An exception is found in the work of Sanders & Redeker (1996: 293), who define perspective as "the introduction of a subjective point of view that restricts the validity of the presented information to a particular subject (person) in the discourse."

In the current thesis, the notion of perspective is used as a tool to operationalize subjectivity in discourse. While subjectivity indicates *that* someone's opinion or attitude about the information is involved, perspective concerns *who* that 'someone' is, for instance the speaker/writer or a character in the discourse, and *how* he/she is involved, for example expressing either an attitude or emotion, or his/her degree of certainty about the utterance. In other words, perspective markers provide a window on the mental representation of subjectivity.

1.2 Subjectivity in causal relations

Subjectivity phenomena can be observed in utterances, as well as in relations between utterances, and in discourse as a whole. In the case of causal relations, a subjective relation concerns the reasoning attributed to a person's mind, in contrast to an objective relation which expresses the causal chains between concrete facts in the real

world (Sanders et al., 2009; Sanders, Sanders, & Sweetser, 2012). A causal relation is objective if the speaker reports that events are connected in the real world, as in example (6). A relation is subjective if the causal connection between two segments is constructed in the speaker's mind instead of in the real world, as in example (7). In the latter case, the speaker is not merely reporting the causal relation “out there in the world”, but constructs the causality based on his/her own subjective inference (Kamalski, Sanders, & Lentz, 2008; Stukker & Sanders, 2012).

- (6) A strong storm stroke the city last night, so the public traffic system was temporarily paralyzed.
- (7) The public traffic system was temporarily paralyzed after the storm, so the storm must have been very strong.

In (6) and (7), the connective *so* is used to explicate that the relation between the connected segments is causal. The fact that this connective can be used in both the objective relation in (6) and the subjective relation in (7) indicates that *so* is underspecified in terms of subjectivity – the connective itself does not encode whether the relation is subjective or objective. Readers need to use the propositional content of the connected clauses to derive that the relation between these clauses is either objective or subjective. For example, *must* in (7) indicates that the second clause is a claim rather than a description of facts, and readers can use this information to come to the interpretation that they are dealing with an argument-claim relation and not with a cause-consequence relation.

1.3 The linguistic encoding of subjectivity

Researchers have studied connectives based on individual examples and intuitions, in order to find out how lexicons of connectives are organized to express relations: which relations can be expressed by which connectives in various languages (Sweetser, 1990). Following the theoretical proposals that came out of this, linguists over the last decades have tested these theories in actual corpora of language use, applying quantitative techniques and methodologies. Recently, automatic analyses such as collocational analysis have obtained attention as tools to explore the semantic and pragmatic profiles of connectives. For several languages, the corpus-based findings have revealed the categorization of connectives in terms of the degree of subjectivity they encode: as illustrated in Figure 1, some connectives are specifically used for subjective relations, some are mainly used to express objective relations, and others are neutral in subjectivity and can be used generically for both subjective and objective relations.

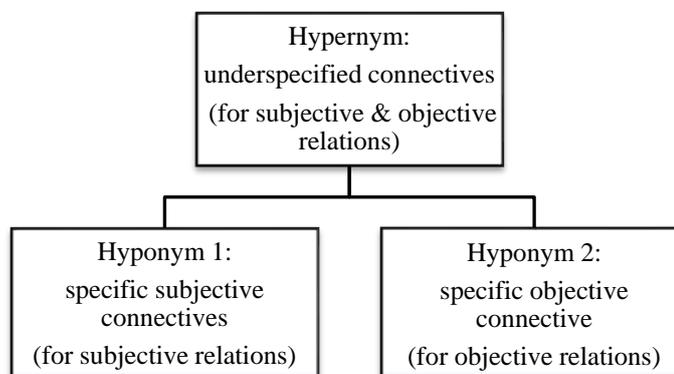


Figure 1. Diagram of connectives encoding different degrees of subjectivity (cp. The taxonomy of linguistic markers of relations in Knott & Dale, 1994; and Knott & Sanders, 1998)

The English connective *because* and the Chinese connectives *yinwei* ‘because’ and *suoyi* ‘so’ fall into the category of connectives that are underspecified for subjectivity (Li, Evers-Vermeul, & Sanders, 2013; Li, Sanders, & Evers-Vermeul, 2016). Other connectives do provide indications about the degree of subjectivity of the relation. These connectives seem to vary systematically in terms of subjectivity: some express subjective relations, others are specialized in objective relations. For instance, French *parce que* ‘because’ (Stukker & Sanders, 2012; Zufferey, 2012), German *weil* ‘because’ (Stukker and Sanders, 2012), Dutch *omdat* ‘because’, *daarom* ‘that’s why’, *doordat* ‘because of the fact that’, and *daardoor* ‘as a result’ (Pander Maat & Degand, 2001; Pander Maat & Sanders, 2000; Sanders & Stukker, 2012) and Chinese *yin’er* ‘as a result’ (Li et al., 2013) are typically used to express objective relations. By contrast, other connectives are prototypical markers of subjective relations, such as French *puisque* ‘because’ and *car* ‘because’, German *denn* ‘because’, Dutch *want* ‘because’ and *dus* ‘so’, and Chinese *kejian* ‘so/therefore’. The Chinese examples (8) and (9) illustrate the use of underspecified connectives and specific connectives in expressing the Chinese counterparts of (6) and (7).

(8)

Qiang fengbao zuoye xiji le zhe ge chengshi, *suoyi/yin’er* gonggong jiaotong xitong zanshi tanhuan.

Strong storm last:night strike ASP(PFV) this CL city, *so/as a result* public traffic system temporarily paralyzed.

A strong storm stroke the city last night, *so/as a result* the public traffic system was temporarily paralyzed.

(9)

Fengbao guohou gonggong jiaotong xitong zanshi tanhuan, *suoyi/kejian* zhe chang fengbao yiding feichang qiang.

Storm after public traffic system temporarily paralyzed, *so/therefore* this CL storm must very strong.

The public traffic system was temporarily paralyzed after a storm, *so/therefore* the storm must have been very strong.

The lexicon of Mandarin Chinese offers a variety of choices (illustrated in Figure 2) to express causal relations (cf. Li et al., 2013; Lü, 1999; Xing, 2001): it has an underspecified connective *suoyi* that can express both objective and subjective causal relations – similar to English *so*; it also has a specific objective connective *yin'er* (similar to the English phrase *as a result* and the Dutch connective *daardoor* ‘as a result’) to express objective relations such as in (8), as well as a subjective connective *kejian* (similar to English *therefore* and the Dutch connective *dus* ‘so’), which is typically used to express subjective relations as illustrated in (9). Although the word *kejian* contains a vision verb *jian* ‘see’, which contributes to a literal interpretation of the word *kejian* – ‘(from this) can see’, the verb *jian* can be considered as a grammaticalized element in the construction (Tao, 2007). It is this use of *kejian* as a connective, as well as the use of the connective *suoyi* that is the focus of this dissertation. The rich lexical system of connectives in Chinese enables us to investigate subjectivity in discourse in more detail.

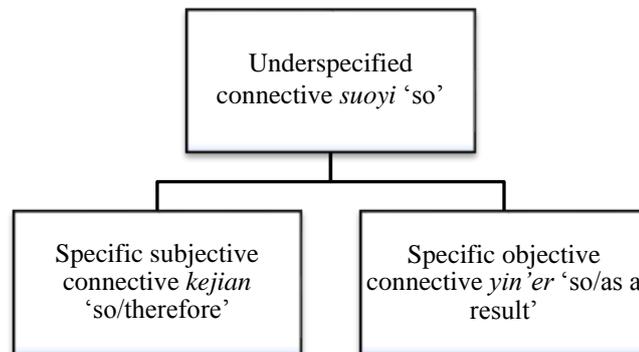


Figure 2. The Chinese connectives encoding different degrees of subjectivity.

Based on the semantic profiles of Chinese connectives proposed by Li et al. (2013, 2016) and Li (2014), this dissertation explores other linguistic markers of subjectivity used in the context of these connectives, such as perspective markers and modal verbs. In addition, this dissertation focuses on the effects of the combined use of connectives,

perspective markers and modal verbs on the processing of causal discourse relations.

1.4 The processing of subjectivity

The degree of subjectivity of a relation has an effect on the way in which readers process the relation. Traxler, Bybee and Pickering (1997) and Traxler, Sanford, Aked and Moxey (1997) provided the first empirical evidence on the complexity of subjectivity in causal relations: longer reading times were needed to process a subjective relation compared to that of an objective relation. The English sentences tested in the experiment were marked by *because*, which is an underspecified connective – processing delays were found in the second clause of the causal relation compared to that same clause in an objective relation. Subjective relations require the construction of a mental discourse representation in which not only the propositional content of the clauses is present, but also a source of information. Longer reading times followed as a result of this extra cognitive effort to construct such a subjective mental representation.

Building on the connective profiles mentioned above, on-line reading studies have investigated their roles in processing. A variety of studies have shown that connectives function as processing instructions for sentence comprehension (Cozijn, Noordman, & Vonk, 2011; Cozijn, 2000; Lorch & Lorch, 1986; Millis & Just, 1994; van Silfhout, Evers-Vermeul, Mak, & Sanders, 2014), and that the specificity in terms of subjectivity encoded in connectives influences processing. For example, Canestrelli, Mak and Sanders (2013) conducted a study on Dutch causal relations. In subjective relations marked by the specific subjective connective *want* ‘because’, Canestrelli and colleagues precisely located the processing delay of subjectivity at the region right after the connective compared to the objective connective *omdat* ‘because’. Subjective connectives instruct readers on the forthcoming subjective nature of the relation, and thus immediately trigger the construction of a mental discourse representation that includes a speaker’s perspective. In a study on Mandarin Chinese, Li, Mak, Evers-Vermeul and Sanders (2017) found results that supported Traxler et al.’s (1997a) findings for subjective relations marked by underspecified connectives: readers slowed down in the second clause of the subjective relations expressed by the underspecified connective *suoyi* ‘so’.

The effects of connectives on processing can be influenced by other linguistic elements in the discourse, such as perspective markers. In both Traxler et al. (1997b) and Canestrelli et al.’s (2013) studies, a facilitation effect of a perspective marker in the clause preceding the connective was found, i.e. the reading delay in the second clause or right after the connective was cancelled out by the presence of perspective markers such as *John thinks*, *perhaps* and *according to Peter*. These perspective markers already indicate that the propositional content arises from someone’s mind,

and thereby provide subjectivity information before the subjective connective marks the subjectivity of the relation. At the connective, therefore, readers do not have to put extra effort in constructing a mental representation that includes a source of information, since that representation is already in place.

The empirical evidence on the processing of subjectivity and the role of connectives and perspective markers leads to the question how different linguistic expressions of subjectivity interact. In language use, are there any collocation patterns between connectives and other linguistic elements, such as perspective markers? In language comprehension, how do different types of perspective markers (e.g. markers of evidentiality, certainty, attitude and emotion, etc.) influence the processing of subjectivity in discourse relations? By investigating these linguistic markers of subjectivity, the aim is to shed light on the encoding of subjectivity in discourse, as well as the mental representation of subjectivity.

Subjectivity, as a conceptual phenomenon, exists in all languages, although the linguistic realizations of subjectivity may differ across languages, as Figure 1 indicated. It is interesting to look at the way different linguistic cues help readers/hearers make interpretations in various languages. I will look at Mandarin Chinese which is typologically very different from Western languages. In addition, we know much less about the role of subjectivity in Chinese. If subjectivity is a basic cognitive notion, we would expect similar effects in different languages, although the linguistic marking of subjectivity may vary across languages. In this sense, the mental representation of subjectivity in Chinese is not expected to be different from the representations in the languages that are well studied in this respect. How subjectivity is encoded in Chinese and how various linguistic cues influence the processing will be the focus of this thesis. Given the rich inventory of connectives (see Section 3), it is promising to conduct more processing studies on the Chinese language.

1.5 Methods and research questions

The variety of linguistic expressions of subjectivity and its complexity in mental representations call for a combined approach with both corpus-based and experimental research. In this dissertation, three different methods will be applied to investigate subjectivity in discourse: collocational corpus analysis, visual world paradigm eye-tracking experiments and on-line reading experiments. With this combined approach, the goal is to answer several important questions on subjectivity in both language use and language comprehension.

1.5.1 Corpus-based study: collocational analysis of Chinese connectives

Previous experimental reading studies have illustrated the effects of contextual cues on the processing of subjective relations and also the interference of contextual cues on the processing effect of connectives (Canestrelli et al., 2013; Traxler et al., 1997b). One way to investigate the subjectivity information encoded in different connectives is to investigate their contextual features. Collocational analysis, as one of the statistical methods in corpus linguistics, has been used as an important tool to investigate word meanings and contextual features (Gries, 2013; Gries & Stefanowitsch, 2004; Speelman & Geeraerts, 2009). This analysis is based on occurrences of collocates appearing in the context of the target word. For instance, the causal connective *yinwei* ‘because’ in Chinese is a collocate in the context of the connective *suoyi* ‘so’, as shown in example (10a), because double marking is very common in Chinese to mark causal relations. However, in languages such as modern English, where double marking of relations is no longer a frequent phenomenon¹, the collocation pattern of *because* and *so* as in (10b) comes across as ungrammatical. Therefore, in English, *because* is probably a word that is not attracted but repelled by the connective *so*.

(10)

a. Chinese

Yinwei Anni shangzhou shuaishang le gebo, **suoyi** ta jintian meiyou lai shangke.

Because NAME last:week break ASP(PFV) arm, **so** 3SGF today NEG come class.

Because Anne broke her arm last week, **so** she didn’t come to class today.

b. English

Ø/ *Because Anne broke her arm last week, **so** she didn’t come to class today.

By measuring the collocational strength between the target word and the words in its context, the words that are in strong association or repulsion with the target word, surface as important contextual features. These features help to distinguish the target word from linguistic alternatives with comparable but not exactly the same meanings.

Up till now, previous corpus studies on the Chinese language have investigated

¹ Double marking of causal relations such as *because P, so/therefore Q* was used in Old English, but was replaced by the form of single marking in Modern English (Traugott, 1989).

individual connectives – how they specify relations (Li et al., 2013; Lü, 1999; Xing, 2001), but they have not seriously looked into the collocates of connectives in the segments. It is hypothesized that connectives encoding different degrees of subjectivity may differ in their collocates. For instance in example (9), the modal verb *must* co-occurs with connectives to express a subjective relation. This study will investigate the underspecified connective *suoyi* ‘so’ and the specific subjective connective *kejian* ‘so/therefore’ in different genres and contexts, exploring how the two connectives are used in combination with other linguistic elements in Mandarin Chinese corpora. These two connectives have been selected firstly because they are both relatively frequent (see the Lancaster Corpus of Mandarin Chinese; McEnery & Xiao, 2004). Secondly, they can be studied using written Chinese because they either occur frequently in both written and spoken language (*suoyi*) or are predominantly used in written language (*kejian*). The following two research questions will be explored:

- RQ1. Do connectives of different subjectivity degrees differ in their types of collocates?
- RQ2. More specifically, do connectives differ in the types of perspective markers they co-occur with?

This collocational study will provide important clues to understanding the mechanism of expressing subjectivity in discourse. In addition, the collocation patterns obtained from the authentic language data will provide directions and materials for further investigations on the comprehension of subjectivity in causal relations.

1.5.2 Visual world paradigm processing study

Theoretical studies on subjectivity suggest a different mental representation of subjective relations in comparison to that of objective relations. According to the definition of subjectivity, the mental representation of subjective relations involves a person making the inference – the *Subject of Consciousness* (Pander Maat & Sanders, 2000; Sanders et al., 2009; Sanders & Spooren, 2015). On-line reading studies provide evidence for such a conceptual difference between subjective relations and objective relations: longer reading times are associated with the processing of subjective relations, which reflect relative processing complexities. Interpreting the involvement of the *Subject of Consciousness* for subjective relations apparently costs extra cognitive effort, and this results in longer reading times. However, differences in reading time found in reading experiments do not tell us what exactly people are processing during the longer intervals and what the resulting mental representations are.

The aim of the experimental study in Chapter 4 is to reveal the process behind

the comprehension of subjectivity by measuring people's fixations in a visual display with or without the speaker during processing. As illustrated in (11), the sentences participants will hear in the experiment encode different degrees of subjectivity (objective vs. subjective relations), and are marked by different connectives (i.e. subjective, objective or underspecified connective). Meanwhile, participants will be presented with pictures on the screen – one of which contains a speaking person. The looks of participants on the picture over time reflect their focuses of attention in language processing.

(11)

a. Objective sentence

Liu Ziming zuo zai diyi pai dou kan bu qingchu heiban, **suoyi/yin'er** ta de chengji shoudao le hen da yingxiang.

NAME sit at first row even see NEG clear blackboard, **so/as a result** 3SGM MOD grade suffer ASP(PFV) very big impact.

Liu Ziming can't see the blackboard clearly even at the first row, **so/as a result** his grades have been seriously affected.

b. Subjective sentence

Liu Ziming zuo zai diyi pai dou kan bu qingchu heiban, **suoyi/kejian** ta de shili **yiding** bu tai lixiang.

NAME sit at first row even see NEG clear blackboard, **so/therefore** 3SGM MOD eyesight **must** NEG very ideal.

Liu Ziming can't see the blackboard clearly even at the first row, **so/therefore** his eyesight **must** be very limited.

By performing a series of visual-world paradigm processing experiments in two typologically different languages – Dutch and Chinese, Chapter 4 focuses on the mental work people conduct when processing subjectivity and causality:

RQ3. Does the processing of subjectivity involve tracking of the sources of information?

RQ4. What is the role of connectives and modal verbs in tracking the information source in situation models?

It is generally accepted that comprehenders make mental representations when processing discourse (Britton, 1994; Graesser, Millis, & Zwaan, 1997; Knott, Sanders, & Oberlander, 2001; Zwaan & Rapp, 2006). If the processing of subjectivity indeed involves tracking the source of information in the mental representation, people should pay more attention to a visual display of the speaker when interpreting subjective relations. Moreover, subjective connectives and modal verbs, which convey the information of subjectivity, are predicted to function as processing

instructions that guide the process of tracking the source of information. We will investigate these hypotheses in Chapter 4.

1.5.3 On-line reading study

A further question is how the combined use of connectives and subjectivity indicators influences language processing. Based on the findings from the collocational study, it is possible to investigate this research question with a specific focus on the collocates that are relevant to the Chinese connectives – some linguistic cues may consistently pattern with one type of connective instead of the other in the corpus. Example (12) shows the combined use of the attitudinal type of perspective marker – *jingran* ‘surprisingly’ with *kejian*; example (13) exhibits a case of collocation between an epistemic perspective marker – *jushuo* ‘it is said’ with *suoyi*. Both types of perspective markers indicate the presence of a speaker. However, they introduce the perspective of the speaker in different ways: *surprisingly* introduces the speaker by expressing the attitude of the speaker, while *it is said* involves the speaker by introducing his/her opinion.

(12)

Chahuanü he Fu’ermosi **jingran** tongshi shou huanying, **kejian** 1920 niandai de Shanghai wenyi jie hen duoyuan.

La Traviata and Holmes **surprisingly** at:the:same:time PASS welcome, **so/therefore** 1920 age MOD Shanghai art circle very diverse.

Surprisingly, La Traviata and Holmes were both very popular (in Shanghai), **so/therefore** the art circle of Shanghai in the 1920s was very diverse.

(13)

Jushuo diqiu shang de shui shi you huixing dailai de, **suoyi** nin de zhexie yunshi he shui yiyang bu xihan.

It is said earth on MOD water COP from comet bring PRT, **so** 2SG POSS these meteoric stones with water same NEG unusual.

It is said that the water on the earth was brought by comets, **so** your meteoric stones are the same as water – nothing unusual.

If speakers and writers display different patterns in their usage of subjective versus underspecified connectives in language use, it can be expected that listeners and readers will be sensitive to such patterns. In other words, in language comprehension, readers are expected to be sensitive to the collocation patterns in language use, as should be reflected in their reading times. With the collocation information from the corpus-based study in Chapter 3, the on-line reading study in Chapter 5 will investigate how different types of perspective markers appearing in collocation with connectives guide the processing of subjective relations. Moreover, the interaction of

these markers with the effect of connectives in processing will be explored.

- RQ5. How does the presence of perspective markers affect the processing patterns at the connective region and at later regions in subjective relations with an argument-claim structure?
- RQ6. Do these effects of epistemic stance markers on the processing of subjective relations also hold for attitudinal stance markers?

Although previous reading studies have shown the effects of connectives and perspective markers in processing subjectivity, the proposed study derived from collocation tendencies in the corpus study will bring important new insights into this field of research from at least three respects. First, the linguistic elements that will be tested in the reading experiment are selected on an empirical basis – they are found as an important type of collocates of the connectives under investigation. Therefore, the effects of these linguistic elements can be linked to their distributions in the corpus and be interpreted from the perspectives of both language production and language comprehension. Second, a wider range of perspective markers will be investigated, such as the epistemic stance markers *it is said* and *apparently*, and attitudinal stance markers *surprisingly* and *(un)fortunately*.

1.6 Chapter overview

This dissertation contains six chapters, starting with an introduction and ending with a conclusion chapter. In the intermediate four core chapters, I will present the empirical investigations on discourse representation. All of these main chapters (Chapter 2 to 5) have been written as individual papers. As a result, there is some overlap in the literature reviews in these chapters. The advantage of this arrangement is that each chapter can be read independently.

Chapter 2 is instrumental in introducing the method of collocational analysis for Chapter 3. I will explore the method of collocational analysis to address potential issues in its application to the Chinese language, especially in the field of discourse studies. In Chapter 3 I will report a collocational corpus study on Mandarin Chinese. I aim to examine with what elements subjective and objective causal connectives co-occur in natural discourse; in other words, how Chinese connectives with different degrees of subjectivity collocate with other linguistic elements in discourse, specifically, perspective markers.

Then I will discuss the on-line processing effects of connectives in interaction with other linguistic elements in both reading and visual world paradigm experiments. Chapter 4 investigates the influence of connectives as linguistic cues on the cognitive representation and interpretation of subjectivity information in discourse. I will report

two processing experiments testing the effect of connectives with the visual world paradigm (VWP). Two languages (Dutch and Chinese) with different connective categorizations will be tested in comparison. Based on the findings from the corpus-based collocational study and the VWP study, I will present an eye-tracking reading experiment in Chapter 5, which tests the effect of perspective markers in instructing readers in the comprehension of discourse relations.

The dissertation ends with a general discussion on the theoretical issues that the empirical evidence sheds light on, such as the definition of subjectivity and the mental representation of subjectivity. The advantages of combining different methodologies will also be discussed.

Chapter 2. Collocational analyses at the discourse level using Chinese corpora

2.1 Introduction

An important advantage of linguistic corpora is that they allow linguists to examine naturally occurring data that are representative of the language population under investigation (McEnery & Hardie, 2012). Modern corpora represent more and more genres and modalities, and have shown an enormous increase in size. Techniques of corpus linguistics have also developed dramatically, with inspiring input from computer science and statistics. These improvements in both corpora and methods open up new opportunities for linguistic research at the discourse level, thereby allowing us to enhance our understanding of human language.

Collocational analysis is one of those quantitative methods of large-scale data analysis in corpus studies (Church, Gale, Hanks, Hindle., 1991; Church & Hanks, 1990; Evert, 2005, 2008; Manning & Schütze, 2000; Stefanowitsch & Gries, 2003). In the past decades, collocational analysis has been applied to investigate syntactic and semantic phenomena in Western languages (Boogaart, Coleman, & Rutten, 2014; Church et al., 1991; Gries & Stefanowitsch, 2004; Mukherjee & Gries, 2009; Stefanowitsch & Gries, 2003, 2008). For instance, Church et al. (1991) investigated the differences in meaning between *strong* and *powerful* by looking at the words in associations with them, and Gries and Stefanowitsch (2004) compared the ditransitive construction and the *to*-dative construction by analyzing the collocates of these two constructions at the verb slot. Similar studies are available in Chinese, where quantitative methods and tools are employed in the field of computational linguistics. For instance, Huang et al. (2005) and Huang, Hong, Ma and Šimon (2015) explored the possibilities of extracting grammatical classes from corpora with the *Sketch Engine* platform (Kilgariff et al. 2014), and Gong, Ahrens and Huang (2007) applied a frequency-based collocational approach to search for lexical mappings in Chinese on the basis of a large-scale corpus.

With this methodology chapter, we provide an overview of potential challenges in performing collocational analyses on Chinese linguistic phenomena at the discourse level, as well as some possible solutions. Instead of looking at similarities and differences in meaning between words/expressions at the semantic level, or the syntactic dependency between words/expressions and constructions at the syntactic level, collocational analysis at the discourse level aims at a more global perspective: the relation between discourse segments such as clauses, sentences and paragraphs, which may influence the choice of words/elements used in the context. For example, more evaluative adjectives can be expected in claim-argument relations than in cause-

consequence relations. In addition, collocational analysis at the discourse level allows researchers to select a search context meaningful at the discourse level, instead of a context composed of superficially adjacent words (an issue that will be taken up for discussion in Section 3.2).

Previous collocational studies at the discourse level mainly focused on Western languages, but in the field of Chinese linguistics, analyzing large scale corpus data with statistical tests has not gained much attention. The majority of studies only provide qualitative analyses, and often refer to anecdotal examples to illustrate their claims (e.g., Wang, 2006; Yin, 2008; Zhang, 2011). Of course, this qualitative approach is valuable in itself, because the categories for classification must first be identified before they can be counted (McEnery & Wilson, 2001; Schmied, 1993). Other studies do provide quantitative data, but are restricted in the sense that they only present percentages or frequencies (e.g., Qi, 2007; Tang & Zhu, 2008; Yin, 2011). However, inferential statistics are necessary to enable researchers to generalize conclusions beyond actual observations (McEnery & Hardie, 2012; Núñez, 2007).

For example, Li (2014) has reviewed Chinese studies on the categorization and properties of present-day Chinese causal connectives such as *yinwei* ‘because’ and *jiran* ‘since’. The main methods these studies use are providing individual examples (e.g., Deng, 2007; Guo, 2008; Xing, 2002), and presenting corpus-based studies with frequency counts (e.g., Li, 2011; Li & Liu, 2004). For example, Xing (2002) has conducted a small-scale corpus-based study into novels and texts on political theory, finding that *youyu* ‘because’ is mainly used in argumentative texts, and seldom in narrative texts. Corpus-based analyses also reveal that certain connectives, such as *jiran* ‘since’, tend to co-occur with discourse clauses expressing subjective opinions (Li & Liu, 2004). Li and colleagues (Li, Evers-Vermeul, & Sanders, 2013; Li, Sanders, & Evers-Vermeul, 2016) have taken the analysis of Chinese causal connectives a step further by investigating their use in different genres, and by applying inferential statistics to the data. They show that certain connectives display a profile that is robust across informative, narrative and argumentative genres, whereas other connectives appear to be genre-sensitive.

With the advantage of corpora and methods for conducting corpus-based analyses, the study of the Chinese language, including the use of causal connectives, can be taken a step further. One attractive option for studying the use of discourse connectives from a more comprehensive view is to investigate discourse connectives in relation with discourse features or other discourse elements. Studying a word in its context provides more insight into the properties of the word, as Firth (1957: 11) has already argued: “you shall know a word by the company it keeps”. From this perspective, collocational analysis based on associations between words is considered a good choice.

The purpose of the current chapter, therefore, is to provide an overview of potential issues in the practice of performing a Chinese collocational analysis, as well

as some possible solutions. In Section 2.1, we will give a more precise definition of the notion collocation. In Section 2.2, we will address statistical methods to assess whether words in the context of a target word should be considered a collocate. Section 3.1 and 3.2 introduce issues related to the segmentation of words and of discourse segments. Section 4 will address some considerations for selecting Chinese corpora that are suitable for conducting collocational analyses, and provides an overview of Chinese corpora currently available.

We have two reasons to devote special attention to the Chinese language: (i) not much research has been done with the new collocational method on Chinese; and (ii) some special technical details need to be considered for Chinese collocational analysis. However, the issues raised, and the methods and strategies introduced for conducting collocational analyses can also be applied to other linguistic phenomena, as well as to other languages.

2.2 Introduction to the study of collocations

As Gries (2013: 138-139) has pointed out, all researchers who would like to conduct a collocational study have to make decisions on a variety of dimensions or parameters. In a way, they thereby create their own definition of the notion of collocation. Section 2.1 will discuss five parameters that determine the type of elements under investigation. Section 2.2 discusses ways to determine the frequency with which these elements have to co-occur before they are considered a collocate, a sixth parameter mentioned by Gries (2013).

2.2.1 Defining collocations

Obviously, researchers first have to select “the nature of the elements” to be observed (Gries, 2013: 138). Originally, the notion of collocation was introduced for characteristic and frequently recurring word combinations (Firth, 1957). This focus on words is also apparent in Evert’s (2008: 1214) definition – “a combination of two words that exhibit a tendency to occur near each other in natural language, i.e. to co-occur” – which includes, but is not restricted to lexicalized word combinations. As he points out, however, a restriction on the word level is not necessary: “while collocations are most commonly understood as combinations of orthographic words, delimited by whitespace and punctuation, the concept and methodological apparatus can equally well be applied to combinations of linguistic units at other levels, ranging from morphemes to phrases and syntactic constructions” (Evert, 2008: 1215). In the literature, the syntactic type of collocations is often referred to as *collostructions* (Evert, 2008; Gries & Stefanowitsch, 2004).

The second and third parameter can broaden or restrict the type of elements that

are considered collocates. The second parameter deals with the degree of lexical and syntactic flexibility of the collocates involved (Gries, 2013: 139). For instance, in the case of words, researchers may be interested in co-occurrence with exactly the same form – e.g. looking at collocates of the noun *woman* – or they may increase the flexibility of their approach by focusing on lemmas (e.g. by including both *woman* and *women* as input for their collocational analyses). The third parameter concerns the role that semantic unity and semantic non-compositionality or non-predictability play in the definition; often, it is assumed that the elements considered as collocates exhibit something unpredictable in terms of form and/or function (Gries, 2013: 139).

A fourth parameter concerns the number of collocates that make up the collocation (Gries, 2013: 138). In most cases, this value is “two”, but the number of collocates is not restricted to this value. An N-gram analysis, for example, allows collocations composed of a sequence of N words in a fixed order, which could result in bigrams (N = 2), trigrams (N = 3), etcetera, depending on the value of N that is chosen (de Kok & Brouwer, 2011).

The fifth parameter concerns the distance and/or (un)interruptability of the collocates (Gries, 2013: 139). The most frequently used option is to focus on elements that are directly adjacent. Alternatively, researchers may be interested in elements that are syntactically or phrasally related but not necessarily adjacent, or they can investigate collocates that are more distant but still co-occur within a window of N words, or within a specific unit, such as a sentence. The choice for one of these options along this fifth parameter largely depends on the type of co-occurrence the researcher is interested in: surface co-occurrences, textual occurrences or syntactic co-occurrences (Evert, 2008). In surface co-occurrences, elements appear within “a certain distance or collocational span, measured by the number of intervening word tokens” (Evert, 2008: 1215). For instance, span size *L2, R3* restricts the search window to two words to the left and three words to the right of the target word.

Instead of setting an arbitrary span size, it is also possible to search for collocates within a certain textual unit, which results in textual co-occurrences. Here, researchers may look for collocates within a sentence, a clause, a paragraph or even a whole document, which allows them to detect weaker dependencies between words. However, this approach presupposes that the researchers rely on or impose segmentation boundaries in the corpora they investigate. This may not always be an easy job, as we will discuss in Section 3.

Syntactic co-occurrences, by definition, refer to the elements that collocate in the contexts of some specific syntactic pattern, such as *verb + object noun*, *adjective + noun*, or the ditransitive construction (Bartsch, 2004; Evert, 2008). Such syntactic co-occurrence analyses reveal *collexemes* that are associated with slots in specific syntactic patterns (Gries & Stefanowitsch, 2004). For example, if the research goal is to find which words appear frequently as modifiers of the noun *female*, it makes sense to look for the adjectives that accompany this noun. Similarly, if one wants to find

out which verbs are the significant collexemes of the verb slot of the ditransitive construction, then the target elements are verbs (present at the lemma level), such as the verb *give* in the construction *I give her a gift*.

In sum, by making choices along the parameters distinguished by Gries (2013), researchers develop their own definition of the collocations that feature in their study. To improve the transparency of research and facilitate comparison across studies, researchers need to be explicit about the choices they make along each dimension. Also, awareness of the parameters can help researchers to link their research questions with what the method enables them to do, and facilitate interpretation of the output.

2.2.2 Applying statistics in order to determine collocations

In addition to the parameters that define the type of elements that are examined in a collocational study, Gries (2013: 138) also distinguishes a parameter that concerns the frequency of the elements under investigation: the researcher needs to decide upon the number of times an expression must be observed before it counts as a collocate.

Some previous Chinese studies investigated the meaning of linguistic elements by looking at the expressions they co-occur with (e.g., Wang, 2006; Yin, 2008; Zhang, 2011). Some calculate raw frequencies of the co-occurrences of expressions in texts (Qi, 2007; Tang & Zhu, 2008; Yin, 2011), as is also common in the case of N-gram analyses, but they do not set any threshold or apply any inferential statistics. Although these studies have offered appealing accounts for linguistic phenomena in Chinese, we prefer to use more sophisticated statistical methods. Researchers could start by looking at collocates that occur more frequently than expected by chance, but there are thresholds and statistical scores other than raw frequencies of co-occurrence to measure the associations between target elements and collocates (e.g. PMI, Dice, Delta P, Odds Ratio, Chi-square, G^2). By calculating additional statistics for collocates, one can rank the relevant collocates and set the threshold above a certain value as the cut-off line for “important” collocates, or simply select the top N items (e.g. top 100).

The statistical examination involves measures for the association between target words (nodes) and candidate collocates. Despite the variations of different collocation types, the measures for the association are all derived from the same contingency table: Table 1 (taken from Gries (2013: 140)), which is comparable to Evert’s (2008) contingency table). In the case of surface co-occurrence, every word pair is referred to as word₁ and word₂, and *a*, *b*, *c*, *d* are the observed co-occurrence frequencies of the respective combinations. Suppose the research question is: what are the important collocates of the target word (word₁)? Each word₁ creates a target context with a certain search span, i.e. the neighborhood of word₁. All words that appear in the target context at least once contribute to a list of words₂, the candidate collocates. For every word in the whole corpus, two questions are asked to count the observed frequency. First, is this word word₂? If this word is word₂, it will be located in either of the two

left cells (contributing to the frequency of a or c); if this word is not word_2 , it will be in either of the right cells (contributing to b or d). Second, does this word appear in the context created by word_1 ? If it appears in the neighborhood of word_1 , it will end up in either of the upper cells (contributing to a or b), and if it appears outside the neighborhood of word_1 , it will be in the lower cells (contributing to c or d). The advantage of this method is manifest: the observed frequency of a word pair (a) is never calculated in isolation, but rather with regards to the reference levels (b, c, d).

Table 1

Co-occurrence table

	word ₂ : present	word ₂ : absent	Totals
word ₁ : present	a	b	$a+b$
word ₁ : absent	c	d	$c+d$
Totals	$a+c$	$b+d$	$a+b+c+d$

An example can illustrate the logic behind Table 1. If the target word of the research is the connective *suoyi* ‘so’ (word_1), this means that one is interested in the context features of this word. Suppose two of the occurrences of this connective *suoyi* in the corpus are the ones in (1).

(1)

- a. 因为这家餐厅物美价廉，所以它的生意一直很好。
 Yinwei zhe jia canting wumei jialian, **suoyi** ta de shengyi yizhi hen hao.
 Because this CL restaurant good:product cheap price, so 3SGN MOD
 business always very good.
 This restaurant has delicious food with reasonable price, so its business
 has always been good.
- b. 因为莉莉说她讨厌比萨饼，所以他们决定去吃寿。
 Yinwei Lili shuo ta taoyan bisabing, **suoyi** tamen jue ding qu chi shousi.
 Because NAME say 3SGF hate pizza, **so** 3PL decide go eat sushi.
 Lily said she hated pizza, **so** they decided to go for sushi.

If the search span is set to ten words to the left and ten words to the right, the sequence of words around the connective *suoyi* ‘so’ in (1a) generates the following word list: *yinwei* ‘because’, *zhe* ‘so’, *jia* (classifier), *canting* ‘restaurant’, *wumei* ‘good product’, *jialian* ‘cheap price’; *ta* ‘3SG’, *de* (modifier), *shengyi* ‘business’, *yizhi* ‘always’, *hen* ‘very’, *hao* ‘good’. For (1b), the list contains *yinwei* ‘because’, *Lili* (name), *shuo* ‘say’, *ta* (3SGF), *taoyan* ‘hate’, *bisabing* ‘pizza’; *tamen* (3PL), *jue ding* ‘decide’, *qu* ‘go’, *chi* ‘eat’, *shousi* ‘sushi’. The twenty-five words form a list of candidate collocates of *suoyi*

‘so’. Twenty-five individual tests need to be done: one for each of the twenty-five words to measure the association strength between each word and the connective *suoyi* ‘so’. In the test with *yinwei* ‘because’ as word₂, for example, every word in the corpus except the target word *suoyi* ‘so’ is examined with two questions: whether this word is *yinwei* or not; and whether this word appears in the context of *suoyi* or not. If the answers to both questions are ‘yes’, then this word adds to the frequency of *a*; if the answer to the first question is ‘yes’ and to the second question ‘no’, then it adds to *c*, etc. The same procedure is repeated for the words *zhe* ‘so’, *jia* (classifier), *canting* ‘restaurant’, etc. as word₂. *Yinwei* ‘because’ will have a higher association score compared to other words in this small corpus, because it appears twice in the context of *suoyi* ‘so’. Therefore, the association strength test will identify *yinwei* ‘because’ as an important collocate of *suoyi* ‘so’, which makes sense given that double marking of causal relations is very common in Mandarin Chinese.

Expected frequencies of occurrences are the occurrences of each combination under the null hypothesis that word₁ and word₂ are independent of each other (Evert, 2005). The observed frequencies (*a*, *b*, *c*, *d* in Table 1) and the expected frequencies can be used to calculate the strength of association between the target word and each of the particular candidate collocates. Note that if the two sentences in (1) would constitute the entire corpus, none of the words would show a relatively strong association to *suoyi*, because all the words apart from *yinwei* only occur once in the context of *suoyi*.

Association coefficients can be approached from two perspectives: statistical significance and effect size. Significance measures evaluate the difference between the observed co-occurrences and expected co-occurrences from the perspective of a statistical test: how much evidence do we have to establish an actual difference? Statistical association measures based on the amount of evidence include tests such as the Chi-squared (χ^2) test, G^2 test, t-test, and Fisher-test. With a low p-value (usually $< .05$): attraction between a collocate and the target word is established if the observed frequencies are higher than the expected frequencies; repulsion is at stake if observed frequencies are lower than the expected frequencies.

Effect size measures such as PMI, Dice, Delta P, and Odds Ratio, evaluate the magnitude of the difference between the observed co-occurrences and expected co-occurrences (Evert, 2008; Gries, 2013). Association scores produced by effect size measures indicate how strong the attraction or repulsion is between the target word and the collocate. A brief summary of the interpretations of association scores from different measures is listed in Table 2.

Table 2

Summary of some effect size measures (based on Evert, 2008; Speelman, to appear)

Effect size measures	Attraction	Repulsion	Neutral
PMI	>0	<0	0
Odds Ratio	>1	<1	1
Delta P	[0, 1]	[-1, 0]	0
Dice ²	approaching to 1	diverging from 1	n.a.

The evaluation system of association coefficients can be applied to measure attractions or repulsions of all types of collocations/co-occurrences. For further details on the calculations of association scores between different types of co-occurrences, see Evert (2008, section 3). Theoretical and technical accounts for the differences among various association measures have been discussed by Evert (2005, 2008), Gries (2013), Gries and Stefanowitsch (2004), Pecina (2009) and Wiechmann (2008), which is why we will not go into details here.

As an extension of a simple analysis on collocates, a distinctive-collocate analysis (Church et al., 1991) is a method to detect differences between comparable words (e.g. *kejian* ‘so’ vs. *suoyi* ‘so’) in terms of their collocates. The key difference between a simple collocate analysis and a distinctive-collocate analysis is that the latter takes the context of the competitor word as the reference context instead of the rest of the corpus. For example, two separate simple analyses for the connectives *kejian* and *suoyi* would provide two lists of collocates, one for each of the two connectives, given the words occurring in the entire corpus, whereas a distinctive-collocate analysis will generate one list of collocates that prefer the context of one connective over the other. This analysis can be applied to words as well as constructions. In the latter situation, a specific terminology – distinctive collexeme analysis – is suggested by Gries and Stefanowitsch (2004, 2010).

2.3 Segmentation issues

While conducting collocational analyses, several segmentation issues may occur. First, the segmentation of words matters to the identification of target words and their collocates. This point is especially important for Chinese collocational studies because written Chinese is composed of continuous characters without white spaces between words. In order to apply collocation measures on the association of Chinese words, the first thing to do is to identify the word boundaries. Second, given that the distance

² Dice scores range from zero (perfect lack of attraction) to one (perfect attraction), but there is no single Dice score that at all times can be interpreted as the neutral value in between attraction and repulsion.

between collocates is relevant (Gries, 2013), the segmentation of discourse segments also plays a role (recall the fifth parameter of Gries (2013) discussed in 2.1 on the distance and/or (un)interruptability of the collocates). We will address these issues in turn. Being aware of these issues allows researchers to select tools and techniques, and have a better idea of the limitations of these techniques in approaching natural human language.

2.3.1 Word segmentation

In order to investigate the collocation patterns of words, we first need to define what a word is. Many Western languages use white spaces to separate words. Differences in spacing, however, may result in different outcomes. For instance, in English, *football player* is an expression composed of two individual words, while its Dutch counterpart *voetbalspeler* is written as a compound without a space between the component words. In a search for the collocates of the target *coach* in texts with the words *football/voetbal*, *player/speler* and *football player/voetbalspeler*, the English words *football* and *player* will only appear as two separate collocates, but their combination will not show up in the analyses. In a Dutch search, *voetbal*, *speler* and *voetbalspeler* can all three appear as collocates. This illustrates that word segmentation matters to the identification of target words and their collocates.

Unlike most Western languages, the Chinese writing system does not use white spaces to separate words, which makes the identification of words a bigger issue. Even among native speakers agreement on the definition of words is not high: only 76%, according to a test done by Shih and Sproat (1996). Even though collocational analyses on characters are technically possible, analyses on segmented texts can produce neater output for later interpretation. Fortunately, automatic word segmentation has been introduced as an important technique in Chinese information processing. In the past two decades, three major methods towards Chinese word segmentation have been developed (Liu & Wei, 2008; Long, Zhao, & Tang, 2009; Wang & Guan, 2005):

- (i) a method based on sub-string matching;
- (ii) a method based on the understanding of words;
- (iii) a method based on statistics.

The method based on sub-string matching identifies words according to the matching of each sub-string to the word glossary of existing dictionaries. The method based on the understanding of words trains a machine with syntactic and semantic information in order to get an artificial processing model for word segmentation. Instead of adopting an existing dictionary with the repertory of words, the statistic approach identifies words based on the association statistics of the co-occurring characters in

authentic language data: when the association statistics are higher than a given threshold, the combination of characters is counted as a word. Most of the modern word segmentation algorithms combine the three methods to get higher accuracy rates.

The lack of boundary identifiers (such as white spaces) can result in two types of ambiguity: pseudo-ambiguity and true ambiguity. Pseudo-ambiguity refers to the ambiguity of word boundaries due to the presence of different logically possible combinations of characters in a sequence of strings, where only one alternative is actually realistic. For example, in the sequence of words *he ruan jian* ‘and software’ in (2a), both segmentation (2b) and (2c) are logically possible, but only (2c) provides a reasonable segmentation between the word *he* ‘and’ and *ruanjian* ‘software’. The output in (2b) does not separate the sequence into two sensible words.

(2)

- a. 和软件
He ruan jian
- b. *和软 | 件
Heruan | jian
??? | article
- c. 和 | 软件
He | ruanjian
And | software

The problem of pseudo-ambiguity can be neutralized by applying statistics to the word segmentation system: this type of ambiguity can be eliminated by ruling out the infrequent combinations of characters that just co-occur by chance. True ambiguity, however, cannot be neutralized with simple statistics, because both of the possible segmentations give sensible words. The combination of characters in example (3a) can be segmented as (3b) and (3c), which both generate true Chinese words and can be verified in given contexts. For example, (4a) and (4b) provide the respective contexts in which the segmentations in (3b) and (3c) would make sense (examples are taken from Sun & Zou, 2001: 3). More linguistic information is needed to detect true ambiguity. Some statistic models such as the N-gram model and the Markov model nowadays can produce highly reliable disambiguation of words.

(3)

- a. 地面积
Di mian ji
- b. 地 | 面积
Di | mianji
Ground | square measure
- c. 地面 | 积
Dimian | ji
Surface-of-ground | gather

(4)

- a. 这几块 | 地 | 面积 | 还真不小。
Zhe ji kuai | di | mianji | hai zhen bu xiao
This several CL | ground | square-measures | even really NEG small
These pieces of | grounds | square measures | are not small.
- b. 地面 | 积 | 了厚厚的雪。
Dimian | ji | le houhou de xue
Ground-surface | gather | ASP(PFV) thick MOD snow
The surface of ground | gathered | thick snow.

Another major issue for Chinese (and other) word segmentation tools is the recognition of unknown words. Of course, unknown words are not profiled in the word list of dictionaries, but should still be recognized as separate words. For example, in (5), *Liu Hua Qing* should be treated as the name of a person, and *Chu di* as the name of a place (the example is taken from Sun & Zou, 2001: 7).

(5)

- 刘华清 | 楚地 | 重游。
Liu Hua-qing | Chudi | chongyou
Name | Chu-region (name of a place) | re-travel
Liu Hua Qing | travelled to the Chu region | for the second time.

Various strategies have been introduced to better identify unknown words, such as creating maximally comprehensive word lists based on large-scale corpora, providing input of specialized word repertoires in specific fields, as well as the development of some statistical measures and word generation rules to promote the identification of unknown words.

The discussion above makes clear that word segmentation may affect the outcomes of the collocational analysis. When using corpora without indications of

word boundaries, it is therefore important to select suitable word segmentation tools for the analysis, which minimize the threats posed by these potential problems. Nowadays, several word segmentation tools are available for Chinese, as well as for some other continuous writing systems (e.g. Arabic), thanks to the development of information processing in computer science. We have listed some major segmentation tools in Appendix 1. Selecting a segmentation tool that is suitable for the selected corpus and research question will facilitate the process of performing a Chinese collocational study.

The choice of word segmentation tools depends largely on the choice of the target corpus. In the first place, the word segmentation tool should have a high tested accuracy rate on the chosen corpus or corpora of similar types. For instance, if the research is done with a narrative corpus based on fictions, it may not be wise to use LTP-cloud, which has high accuracy rates tested in *People's Daily* newspaper data. Secondly, researchers always need to weigh the sophistication of the segmentation tool and the speed of the segmentation process. Some tools provide segmentations based on complex dictionaries and profound statistics, such as *NLPIR-ICTCLAS*, but these tools exert high processing load for the PC and require some programming knowledge.

Additionally, POS-tags offered by some segmentation systems bring benefits to collocational studies (Huang et al., 2015). For instance, the Chinese connective *kejian* 'so' is not fully lexicalized as a connective, i.e. it can still be used as a verbal phrase to express a meaning of '(from this) can see'. *NLPIR-ICTCLAS* adds different tags to *kejian*: /c representing a connective use, and /v representing verbal use. The distinction between the two types of *kejian* allows researchers to investigate the nature and use of this word in different contexts.

In order to test the effect of the choice of segmentation tool on the collocation results, we performed a distinctive collocational analysis on two Chinese connectives: *suoyi* 'so' and *kejian* 'so', taking a context of ten words to the left and right into account. Texts were originally retrieved from a raw Modern Chinese corpus without segmentation – the CCL corpus (from the Chinese Center for Linguistics at Peking University), and then segmented twice: once by the *Corpus Word Parser* – which is light and handy – and once by *NLPIR/ICTCLAS*, which is more accurate but requires more computational load.³ Both tools segmented texts by adding spaces between words, and tagging words and punctuations. Association strengths were calculated in R with the *mclm* package⁴ (Speelman, to appear), which provides a collection of

³ We left out the word tags automatically generated by the two segmentation tools, because one word can be tagged with different labels according to different POS-tag schemes.

⁴ The *mclm* package is not available from CRAN yet, but can be obtained from Dirk Speelman, the author of the package, via email request.

functions to support quantitative corpus linguistics.

Table 3

Amount of overlap of two collocates lists from texts segmented by different word segmentation tools (Corpus Word Parser and NLP/ICTCLAS)

	Collocates ranked by G^2		Collocates ranked by Fisher-1 ⁵		Collocates ranked by PMI	
	Count	Percentage	Count	Percentage	Count	Percentage
Top 200 items	146	83%	164	82%	137	69%
Top 100 items	81	81%	80	80%	62	62%
Top 50 items	44	88%	42	84%	28	56%
Top 25 items	21	84%	22	88%	11	44%

Results in Table 3 suggest a considerable overlap between the two collocates lists ranked by the measures that are based on the amount of evidence: G^2 and Fisher-1. However, the percentage of overlap between the two lists is relatively low when the collocates are ranked by PMI, an effect size measure. This difference may be due to the fact that different segmentation tools segment words based on different dictionaries and principles: *NLP/ICTCLAS* integrates a larger dictionary and more principles in the identification of Chinese words, while the *Corpus Word Parser* takes the computational speed as a priority. As a result, some linguistic elements identified as words by *NLP/ICTCLAS* may not be recognized by the *Corpus Word Parser*, and some elements segmented as a one-character word in the *Corpus Word Parser* may be identified as part of a two-character word by *NLP/ICTCLAS*. This results in differences in the word lists generated by the two segmentation tools and used as input for the subsequent collocational analyses.

A comparatively large variation between the two lists ranked by PMI scores can be observed from Table 3, which is mainly due to the nature of the PMI measure. As an effect size measure, PMI evaluates the strength of association regardless of the number of occurrences. Effect size measures are more sensitive to variations in word identification. This explains the variation in the top collocates ranked by PMI. Measures based on the amount of evidence, such as G^2 and Fisher-1, on the other hand, always take into account the frequencies of occurrences. Therefore, they are less

⁵ Fisher-1 refers to the p-value of a one-tailed Fisher Exact test for attraction.

influenced by different outcomes of the word segmentation tools.

To summarize, in general it is preferable to adopt texts segmented by a comparatively sophisticated segmentation tool. Given the absence of strict guidelines on acceptable degrees of overlap in this relatively new field of research, and given that using a lighter tool will save considerable amounts of time, our suggestion is to consider lighter tools such as the *Corpus Word Parser* with around 80% overlap acceptable as well. Given the variation in top collocates that came out of our first comparison of two segmentation tools, however, it seems safer to use the output that results from measures based on the amount of evidence instead of the effect size measures. Special attention needs to be paid when collocation patterns from corpora are tested in processing studies: if the effect size of an association between two words is too small, the predictability of one word on the other is low, i.e. the presence or absence of one word may not help the reader/listener process the other word.

The method of using word segmentation tools facilitates further collocational analyses, but it also has two limitations. First, the accuracy rates of segmentation tools vary across text types, but none of the tools are 100% accurate. Therefore, we cannot expect the segmentation tools to produce perfectly annotated/segmented texts. Second, by performing collocational analyses with segmented texts, we by default accept the definition of words used by the segmentation system. That is, we are forced to select words that are defined as “words” according to the segmentation system. If certain elements are missed out because they are not identified by the segmentation system, certain collocates may go unnoticed in the collocational study. However, this risk is relatively small, because the majority of words in a dictionary will also be parsed as separate words by the word segmentation tools presented in Appendix 1.

2.3.2 Segmentation of target contexts

In Section 3.1, we have discussed the use of word segmentation tools for detecting target words as well as other words in Chinese corpora. The next question for collocational studies is: what is the context within which the collocates of the target word are detected (cf. the fifth parameter we discussed in Section 2.1)?

One of the approaches researchers usually take is to set an arbitrary span size of the search window, for example three words/characters to the left and three words/characters to the right of the target word. The words/characters frequently occurring within that span size are considered to be collocates. This approach is easy and intuitive, but the arbitrary nature of the search span borders may generate two potential problems. First, if no word segmentation has been applied to the corpus, an arbitrary search span may break a word at the border of a context span into two separate characters. The major problem with this compulsory separation of words is that every split character is treated as a word in the word list created for calculating the association strengths. Consequentially, these fake or partial “words” (characters in

fact) created by splitting the context are measured together with other real words. Due to the fact that in principle most Chinese characters can be taken as a word, the system to calculate association strengths cannot tell ‘fake’ words from real words of the same character. Moreover, the association statistics of other words in the corpus are also affected by the existence of such ‘fake’ words. Thus, the accuracy of association scores is influenced as a whole. Segmenting texts beforehand as we illustrated in Section 3.1 and defining search span by words with spaces can fix this problem.

Second, setting an arbitrary search span disregards meaningful discourse boundaries, thereby increasing unexpected noise in the data. For example, the border of the context may be put in the middle of a long sentence or clause, which creates a loss of data in comparison to an analysis in which the entire sentence is used as the span size. Similarly, sentences shorter than the set span size generate extra collocates from the preceding or following contexts, which also affects the calculation of association strengths.

One proposal to partially fix these two problems is to set the search condition as “stop at punctuation/ sentence break/ paragraph break/ section break” in some analytical software such as *WordSmith Tool*, or by defining the context in the *re.boundary* function in the *mclm* R package as intended, e.g. by referring to commas, full stops, etc. However, for sentences longer than the intended search span, the analysis still suffers from information loss.

In discourse analysis, it is therefore often preferable to adopt a span size that makes sense at the discourse level, such as a sentence, a clause, a paragraph or even a whole document (see Section 2.1). This is, for instance, relevant to the study of linguistic elements that frequently co-occur with connectives such as *suoyi* ‘so’ or *yinwei* ‘because’. Technically, discourse segments can be approached by conducting a textual co-occurrence analysis taking one discourse segment as a unit, or by running a surface co-occurrence analysis with special search conditions (e.g. with only the materials between the specified segment boundaries included). These two approaches originate from different theoretical questions. Surface co-occurrence analyses address the question: what are the words appearing in the neighborhood of the target word? In this case, the type of neighborhood does not receive any specification. Textual co-occurrence analyses answer the question: what words tend to appear in the same clause/sentence/paragraph/document as the target word? From a technical perspective, textual co-occurrence analyses have less statistical power than surface co-occurrence analyses, simply due to the fact that the former have fewer data points; textual co-occurrence analyses treat each clause/sentence/paragraph/document as a token, while surface co-occurrence analyses take each word in the corpus as a token.

Discourse segments have been defined as chunks of text expressing a common purpose (Grosz & Sidner, 1986) or a common meaning (Hobbs, Stickel, Martin, & Edwards, 1988). Two general routines have been taken as the minimal unit: sentences (Hobbs et al., 1988) and clauses, as is common in the Cognitive approach to

Coherence Relations (Sanders, Spooren & Noordman, 1992) and annotations based on the Rhetorical Structure Theory (Carlson & Marcu, 2001; Mann & Thompson, 1988). One practical concern in this area is related to the specialty of Chinese discourse structure. The sentence boundaries in Chinese discourse are not as strict as in Western languages. Example (6), taken from the CCL corpus, gives a brief idea of what a Chinese ‘sentence’ could look like.

(6)

由于中期报告所载明的内容涉及到公司最基本的情况，关系到广大投资者的权益，**所以**，股票或者公司债券上市交易的公司依法制定中期报告后，应当依法将中期报告提交给国务院证券监督管理机构和证券交易所，以使上述机构加强对上市交易的股票或者公司债券的监管，保护广大投资者的合法权益。

Youyu zhongqi baogao suo zaiming de neirong sheji dao gongsi zui jiben de qingkuang, guanxi dao guangda touzizhe de quanyi, **suoyi**, gupiao huozhe gongsi zhaiquan shangshi jiaoyi de gongsi zai yifa zhiding zhongqi baogao hou, yingdang yifa jiang zhongqi baogao tijiao gei guowuyuan zhengquan jiandu guanli jigou he zhengquan jiaoyisuo, yishi shangshu jigou jiaqiang dui shangshi jiaoyi de gupiao huozhe gongsi zhaiquan de jianguan, baohu guangda touzizhe de hefa quanyi.

Since the content of interim report concerns the basic situation of the company, concerns the benefit of many investors, **so**, companies which issue public-traded stocks and corporate bonds, after they have made the interim report according to the law, should submit the reports to the securities regulatory body of the State Council and the Stock Exchange, in order to ensure the supervision and regulation of such institutes on the public-traded stokes and bonds, to protect the legal rights of investors.

The whole paragraph in (6), with so much information, is presented as one long sentence (marked by a full stop) in on-going Chinese language. If the whole sentence is taken as one single discourse segment in a collocational analysis where the researcher is interested in collocates within the sentence, the results may be too noisy. The association between some words may be much weaker than that of other words, simply due to the fact that they appear later on in the sentence. Therefore, special attention needs to be paid to the search span in textual co-occurrence analyses. In this case, the clause is preferably taken as the search span instead of the sentence.

An easy way out of this segmentation problem is to use the segmentation system of the corpus itself to define discourse segments. In some annotated corpora sentences/clauses are tagged with sentence IDs as a unique identifier of sentences. For instance, the *HIT IR-Lab Chinese Dependency Treebank* codes the sentence ID as

<sent id="0">, and the *Lancaster Corpus of Mandarin Chinese* (LCMC) applies a similar coding: <s n="0001">. Annotated corpora offer neat and manifest annotation systems, which would definitely save lots of time and efforts in segmenting the context. However, researchers can also be limited by the coding systems of the annotated corpora in detecting boundaries between sentences or clauses. Applying the annotation system of the corpus means that the researcher accepts the segmentation system of the corpus, and disregards more flexible approaches.

Some raw corpora such as the CCL corpus, however, do not provide any annotations. For these corpora, the segmentation tools discussed in Section 3.1 can be applied to get similar tags for punctuations as one can get from the annotated corpora. For example, the *ICTCLAS* and the *Corpus Word Parser* mark the punctuations as /w in the annotated texts they offer. The /w marks commas or full stops, and can be applied to identify the boundary of clauses and/or sentences. *NLPIR-ICTCLAS* also offers a more detailed repertoire of tags (e.g. /ww for a question mark, /wj for a full stop, /wf for a semicolon, etc.), with which individual decisions can be made on boundaries of discourse segments.

Applying segmentation tools to raw corpora cannot produce results as neat as those from a well-annotated corpus. For instance, the *Corpus Word Parser* also uses /w to mark commas, full stops, quotations, parentheses, colons, as well as many other symbols (e.g. %, &, °C, Σ). If /w would be used to define the relevant search span for a collocational analysis on *suoyi* ‘so’, in this case the clauses around the target word, this may result in non-sensical boundaries of the discourse segment. One simple example with the coding /w from the *Corpus Word Parser* can illustrate the point. In (7), both the comma and the quotation mark are coded with /w. Merely depending on the coding system of /w to identify discourse segments, would result in a right context of *suoyi* containing nothing but a quotation mark.

(7)

上海已规定今后在距离轻轨线轨道中心 30 米内不准建造永久性建筑，
/w 所以 “/w 明珠线” /w 建成后，/w 噪音应该对居民的影响不大。
/w

Shanghai yi guiding jinhou zai juli qingguixian guidao zhongxin 30 mi nei
buzhun jianzao yongjiuxing jianzhu./w *suoyi* “/w Mingzhu Xian”/w
jiancheng hou, zaoyin yinggai dui jumin de yingxiang bu da./w

Shanghai (government) has made the regulation that no permanent buildings
are allowed within 30 meters along the metro line, /w *so* “/w Mingzhu Xian”
/w [name of a metro line] should not heavily influence the life of residents
after it is put into use. /w

(CCL corpus)

In such cases, additional adjustments are needed to “clean up” the texts in order to get rid of the noisy marks for symbols that the researcher does not want to take as boundaries for discourse segments, and only keep those that are meaningful. One important note is that we can never achieve perfection in the detection of discourse segments, because with an automatic coding system, there are always cases that would have been segmented in another way in a manual coding. The advantage of using raw corpora along with segmentation tools is that researchers have the freedom and flexibility to define discourse segments at their own intention. However, such benefit comes with the cost of time and effort.

It is not clear how setting the span size actually affects the outcome of collocational analyses. In order to compare an arbitrary span analysis and an analysis with a meaningful discourse segment span, we conducted the same distinctive collocational study of the two connectives as in Section 3.1 (*suoyi* vs. *kejian*), but using two span sizes. In the first analysis, the arbitrary search span was set as 10 words as in 3.1, which is roughly the length of a clause. In the second analysis, the search span was defined as one clause before the target word, and one clause after.

Table 4
Amount of overlap of collocates lists resulting from an arbitrary (plus or minus 10 words) and a meaningful span size (plus or minus one clause)

	Important collocates	Collocates ranked by G ²		Collocates ranked by Fisher-1		Collocates ranked by PMI	
		Count	Percentage	Count	Percentage	Count	Percentage
Top items	200	146	73%	155	78%	106	53%
Top items	100	79	79%	82	82%	43	43%
Top items	50	43	86%	43	86%	19	38%
Top items	25	21	84%	21	84%	8	32%

As Table 4 indicates, considerable overlap can be observed in the lists of important collocates generated by the two types of analysis. Table 4 shows that the degrees of overlap between the clause-span analysis and the arbitrary-span analysis are quite high in Fisher-1 and G² ranking. However, the degrees of overlap between arbitrary and meaningful span size are low in PMI ranking. Again, the effect size measures are influenced more by altering the method than the measure that takes frequency of

occurrence into account. All in all, this implies that researchers should carefully consider the size of the target context they would like to take into account in their collocational analysis.

2.4 Considerations for selecting Chinese corpora

For every corpus-based study, several decisions have to be made in order to find a corpus that suits the research question best. We will elaborate on three dimensions that may influence the selection of corpora.

2.4.1 Availability of segmentations and annotations

The aim of a collocational analysis is to examine the association between words/expressions. As mentioned before, the Chinese written system does not separate words with white spaces. That is why the availability of word segmentations and/or annotations is especially important. Although we have shown that applying a word segmentation tool to a ‘raw’ corpus will probably generate highly similar outcomes (if G^2 and Fisher-1 are used as ranking measures), annotated or word-segmented corpora are preferred for Chinese collocational studies. For surface collocational studies, corpora with word segmentations are sufficient. For syntactic collocational studies, annotations of the syntactic functions are also necessary.

Corpus size is another important criterion for performing an informative collocational analysis. Despite the advantages of segmented and/or well-annotated corpora, many annotated corpora are limited in corpus size. For reasons of statistical reliability, larger corpora are preferred for collocational studies. Some association measures are vulnerable to a low frequency of occurrence of items. For instance, the G^2 test always requires a minimal frequency of three for statistical stability, and p-values provided by a Chi-squared test become unreliable as soon as one of the expected values in the contingency table is less than five (Dunning, 1993). Besides, words within one document may be correlated. For instance in a document about traffic, the occurrence of the word *truck* in the text may not be entirely independent of *train*. To overcome such interdependencies, corpora with a large number of documents are preferred.

Some word segmentation tools in Chinese are available now, which make collocational studies on large-scale raw Chinese data possible. Such word segmentation tools usually take some dictionaries or POS tagging systems of the word categories as input, and provide the output of segmented and/or annotated texts (technical details have been discussed in 3.1). Choosing between a well-annotated/segmented but comparatively small corpus and a large-scale raw corpus which requires more segmentation work is a decision to be made by researchers. What

is more, the choice of corpora is also related to the method of identifying discourse segments, as discussed in Section 3.2. Compared to annotated corpora, analyzing raw corpora takes more time and effort.

2.4.2 Properties of corpora: genres, sub-corpora and time spans

Corpus linguists need to consider whether they need a specialized corpus or not. This is dependent on the nature of the study. In general, selecting a balanced corpus with different types of genres and/or modalities available seems to be a safe choice. However, if the research question addresses specific genres or modalities, it makes more sense to select a representative corpus or a sub-corpus from a balanced bigger corpus. A prior identification of the genre categories can guarantee a good representation at the genre level (Biber, 1993). If one wants to compare two or more genres or modalities, then either a balanced corpus with relevant sub-corpora or two or more specialized corpora is applicable.

Genres, channels, registers or other subparts of a corpus matter to a collocational study in at least two ways. First, researchers can add credence to the findings by checking whether collocation patterns are consistent across genres, channels, registers, or sub-corpora (Gries, 2013). Second, they allow researchers to formulate hypotheses about the effects of genres/channels/registers on the collocation patterns. In other words, genre, channel, or register type can be treated as an independent variable influencing the choice of words/word forms as the collocates of the target word (Gries & Stefanowitsch, 2004). For example, in their study of Chinese causal connectives, Li et al. (2013) selected sentences containing connectives from three different genres: news reports, opinion pieces, and novels. The first two genres were taken from *People's Daily Online*, and the narrative genre was from the CCL corpus. Li et al. (2013) hypothesized that genre has an impact on the degree of subjectivity of a text, and might therefore affect the meaning and use of Chinese causal connectives. For example, opinion pieces typically express the writer's point of view and aim to convince the reader by presenting arguments, and are therefore likely to display an overall higher degree of subjectivity than news reports, which are more descriptive and informative in nature. Results of this corpus-based analysis reveal that three connectives (i.e. *kejian* 'so', *yin'er* 'as a result', and *yushi* 'so/therefore') display robust profiles, whereas two other connectives (*suoyi* 'so/therefore' and *yinci* 'so/therefore') appear to be genre-sensitive (see Li, Sanders & Evers-Vermeul, 2016 for a similar analysis on *jiran* 'since', *yinwei* 'because' and *youyu* 'as'). In the domain of collocational studies, Stefanowitsch and Gries (2008) find some constructions to be more channel sensitive than others: the active construction exhibits sensitivity to differences related to spoken vs. written channels, while the passive construction consistently

shows construction-specific preferences for certain types of collocates regardless of the channel.

Depending on whether one wants to investigate the diachronic development or make a synchronic comparison between elements, the time span of a corpus may be relevant. Some studies on the diachronic development of causal connectives examined the process of grammaticalization and subjectification, using data from different historical periods (Jiang, 2010; Li, 2012; Li, 2009; Liu & Yao, 2011; Wang, 2002; Zhang, 2012). Wei and Evers-Vermeul (2014) adopted a diachronic approach to investigate the use of the Chinese temporal connectives *erhou* ‘and after’, *ranhou* ‘then’ and *yushi* ‘since then’, based on the data from different periods of time. They observed distinct degrees of lexicalization of the three connectives, predicting different possibilities of getting a causal interpretation from these temporal connectives. Collocational analysis on the use of connectives in different historical periods may provide an interesting extension for this type of study.

2.4.3 Accessibility of texts

Accessibility is another important criterion of corpus selection. Collocational studies always involve analytic tools for large scale investigation, such as *AntConc*, *WordSmith* and *R*. These tools are helpful for running queries, cleaning up datasets and computing association scores. However, they require that linguistic data are retrieved from corpora and converted into certain text formats that are readable as input for the software. Therefore, corpora with full texts available are preferred for convenience. With the full texts, association scores can be calculated based on the information about the total size of the corpus and the observed frequencies automatically counted by the analytic tools.

Some corpora only provide online searches without full texts available online. In that case, it is suggested to perform analysis using the *R* software with certain packages installed (for instance, *mclm* by Speelman, to appear). Those statistical packages can produce reliable results of association strengths based on estimated occurrences of words, given the total size of the corpus. Therefore, for corpora that only allow online searches, information on corpus size is necessary.

For corpora without full text provided, availability of downloading online searches is relevant for the choice of corpora. No limitation on the number of items to be downloaded would be favored for two reasons: (i) to have the ability to collect enough data; (ii) to avoid being biased by observations based on texts restricted to certain domains. In other words, some observations are there just because of a biased selection of texts. For example, if the first 100 texts are all from the same author, the effect found in these 100 texts cannot be generalized. If one cannot get access to the full download of all items, random searches are possible and preferred in order to avoid such a bias.

An overview of the available Modern Chinese corpora can be found in Appendix 2. This appendix also provides information on the possibilities and characteristics of each corpus, such as the possibility of random downloads, the availability of full texts and/or information on the corpus size.

Some modern corpora provide possibilities for on-site collocational analysis, such as the *Academia Sinica Balanced Corpus of Modern Chinese*, the *Lancaster Corpus of Mandarin Chinese* and the *UCLA Written Chinese Corpus*. With online inquiries on the latter corpus, for instance, top 30 collocates can be listed and ranked by the values of MI, MI3, Dice, t, z and log-likelihood. Additionally, the online corpus software interface *Word Sketch Engine* (Kilgarriff et al., 2014) also provides information on collocation behaviors of words, which can be widely applied to lexicography, language teaching, as well as linguistic research.

2.5 Conclusion

In conclusion, we advocate the study of collocates in the Chinese language, including collocates at the discourse level. Statistical measurements of the attraction or repulsion between words/expressions can be more stable and reliable than anecdotal examples to establish collocation relations. In the application of collocational analysis, a crucial starting point is to formulate research questions based on the parameters discussed in Section 2.1 (Gries, 2013), such as the nature of collocates (e.g. morphemes, words, or phrases), the distance (e.g. directly adjacent or not), and lexical and syntactic flexibility (e.g. word or lemma). Suitable methodologies can then be developed regarding the choice of corpora, segmentation tools and relevant target contexts.

This chapter has provided an overview of potential issues in the practice of performing a Chinese collocational analysis, as well as some possible solutions. We expect future discussions and practice to cover more theoretical and technical details relevant to the analysis of the Chinese language, as well as other languages.

Chapter 3. The use of perspective markers and connectives in expressing subjectivity: Evidence from collocational analyses

3.1 Introduction

When communicating, speakers choose words to express the relations between consecutive discourse segments (Sanders, Spooren & Noordman, 1993: 94; cf. also Sanders & Spooren, 2007; Schilperoord & Verhagen, 1998). For instance, they can use connectives such as *because* and *therefore* to mark causal relations. In expressing such discourse relations, people also continuously choose other linguistic elements to embed their opinions and feelings in these relations. When making those choices, speakers and writers decide how informative they should be in order to provide sufficient cues for others to comprehend them. Meanwhile, they should also avoid being too wordy. This tension has been systematically described by Horn's framework for pragmatic inference: his *Q* (Quality) *Principle* describes the need to 'make your contribution sufficient'; the *R* (Relation) *Principle* describes the need to 'make your contribution necessary' (Horn, 1984: 13). According to Horn (1984), speakers should find a balance between the speaker-based economy (saving the speaker's production efforts) and the hearer-based economy (saving the hearer's processing efforts).

A highly similar point has been made by the *Uniform Information Density Theory* (UID), which is about the speakers' strategy of choosing between alternative linguistic forms at several levels of linguistic representations: phonetic, syntactic, pragmatic, etc. (Frank & Jaeger, 2008; Jaeger, 2010; Levy & Jaeger, 2007). The UID suggests that speakers modulate their word choice according to the amount of information in the utterance: full linguistic forms are more often used at the point where the content conveyed by the form is unexpected in its context, i.e. the point with a low probability and a high information density (for details, see Frank & Jaeger, 2008). For instance, connectives can be omitted if the information they convey is highly predictable given other linguistic cues in the context (Asr & Demberg, 2015). Through such modulation of word choices, the density of information of the utterance is kept at a uniform level – a roughly equal amount of information at each unit of the sentence (Levy & Jaeger, 2007). The UID theory echoes Horn's pragmatic theory in the sense that both theories predict a modulated process of word selection to optimize communication. In terms of discourse relations and connectives, these theoretical discussions raise the question as to which information is exactly conveyed by connectives, and how that information may become predictable given other cues in the context.

Connectives and other coherence markers provide the reader with information

on the type of coherence relation to be established (Britton, 1994; Graesser & McNamara, 2011; Mak & Sanders, 2010; van Silfhout, Evers-Vermeul, Mak & Sanders, 2014; van Silfhout, Evers-Vermeul, & Sanders, 2015). Such information facilitates the reading process. For instance, it triggers faster processing of information immediately following the connective (Cain & Nash, 2011; Cozijn, Noordman & Vonk, 2011; Sanders & Noordman, 2000; van Silfhout et al., 2014, 2015) compared to the processing of that same information in unmarked relations.

The influence of connectives on the processing of discourse is not restricted to establishing the type of coherence relation. Certain connectives also provide information on the degree of subjectivity of the relation, which concerns the degree of involvement of a locutionary agent or a *Subject of Consciousness* (Finegan, 1995; Lyons, 1977; Sanders, Sanders & Sweetser, 2009). For example, causal relations can be of a subjective type or an objective type. Objective causal relations are causal links between events in the real world such as (1a), while subjective relations, illustrated in (1b), involve someone's reasoning (Langacker, 1990; Pander Maat & Sanders, 2000; Verhagen, 2005). Subjective relations are not observable in the real world – one needs to take into account another person's (e.g., the speaker's or another agent's) perspective (Sanders et al., 2009, 2012) to process the reasoning, and thus one needs to track the source of information.

- (1)
- a. This restaurant is decorated with several art works of Mondriaan, **so** it attracts lots of fans of Modern art.
 - b. This restaurant is decorated with several art works of Mondriaan, **so** its owner must be a fan of Modern art.

As examples (1a) and (1b) illustrate, English *so* can be used in objective and subjective causal relations. It only marks the causal nature of the relation, and does not indicate the degree of subjectivity of the relation. However, certain connectives in other languages do code information of subjectivity. For example, some connectives are only used for objective relations, such as Dutch *daardoor* 'as a result' and Chinese *yin'er* 'as a result', as is illustrated in the Dutch (2a) respectively Chinese (3a) translation of (1a). By contrast, the Dutch connectives *want* 'because' and *dus* 'so' (Degand & Pander Maat, 2003; Sanders & Spooren, 2015; Spooren, Sanders, Huiskes & Degand, 2010; Stukker & Sanders, 2008; Verhagen, 2005), and Mandarin Chinese *kejian* 'so' prototypically express subjective coherence relations (Li, Evers-Vermeul & Sanders, 2013). This is illustrated by the Dutch (2b) respectively Chinese (3b) counterparts of the subjective relation in (1b). Just like English *so* in example (1a) and (1b), some connectives leave the subjectivity information underspecified, i.e. they can be used for both subjective and objective relations (e.g. Chinese *suoyi* 'so' in example (3a) and (3b)).

- (2) Dutch
- a. Dit restaurant is versierd met diverse kunstwerken van Mondriaan, **daardoor** trekt het veel fans van moderne kunst.
 - b. Dit restaurant is versierd met diverse kunstwerken van Mondriaan, **dus** de eigenaar moet wel een fan zijn van moderne kunst.
- (3) Chinese
- a. Zhe jia canguan zhuangshi zhe hao ji fu Mengteli'an de huazuo, **yin'er/ suoyi** ta xiyin le henduo xiandai yishu mi.
This CL restaurant decorate ASP(IPFV) CL Mondrian MOD painting, **as a result/ so** 3SG attract ASP(PFV) many modern art fan.
 - b. Zhe jia canguan zhuangshi zhe hao ji fu Mengteli'an de huazuo, **kejian/ suoyi** ta de zhuren keneng shi yi ge xiandai yishu mi.
This CL restaurant decorate ASP(IPFV) CL Mondrian MOD painting, **in conclusion/ so** 3SG MOD owner must be a CL modern art fan.

The degree of subjectivity expressed by connectives is found to affect the processing of coherence relations. For instance, the Dutch subjective connective *want* 'because' leads to longer processing times directly after the connective compared to the Dutch objective connective *omdat* 'because' (Canestrelli, Mak, & Sanders, 2013). Such processing effects can be attributed to the difficulty of interpreting subjectivity: the reader needs to track the source of information to interpret subjectivity. Specific subjective connectives such as *want* 'because' instruct the reader at an early stage that there is a coherence relation, and that the relation is subjective, before the entire sentence is processed. In terms of the information density, subjective connectives encode more information compared to underspecified connectives.

Given Asr and Demberg's (2015) claim that the information conveyed by connectives can become highly predictable given other contextual cues, it is worthwhile exploring which linguistic markers also provide information on the degree of subjectivity of a relation. If other markers already indicate the degree of subjectivity, this will reduce the need of information on subjectivity to be expressed at the connective. This seems to be the case for expressions such as *probably*, *surprisingly* and *according to Peter*, which are addressed as markers of stance (Biber, Johansson, Leech, Conrad & Finegan, 1999; Conrad & Biber, 2000), evaluation markers (Bednarek, 2006, 2009; Thompson & Hunston, 2000), or appraisals (Eggins & Slade, 1997; Martin, 2000). Conrad and Biber (2000) suggest three sub-types of stance markers (see Bednarek 2006, 2009, and Thompson & Hunston 2000 for similar classifications):

- i. Epistemic stance, which indicates how certain the speaker or writer is, or where the information comes from (e.g. *probably, according to the President*).
- ii. Attitudinal stance, which indicates feelings or judgements about what is said or written (e.g. *surprisingly, unfortunately*).
- iii. Style stance, which indicates how something is said or written (e.g. *honestly, briefly*.)

(Conrad & Biber, 2000: 57)

Stance markers introduce the viewpoint of the speaker or other agents, and hence can be termed as *perspective markers* (Sanders & Redeker, 1996). Perspective markers expressing epistemic stance show overlap with specific subjective connectives. Both indicate subjective reasoning, either from the speaker or from a character. Canestrelli et al. (2013) and Traxler, Sanford, Aked, and Moxey (1997) found that the processing effects of connectives are influenced by epistemic stance markers: by adding *volgens Peter* ‘according to Peter’ to the first clause connected in a subjective relation, as in example (2c), the extra processing time associated with the subjective connective *want* ‘because’ disappears.

(2)

- c. **Volgens** Peter is de eigenaar van dit restaurant een fan van moderne kunst, **want** het restaurant is versierd met diverse kunstwerken van Mondriaan.
According to Peter the owner of this restaurant is a fan of Modern art, **because** the restaurant is decorated with several art works of Mondrian.

In terms of Horn’s pragmatic theory, the reader/hearer has obtained sufficient information about the degree of subjectivity by the introduction of epistemic perspective markers. Upon encountering the subjective connective the reader/hearer does not have to establish an entirely new subjective mental representation, but rather only has to make a link to an already established mental representation introduced by the perspective marker in the first clause. In other words, epistemic stance markers in the first clause make it clear that the first clause is a claim and thereby create the expectation that the next clause will be an argument for this claim.

The empirical findings of Canestrelli et al. (2013) and Traxler et al. (1997) suggest an overlap between specific subjective connectives and perspective markers in their function of instructing readers on the degree of subjectivity of the relation. The question is whether this holds true for perspective markers in general, including all types of stance markers, or only pertains to markers of epistemic stance. Epistemic stance markers explicate the dimension of reliability/certainty and evidentiality, which directly introduces a source of information. However, attitudinal stance markers and

style stance markers introduce a source in an indirect way: by indicating attitudes, feelings and styles of writing/speaking that can be attributed to a source. Although all three types of stance markers presuppose a source of information, they differ in the way in which this source of information is involved. How these perspective markers overlap with connectives marking different degrees of subjectivity may shed light on the relation between subjectivity and perspective marking.

In this paper, we investigate this issue in natural language data. Starting from the assumption that language users will tend to avoid a doubling of information in terms of marking subjectivity in discourse relations, we may expect authors/speakers to observe some pragmatic strategies (e.g. apply Horn's *R principle* or try to produce an information flow with a *Uniform Information Density*) to achieve a successful communication (both sufficient and necessary). Avoiding repetition of information in the same dimension fits the *R principle* as well as the UID. Therefore, in natural language data we may expect connectives marking different degrees of subjectivity to vary in their co-occurrence patterns with perspective markers.

Methods in corpus linguistics have provided various opportunities to investigate language use at a large scale. As Firth (1957: 11) has argued: "you shall know a word by the company it keeps." The method of collocational analysis (Evert, 2008; Gries & Stefanowitsch, 2004; see also Chapter 2) provides insightful information on the context of given linguistic elements. It measures the association strengths between words or expressions, and produces a list of important collocates in attraction or repulsion with a target word. Collocational analysis can advance our knowledge about the properties of a connective on the basis of its contextual features. We therefore conducted a corpus-based study using collocational analyses to examine the use of connectives and perspective markers in discourse, aiming to answer the following research questions:

- 1) Do connectives of different subjectivity degrees differ in their types of collocates?
- 2) More specifically, do connectives differ in the types of perspective markers they co-occur with?

We focused on two Chinese causal connectives. *Kejian* 'so' indicates that the causal reasoning arises from someone's mind, i.e. it encodes the epistemic stance apart from its discourse function of connecting two segments. Such subjectivity information is underspecified with the generic connective *suoyi* 'so', which can be used in both objective and subjective relations (Li, Evers-Vermeul & Sanders, 2013). On the basis of Horn's theory of speaker economy, it can be expected that *kejian* co-occurs less with perspective markers of the epistemic stance than *suoyi*. Since neither the specific subjective *kejian* nor the generic connective *suoyi* encode attitudinal or style stance, no differences in collocation tendencies are expected between the connectives for the

other two types of perspective markers.

3.2 Method

We conducted a series of distinctive collocates analyses on the two Chinese causal connectives *suoyi* ‘so’ and *kejian* ‘so’, with the aim to investigate the contextual features of the two connectives. Collocational analysis in general allows measures of associations between target words and their collocates. A distinctive-collocate analysis, as a specific type of collocational analysis, can distinguish words that tend to appear in the context of one word or structure compared to another word or structure (Gries & Stefanowitsch, 2004; Stefanowitsch & Gries, 2003). With this type of analysis, words with high association scores are not associated with the target word/structure in a general sense, but only associated with the target word/structure in the sense that they are attracted more by the context of the target word/structure compared to the reference contexts (i.e. the contexts of an alternative word/structure). In the current study, we identified words that tend to ‘sit’ in the context of *suoyi* more often than in the context of *kejian* and vice versa.

In the current study, we paid special attention to the linguistic elements expressing subjectivity, including all kinds of indications of the source of information: verbs, nouns, adjectives, adverbials, etc.

3.2.1 Sample of texts

We used a balanced modern Chinese corpus: the CCL corpus (*Center for Chinese Linguistics of Peking University*, web-version 2015), which covers a variety of written texts: fiction, newspapers, conferences, translated literature, blogs, etc. The total size of the CCL corpus is 581,794,456 characters.

To make sure all the texts were homogeneous in terms of mode (written), we excluded the sources of oral texts (spoken), TV (written to be spoken), dictionaries, etc. From the remainder of the corpus, we selected texts from three types of genres: narrative genres on the one hand, and informative and argumentative genres on the other. Narrative genres included literature, drama, biographies and fiction magazines; informative and argumentative genres included newspapers, legal documents, academic works of natural science and social science, governmental reports and other texts labeled as practical writing.

From the afore-mentioned parts of CCL, we then generated two raw datasets: text files containing all the sentences with the words *suoyi* or *kejian*, with a search scope of 200 characters to the left and 200 to the right. This scope was much wider than the length of a sentence so that we would have enough contexts for the analysis on the intended discourse unit.

In line with the parameters of collocation (Gries, 2013; Chapter 2), we first decided to investigate words as the linguistic units of collocates. However, natural Chinese texts do not have spaces between words. Therefore, we used the Chinese word segmentation tool *NLPIR-ICTCLAS* (2016, tag: *ICT_POS_MAP_SECOND*) to separate the word boundaries of characters in the text (see Chapter 2). In this segmentation system, white spaces were added between words, and words were tagged based on their semantic types. Meanwhile, punctuations such as commas, full stops, parentheses, colons were also marked with tags. The word segmentation tool generated segmented and annotated texts for later analysis.

In terms of the distance of collocates, a collocate did not need to be directly adjacent to the connective. Any words appearing within one clause before or one clause after the connectives were considered collocates. Instead of adapting an arbitrary number of words as the context, we set the context of the target word in such a way that it is meaningful at the discourse level: discourse clauses were taken as the units for the analysis.

3.2.2 Sample of connective fragments

From the two segmented datasets of all sentences containing *suoyi* or *kejian*, we compiled a sample of connective fragments. This step was necessary, because *suoyi* does not only occur as a connective, but can also be used in an inversion construction *zhisuoyi* ‘why there is a consequence of’. In addition, the word *kejian* is not fully lexicalized as a connective, but can still be used as a verb, sometimes resulting in modified constructions such as *qingxikejian* ‘clearly can see’. To exclude the majority of these verbal uses of *kejian* and all the inversion constructions *zhisuoyi*, we restricted the sample of target items to cases preceded by a punctuation marker (namely comma, full stop, semicolon, question mark, exclamation mark or ellipsis) in the software *AntConc_3.4.4.0* (Anthony, 2016). After the rough automatic screening process, 67,147 sentences with *suoyi* and 3,902 sentences with *kejian* were included for further analyses.

We then manually checked the remaining sentences marked by *kejian*, in order to exclude all other verbal instances of *kejian*. The verbal status of *kejian* could easily be derived from the absence of the main verb in the clause headed by *kejian*. For example, in (4), interpreting *kejian* as a connective with the meaning ‘so’ would only leave a noun phrase as the remainder of the second clause: *the status of German cars in the minds of Chinese*. By contrast, interpreting *kejian* as a verb ‘can see’, results in a grammatical clause, because in Chinese, the subject can be dropped. Hence, only full sentences such as (5) were included in the analyses of the connective use of *kejian*.⁶

⁶ These texts, and all other fragments in this paper are taken from the CCL corpus.

(4)

德国产的大众、奥迪和奔驰占有很大的比例，**可见**德国产的汽车在中国人心目中的地位。

Deguo chan de dazhong, Aodi and Benchi zhanyou hen da de bili, **kejian** deguo chan de qiche zai zhongguoren xinmuzhong de diwei.

Germany produce MOD Volkswagen, Audi and Benz occupy very big MOD proportion, **kejian** 'from this can see'/***kejian** 'so' Germany produce MOD car in Chinese mind MOD status.

The German products Volkswagen, Audi and Benz take a big proportion (of Chinese market), **from this we can see**/***so** the status of German cars in the mind of Chinese people.

(5)

一个能值得一头牛的价钱，**可见**那时候石榴在我国还是稀罕物。

Yi ge neng zhide yi tou niu de jiaqian, **kejian** nashihou shiliu zai woguo haishi xihan wu.

One CL can worth one CL cow MOD price, **kejian** 'so' that-time pomegranate in our-country still-is rare thing.

One (pomegranate) was worth the price of a cow, **so** pomegranate was still very rare in our country at that time (in Ancient China).

All in all, the automatic and manual screening process excluded 20,096 cases of *suoyi* and 10,900 cases of *kejian*. Table 1 shows the resulting distribution of *suoyi* and *kejian* in the narrative and non-narrative texts in the sample.

Table 1
Distributions of 'suoyi' and 'kejian' in two types of genres

Connectives	Narrative texts		Non-narrative texts	
	Retrieved from CCL	Used for analysis	Retrieved from CCL	Used for analyses
<i>suoyi</i>	34,641	29,077	52,445	37,913
<i>kejian</i>	2,494	752	11,688	2,530
Total	37,135	29,829	64,133	40,443

To obtain reliable results, enough data should be collected for collocational analyses. In this study, the number of cases used for analyses was sufficient to calculate association scores (as will be elaborated upon in Section 2.4).

3.2.3 Three sets of distinctive collocates analyses

The actual collocate analyses were conducted using the software *R* (R Core Team, 2015) with the R package *mclm_0.1* (Speelman, to appear). The method of distinctive-collocate analysis was applied three times. We first applied it to a context of one clause before and one clause after the connective in all types of genres. With this first analysis, we obtained a general picture of the words in collocation with one connective compared to the other. Second, we explored the collocates of the two connectives in their preceding context and following context separately, so that contextual features could be located more precisely.

However, the distinctive collocates of *suoyi* versus *kejian* may be different depending on the genre they appear in, because the narrative genre is supposed to have more descriptions of facts, events and actions, while the non-narrative genre contains more argumentations and demonstrations. Therefore, in the third analysis, we took genre into account, distinguishing the collocational patterns in the narrative genre on the one hand, and in the informative and argumentative genres on the other. The argumentative and informative texts were collapsed as the 'non-narrative genre', because of the low number of argumentative texts available in the CCL corpus. Low frequencies of word occurrences with limited data size may lead to statistically unreliable results.

3.2.4 Analysis of the association scores

The attraction and repulsion strength between a given word and the target connectives are measured by association scores (Evert, 2008; Gries, 2013). Such association scores were calculated for all the words in the corpus that appeared at least once in the context of *kejian* or *suoyi*. From the variety of association scores generated by the

mclm R package, the top 100 items ranked by G^2 were treated as important collocates in attraction/repulsion with *suoyi* compared to *kejian*. G^2 only reports association strength without any indications on the direction of associations – the top 100 items were either in attraction with the target word *suoyi* (i.e. in repulsion to *kejian*), or in attraction with the reference word *kejian* (i.e. in repulsion to *suoyi*). We need to rely on the *dir* (direction) values provided by the R package along with the association scores to judge whether a word was attracted to *suoyi* (positive) or repelled by *suoyi* (negative) in reference to *kejian*.

The Delta-P value was used as a secondary criterion for the collocates: the words in attraction to *suoyi* all needed to be above the threshold of 0, and the words in repulsion to *suoyi* (the collocates of *kejian*) needed to be below this threshold (<0). The Delta-P measure was applied because it is considered more psycholinguistically realistic, as it takes into account the directionality of the collocation: ‘whether the word₁ is more predictive of word₂ or the other way round’ (Gries, 2013: 141).

In order to interpret the outcomes, we clustered the collocates within the top 100, by searching for different types of linguistic items that could be related to subjectivity, particularly perspective markers. The results section will focus on these items; a full list of collocates per distinctive-collocate analysis can be found in Appendix 3.

3.3. Results

In this section, we discuss the results of the three distinctive collocates analyses. Section 3.1 illustrates the general collocation patterns of the two connectives. Section 3.2 compares the collocates of the two connectives in the clause preceding the connective and the clause following the connective. A genre-specific analysis shown in Section 3.3 reveals the collocations in different genres.

3.3.1 General collocational analysis

The top 100 of important collocates (either attracted by *suoyi* or attracted by *kejian*) were categorized according to their semantic types. Since our goal was to find out whether language users avoid overlap in the expression of subjectivity in their utterances, we checked the top 100 for linguistic elements that can be related to subjectivity and perspective marking. Table 2 shows the collocates of *suoyi* that are relevant to our discussion, with their observed and expected frequencies and the G^2 scores indicating the distinctiveness of particular collocates in the context of *suoyi* compared to the context of *kejian*. The observed frequencies in Table 2 are higher than expected, indicating that the words listed in this table occur more frequently in the context of *suoyi* than one would expect if the distributions in the contexts of *suoyi* and *kejian* were comparable.

Table 2
Important collocates of 'suoyi' from top 100

Collocates	Frequency (obs. vs exp.)	G²
<i>Pronouns</i>		
wo 'I/me'	22802: 21785	1189.91
ta 'she/her'	8972: 8602	380.85
ni 'you'(singular)	8177: 7840	345.64
ta 'he/him'	24340: 23734	318.83
women 'we/us'	9212: 8864	314.23
tamen 'they'	7183: 6908	253.13
ziji 'self'	6104: 5905	146.47
nimen 'you'(plural)	1129: 1080	53.62
<i>Communication verbs</i>		
shuo 'say'	12301: 12050	103.46
gaosu 'tell'	856: 818	43.12
<i>Cognition verbs</i>		
xiang 'think'	3990: 3829	159.58
zhidao 'know'	3333: 3199	131.28
renwei 'believe'	2794: 2699	74.07
xiwang 'hope'	1213: 1161	55.65
juede 'feel'	1576: 1516	55.04
pa 'be afraid of'	901: 863	39.06
<i>Modal verbs</i>		
hui 'would'	8270: 8030	152.17
neng 'can'	8883: 8714	64.20
keneng 'may'	2711: 2628	55.84
yinggai 'should'	1526: 1469	51.18
keyi 'can'	4025: 3933	43.46
bixu 'have to'	2123: 2059	42.70

From the top 100, certain types of words stood out as important collocates of *suoyi*, the connective that is underspecified in terms of subjectivity. An important cluster is

formed by pronouns of all types (singular and plural, 1st, 2nd and 3rd person). On the one hand, pronouns can be linked to objective relations in which actors carry out certain actions for certain reasons. On the other hand, they can be used in subjective relations in which the pronouns refer to the individuals whose perspective is presented. Therefore, we are not sure whether the higher number of occurrences of pronouns in the context of *suoyi* compared to the context of *kejian* should be attributed to a contextual feature of the objective relations that *suoyi* can express, or to the tendency to avoid doubling of subjectivity information in the context of *kejian*.

This is much clearer for the other clusters that are attracted by *suoyi*, but repulsed by *kejian*: communication verbs, cognition verbs and modal verbs. Both communication verbs and cognition verbs can express the epistemic stance of the speaker, to be specific, the evidentiality of the information. Modal verbs indicate the (un)certainly of the author/character towards the proposition, which is also one of the dimensions of epistemic stance.

This observation can be accounted for in terms of subjectivity. With *suoyi* in the sentence, the subjectivity information is underspecified. When subjectivity needs to be expressed, cognition verbs (marking evidentiality) and modal verbs (marking certainty) are used to help readers/hearers track the source of information. These would be repetitive for readers/hearers, however, in *kejian* contexts. *Kejian* already implies someone is making the inference (normally, the speaker), so the use of cognition verbs and modal verbs would be a repetition of information on subjectivity.

Communication verbs were not exclusively used to express the epistemic stance of the speaker as in example (6). They could also be used for reporting an objective description of real-world events such as in example (7).

(6)

它没有舞蹈一样美丽的动作，往往动作的开始就是打斗的结束，所以李小龙说，截拳道绝对不适宜表演。

Ta meiyou wudao yiyang meili de dongzuo, wangwang dongzuo de kaishi jius shi dadou de jieshu, *suoyi* Li Xiaolong *shuo*, Jiequandao juegui bu shiyi biaoyan.

It (Jiequandao 'Jeet Kune Do', a type of Chinese Kong Fu) NEG:have dance alike MOD motion, often motion MOD start just COP fight MOD end, **CONJ** NAME **said**, Jiequandao absolutely NEG suit performance.

It (Jeet Kune Do) does not have beautiful motions like a dance; the start of a motion is often the end of a fight, **so** Li Xiaolong **said** Jeet Kune Do is absolutely not suitable for performances.

(7)

庄静庵从怀里掏出一张签好字的单子，郑重地交给他**说**：阿诚，两家银行我昨天已经去谈过了，他们都还相信我庄某人的经济实力和做人的信用，**所以**他们已经同意马上解封你的工厂。

Zhuang Jing'an cong huaili tao chu yi zhang qian hao zi de danzi, zhengzhong di jiaogei ta **shuo**: Acheng, liang jia yinhang wo zuotian yijing qu tan guo le, tamen dou hai xiangxin wo Zhuangmouren de jingji shili he zuoren de xinyong, **suoyi** tamen yijing tongyi mashang jiefeng ni de gongchang.

NAME from pocket take out one CL sign good signature MOD sheet, seriously MOD hand 3SGM **say**: NAME, two CL bank I yesterday already go talk ASP(PFV) ASP(PFV), they all still believe me NAME POSS economical capability and behave MOD credibility, **CONJ** they already agree immediately unban 2SG MOD factory.

Zhuang Jing'an took out a sheet with a signature from his pocket, seriously handed it to him and **said**: Acheng, I went to the two bank yesterday and talked with them. They still believed in my economical capability and personal credibility, **so** they already agreed to immediately unban your factory.

Therefore, we cannot be sure of the reason for the collocation of communication verbs and *suoyi*. This collocation pattern could be due to the speaker/author's strategy to avoid repetition of subjectivity information in subjective relations, just as for the cases with cognition verbs. Alternatively, communication verbs could be a feature of the context typical of the objective relations expressed by *suoyi*.

Some of the words in the top 100 list were repelled by *suoyi* and should therefore be seen as distinctive for *kejian* instead of *suoyi*, as illustrated in Table 3. As mentioned in Section 2, we included all kinds of indications of subjectivity, irrespective of their grammatical categories. The noun *jiazhi* 'value' was treated equally as the adjective *zhongyao* 'important', because *jiazhi* 'value' is often associated with evaluations that are made from a person's perspective. The observed frequencies in Table 3 are lower than the expected frequencies, indicating that the words listed in this table occur less frequently in the context of *suoyi* (and hence more frequently in the context of *kejian*) than one would expect if the distributions in the contexts of *suoyi* and *kejian* were comparable.

Table 3.
Important collocates of 'kejian' from top 100

Collocates	Frequency (obs. vs exp.)	G ²
Exclamatory adverbials		
duome 'so much'	147: 235	263.63
hedeng 'how much'	33: 75	164.21
Expressions of importance		
jiazhi 'value'	768: 843	84.06
zhongyao 'important'	1423: 1489	42.90
Expressions of expectation		
jing 'surprisingly'	245: 272	33.95

The exclamatory adverbials, expressions of expectation and expressions of importance can be related to subjectivity: they indicated that someone's feeling or evaluation is involved, and that the hearer/reader is not merely dealing with a description of real-world facts. These collocational patterns indicate that language users do not necessarily avoid a doubling of information, as both *kejian* and these collocates express *that* subjectivity is involved. However, from this list of collocates of *kejian*, it can also be derived that language users do pay attention to the type of subjectivity information, in other words *how* the perspective of a speaker/character is involved. While the important collocates of *suoyi* (cognition verbs, communication verbs and modal verbs) could be related to epistemic stance marking, the important collocates of *kejian* – expectation markers and importance markers – can be related to attitudinal stance marking. Hence, there is no doubling of epistemic stance marking information, the crucial type of subjectivity expressed by the connective *kejian*.

3.3.2 Collocational analysis on different clauses

Given the general information on the contextual features in the analysis across clauses and genres, we obtained a basic understanding of the types of collocates that appear in the context of *kejian* and *suoyi*. However, we do not know from the overall analysis where these collocates appeared exactly – do they appear in the clause preceding the connective, or do they appear in the clause following the connective? By precisely identifying the locations of different types of collocates, we can be more informed on how language users combine different linguistic cues to express subjectivity in discourse. Moreover, for further psycholinguistic experiments, collocation distributions by clause provide insights into how linguistic stimuli should be designed to closely reflect authentic linguistic data. The current section therefore elaborates on the distribution of collocates in different clauses.

Table 4
Important collocates of 'suoyi' in preceding and following clauses

Collocates	Preceding clause		Following clauses	
	Frequency (obs. vs exp.)	G ²	Frequency (obs. vs exp.)	G ²
<i>Pronouns</i>				
wo 'I/me'	12056: 11547	497.78	10777: 10277	740.49
ta 'she/her'	5023: 4816	193.22	3959: 3796	192.48
ni 'you'(singular)	4056: 3878	182.67	4128: 3974	155.00
ta 'he/him'	13497: 13154	171.33	10870: 10609	145.81
women 'we/us'	4721: 4534	166.11	4505: 4349	142.19
tamen 'they'	3826: 3685	113.05	3372: 3240	144.12
ziji 'self'	3501: 3382	86.09	2617: 2537	61.23
nimen 'you'(plural)	--	--	601: 575	32.74
<i>Communication verbs</i>				
shuo 'say'	--	--	4720: 4549	166.63
jiao 'call'	--	--	866: 824	65.78
ting 'listen'	--	--	542: 516	38.84
chengwei 'be stated as'	--	--	442: 421	33.77
wen 'ask'	--	--	413: 393	30.92
<i>Cognition verbs</i>				
mingbai 'understand'	372: 353	24.87	--	--
zhidao 'know'	2452: 2342	117.21	--	--
xiang 'think'	2168: 2072	98.91	1825: 1761	59.68
renwei 'believe'	1898: 1833	46.96	903: 870	32.49
pa 'be afraid of'	666: 636	32.87	--	--
liaojie 'understand'	560: 533	31.11	--	--
juede 'feel'	903: 870	26.87	675: 648	29.90
xiwang 'hope'	--	--	713: 680	46.61
gan 'dare'	--	--	521: 497	32.69
<i>Modal verbs</i>				
hui 'would'	4535: 4404	76.03	3747: 3639	76.24
keneng 'may'	1562: 1507	40.97	--	--
yinggai 'should'	501: 478	25.46	--	--
yiding 'must'	898: 858	41.32	--	--
xiande 'seem'	--	--	273: 259	29.38
neng 'can'	--	--	4621: 4513	57.71
bixu 'have to'	--	--	1452: 1410	29.55

* The dashes in the table indicate that the words are NOT a distinctive collocate in the relevant corpus (same for Table 5, 6 and 7).

Table 4 and Table 5 show the collocates of *suoyi* and *kejian* we derived from the top 100 in preceding clauses and in following clauses.

Most of the general collocation patterns also held in the analysis per clause except for the communication verbs. In both preceding and following clauses, pronouns, cognition verbs, modal verbs co-occurred with *suoyi*. Most of these perspective markers may serve as the supplement of subjectivity information supplied by *suoyi*, regardless of whether they appear before or after the connective. Sentence (8) and (9) show examples of the combined use of *suoyi* ‘so’ and the perspective marker *renwei* ‘believe’, which can appear in both the clause before and after the connective.

(8)

人生难得有痛痛快快享受的日子，你错过了，将来老了时，想享受都没有能力，眼力不够，没有牙齿，听觉又不好，你想去享受一下，也力不从心，所以我**认为**应该趁年轻的时候，及时行乐。

Rensheng nande you tongtongkuai kuai xiangshou de rizi, ni cuoguo le, jianglai lao le shi, xiang xiangshou dou meiyou nengli, yanli bugou, meiyou yachi, tingjue you buhao, ni xiang qu xiangshou yixia, ye libucongxin, **suoyi** wo **renwei** yinggai chen nianqing de shihou, jishi xingle.

Life difficult have joyful enjoy MOD day, you miss ASP(PFV), future old ASP(PFV) time, want enjoy even NEG:have ability, eye NEG:enough, NEG:have tooth, hearing also NEG:good, you want go enjoy a:bit, also incapable, **CONJ I believe** should when young MOD time, in:time enjoy:life. It's difficult to have joyful days to enjoy. If you miss them, you cannot enjoy them anymore when you get old: with failing eyesight, few teeth, and defective hearing, it is impossible to enjoy even for a little bit, **so I believe** (we) should enjoy life at youth.

(9)

古典学派的学者们**认为**，哲学上的真理都已被亚里士多德和柏拉图所掌握，而神学上的真理都被《圣经》和奥古斯丁所掌握，**所以**要学得真理，只要读这些圣贤的书就够了。

Gudian xuepai de xuezhemeng **renwei**, zhexue shang de zhenli dou yi bei Yalishiduode he Bolatu suo zhangwo, er shenxue shang de zhenli dou bei Shengjing he Aogusidu suo zhangwo, **suoyi**, yao xuede zhenli, zhiyao du zhaxie shengxian de shu jiu gou le.

Classical school MOD scholars **believe**, philosophy on MOD truth all already PASS Aristotle and Plato which master, and theology on MOD truth

all already PASS Bible and Agustin which master, **CONJ** want learn truth, only:need read these virtuous MOD book just enough ASP(PFV).

Scholars of the Classical School **believed** all truths in philosophy had been mastered by Aristotle and Plato, and all truths in theology had been mastered by the Bible and Agustin, **so** reading these books of the virtuous people was enough to learn truths.

An important difference with the general collocation pattern is that the communication verbs appeared as important collocates of *suoyi* only in the clauses following this connective. This means that for the co-occurrence with such reportative verbs, no significant difference between *suoyi* and *kejian* can be found in the first clause.

A clear-cut difference between the collocates of *kejian* in preceding and following clauses is suggested in Table 5.

Table 5

Important collocates of 'kejian' in preceding and following clauses

Collocates	Preceding clause		Following clauses	
	Frequency (obs. vs exp.)	G ²	Frequency (obs. vs exp.)	G ²
Exclamatory adverbials				
duome 'so much'	--	--	67: 153	343.15
hedeng 'how much'	--	--	16: 60	207.58
xiangdang 'considerably'	--	--	269: 303	49.06
Expressions of expectation				
jing 'surprisingly'	110: 143	69.67	--	--
juran 'unexpectedly'	44: 59	34.71	--	--
jingran 'surprisingly'	30: 41	27.07	--	--
Expressions of importance				
zhongyao 'important'	--	--	757: 841	109.15
juzuqingzhong 'crucial'	--	--	3: 9	29.39
jiazhi 'value'	432:463	26,61	336:380	64.20
Communication verbs				
cheng 'state'	231: 265	47.84	--	--
yue 'say' (formal)	42: 57	35.93	--	--
yan 'speak'(formal)	165: 185	25.00	--	--

Exclamatory adverbials and expressions of importance were only distinctive for *kejian* in the clauses following this connective. This finding may be due to the tendency to express an evaluation in the second clause in a forward causal relation: the evaluation of importance is expressed in the second clause based on the

events/phenomenon described in the first clause, as is illustrated in (10).

(10)

怀孕后焦虑不安的母亲更容易难产和生出异常的孩子，可见孕期中注意心理卫生是**多么重要**。

Huaiyun hou jiaolü bu'an de muqin geng rongyi nanchan he shengchu yichang de haizi, **kejian** yunqi zhong zhuyi xinli weisheng shi **duome zhongyao**.

Pregnant after anxious disturbed MOD mother more easy dystocia and deliver abnormal MOD infant, **CONJ** pregnancy middle pay:attention:to mental health is **so:much important**.

Mothers who are anxious and disturbed after pregnancy are more likely to suffer dystocia and deliver abnormal infants, **so** paying attention to mental health is **very important** during pregnancy.

Expressions of expectation only appeared as important collocates of *kejian* in the preceding clause. These linguistic elements express an attitude of the speaker towards the situation described in the first clause, such as in example (11): the author is surprised by the fact that Wang Jian, a general, won the battles both in the south and in the north.

(11)

王翦**竟然**能在南北两方的作战中都取胜，可见其在用兵方面应当是属于全方位的武将。

Wang Jian **jingran** neng zai nanbei liang feng de zuozhan zhong dou qusheng, **kejian** qi zai yongbing fangmian yingdang shi shuyu quanfangwei de wujiang.

Wang Jian (a general in Chinese history) **surprisingly** can at south:north two side MOD battle in all win, **kejian** 3SGM at military aspect should COP belong:to extensive MOD general.

Surprisingly Wang Jian won the battles both in the south and in the north, **so** he should be a general with extensive military capabilities.

In contrast to the general collocation pattern, some instances of communication verbs were found as important collocates of *kejian* instead of *suoyi* in the preceding clause. Most of them are formal expressions, which are more characteristic of formal contexts such as informative and argumentative texts. Compared to the findings in Table 4 and 5, communication verbs can be collocates of either *suoyi* or *kejian*, depending on the formalities encoded in different specific communication verbs. Formal communication verbs such as *cheng* 'state' and *yue* 'say' patterned with *kejian*, while

informal communication verbs such as *shuo* ‘say’ patterned with *suoyi*. Therefore, it is not possible to identify a uniform pattern in the co-occurrence of communication verbs in relation to the degree of subjectivity expressed by the connective.

3.3.3 Collocational analysis on different genres

The results discussed so far may be the result of a confound with the genre preference of the connectives under investigation. *Suoyi* is a generic connective that can be used for all types of genres, while *kejian* is not frequent in narrative texts (cf. Table 1). Moreover, several of the collocate clusters found in Section 3.1 and 3.2 may be a side-effect of genre preferences as well. For example, communication verbs can be expected to appear more in the narrative genre, just like pronouns. Therefore, communication verbs and pronouns may pattern with *suoyi* simply because they all share the preference for the narrative genre. To neutralize the influence of genre as a confounding factor, we further examined the collocation of the two connectives in different genres, namely narratives and non-narratives. The collocation distributions of these two connectives with other linguistic elements in different genres are summarized in Tables 6 and Table 7.

Pronouns were observed as important collocates of *suoyi* in both types of genres, which indicated that this collocation pattern is not a side-effect of the genre preference of *suoyi*. Cognition verbs also appeared as important collocates of *suoyi* in both types of genres. Although the exact collocates differ per genre, they all expressed the same cognitive state of knowing and thinking. In addition, modal verbs were still distinctive collocates for *suoyi* in both narratives and non-narratives. Therefore, we may infer that the collocation of the generic connective *suoyi* with cognition verbs and modal verbs is not due to genre differences.

We did find a difference with the general collocation pattern, however, opposite to the hypothesis: communication verbs were no longer important collocates of *suoyi* in narratives, although they were still important collocates in non-narratives. Apparently, *suoyi* and *kejian* do not differ in their preference for co-occurring with communication verbs in the narrative genre, but only in the non-narrative genre.

Table 6.
Important collocates of 'suoyi' in different genres

Collocates	Narratives		Non-narratives	
	Frequency (obs. vs exp.)	G ²	Frequency (obs. vs exp.)	G ²
<i>Pronouns</i>				
wo 'I/me'	16329: 16043	241.62	6473: 6104	407.51
women 'we/us'	3503: 3434	69.49	5709: 5424	257.24
tamen 'they'	3744: 3673	67.21	3439: 3279	129.18
ta 'she/her'	7614: 7514	58.08	1358: 1290	62.14
ta 'he/him'	16328: 16196	43.53	8012: 7868	37.47
ni 'you'(singular)	6439: 6368	33.57	1738: 1632	129.07
ziji 'self'	3202: 3155	31.68	2902: 2789	72.55
ta 'it'	--	--	4099: 3979	54.42
nimen 'you'(plural)	--	--	339: 316	34.80
<i>Communication verbs</i>				
shuo 'say'	--	--	6623: 6433	83.64
chengwei 'be stated as'	--	--	444: 418	28.38
gaosu 'tell'	--	--	313: 293	27.29
<i>Cognition verbs</i>				
xiang 'think'	2776: 2733	30.99	1214: 1156	49.56
zhidao 'know'	2504: 2463	30.56	829: 794	24.15
renwei 'believe'	--	--	1928: 1837	76.42
xiwang 'hope'	--	--	713: 673	44.12
juede 'feel'	--	--	497: 470	26.93
<i>Modal verbs</i>				
hui 'would'	4043: 3990	30.82	4227: 4078	84.80
neng 'can'	3258: 3223	16.11	5625: 5477	58.89
bixu 'have to'	545: 533	15.05	1578: 1511	47.52
zhineng 'can only'	240: 233	13.56	--	--
keneng 'may'	876: 860	13.46	1835: 1758	54.43
yinggai 'should'	--	--	898: 853	41.34
keyi 'can'	--	--	2492: 2415	37.27

Even though the size of effects (observed frequency vs expected frequency) in the results of different analyses differ, the top 100 items still reflect important tendencies of collocation patterns in different genres. As Table 7 indicates, exclamatory adverbials and expressions of expectations remained distinctive for *kejian* in both narratives and non-narratives, although there were some differences per item. These perspective markers on the attitudinal stance dimension of expectedness were

associated more with *kejian* than with *suoyi* across genres. Expressions of importance only appeared as important collocates of *kejian* in non-narrative genres.

Table 7
Important collocates of 'kejian' in different genres

Collocates	Narratives		Non-narratives	
	Frequency (obs. vs exp.)	G ²	Frequency (obs. vs exp.)	G ²
Exclamatory adverbials	--	--	--	--
duome 'so much'	102: 124	62.78	45: 112	238.76
hedeng 'how much'	--	--	15: 53	152.93
Expressions of expectation				
jurán 'unexpectedly'	68: 79	25.59	--	--
jìng 'surprisingly'	140: 151	16.78	105: 124	25.97
guórán 'as expected'	28: 34	16.69	--	--
Expressions of importance				
jiázhì 'value'	--	--	660: 723	57.04

3.4 General discussion and conclusion

The collocational analyses revealed an interpretable pattern in connective use in combination with perspective markers in discourse. With this type of analysis, only relative distinctions can be made between the collocates in the context of *suoyi* and those in the context of *kejian* – the collocates of *suoyi* may also co-occur with *kejian*; they just have a preference for *suoyi* in comparison to *kejian*. Nevertheless, such tendencies of co-occurrences between connectives and perspective markers provide insights into the meaning and use of connectives expressing different degrees of subjectivity. In line with our predictions, the degrees of subjectivity encoded in the two connectives was related to the type of linguistic cues in their contexts.

3.4.1 General collocation patterns in line with pragmatic principles and UID

In general, the underspecified connective *suoyi* 'so', which can express both subjective and objective relations, patterned with more occurrences of cognition verbs and modal verbs in comparison to the specific subjective connective *kejian* 'so'. In the context of *kejian*, we found more exclamatory adverbials, expressions of importance and expressions of expectation compared to the context of *suoyi* as a reference level.

The collocation results showed that perspective markers as a general type of

linguistic cues marking subjectivity can be used in combination with either of the two causal connectives. However, if perspective markers are specifically categorized into sub-types with regards to various dimensions of subjectivity, different collocation patterns surfaced. *Suoyi* turned out to collocate with epistemic stance markers more often, while *kejian* co-occurred with attitudinal stance markers.

The collocation pattern of epistemic stance markers and *suoyi* agrees with Horn's pragmatic theory of *Relation principle* (reducing the speaker's production effort) and *Quality principle* (reducing the hearer's comprehension effort). From the perspective of the *R principle*, if subjectivity information on the epistemic stance is already specified in the connective *kejian*, epistemic stance markers in the context of the connective are redundant, i.e. not efficient from the speaker economy account. *Suoyi*, by contrast, does not provide sufficient information on the epistemic stance, and the use of epistemic stance markers therefore provides valuable information that compensates the lack of subjectivity information in *suoyi*. The *Q principle* is observed and hearers/readers' comprehension process should be facilitated.

The collocation results can also be well explained by the *Uniform Information Density Theory* account. With the two alternative connectives expressing discourse coherence, the presence of epistemic stance markers (e.g. cognition verbs, modal verbs) makes the content of the context highly predictable (high probability and low information), which is why it is more likely to have an underspecified connective, *suoyi* in this case. On the other hand, utterances with fewer occurrences of epistemic stance markers make the content conveyed by the context unexpected (low probability and high information). In this sense, the use of a specific connective is preferred. The prevalence of epistemic stance markers in the context of *suoyi* and their lower co-occurrence with *kejian* fit the need for a uniform information density throughout the sentence in terms of subjectivity. Optimal information density is realized in this way.

However, speakers/authors did not avoid overlap in the expression of subjectivity at all costs. Some attitudinal stance markers such as *jingran* 'surprisingly' and *zhongyao* 'important', which also indicate the involvement of a speaker responsible for an evaluation, occurred as important collocates in the context of *kejian*. Both epistemic stance markers and attitudinal stance markers express *that* a source of information is involved. Apparently, in their use of *kejian*, which also indicates *that* a source of information is involved, speakers and writers do not avoid overlap with that same information provided by attitudinal stance markers. However, *kejian* does not overlap with attitudinal stance markers in the subjectivity dimension it expresses, i.e. in expressing *how* the source of information is involved. The fact that *kejian* patterns with attitudinal but not with epistemic stance markers indicates that language users try to avoid overlap in the expression of these dimensions. The connective *kejian* and epistemic stance markers both indicate how certain the speaker/writer is about the information, while attitudinal stance markers express the attitude or feelings of a person towards the information. In terms of UID, the combination of attitudinal stance

markers and *kejian* does not create high information density in the utterance. Given the two reasons above, it is explainable that attitudinal stance markers and *kejian* were found in collocation, showing a kind of agreement of subjectivity at the discourse level – jointly contributing to a subjective context.

3.4.2 Collocation patterns in different genres and clauses

To test for potential genre influences on the results, we performed collocational analyses on narratives and non-narratives separately. These analyses showed that communication verbs patterned with *suoyi* in the non-narrative genre, but not in narratives. This asymmetry is probably due to a different usage of communication verbs in these genres. As illustrated in section 3.1, communication verbs can be used to express an epistemic stance, as in example (6), or as a reportative verb to introduce a description of real-world events, as in example (7). In the non-narrative genre, we expect a higher frequency of epistemic communication verbs. The fact that communication verbs stood out as important collocates of *suoyi* in this genre is in line with the avoidance of doubling of information: *suoyi* has more needs of epistemic markers to strengthen the epistemic nature of the utterance than *kejian*, which encodes such information by itself. In narratives, with their abundance of descriptions of real-world events, however, we would expect a higher number of reportative communication verbs, which do not create a doubling of information with the information provided by *kejian* when they are used to report objective events in one of the clauses connected by *kejian*. This might explain why communication verbs do not stand out as collocates of *suoyi* in narratives, an explanation that needs to be corroborated in future corpus research on the actual usage of different types of communication verbs in narrative and non-narrative genres.

Apart from communication verbs, all other types of collocates of *suoyi* in the general analysis still surfaced as important in both narrative and non-narrative genres. Although the individual collocates of each cluster slightly vary per genre, the collocation between cognition verbs, modal verbs and pronouns with *suoyi* was robust across genres. As for *kejian*, exclamatory adverbials and expressions of expectation also appeared in both narratives and non-narratives as important collocates, which suggests that the perspective markers related to expectations were indeed an important contextual feature of *kejian*.

In order to locate the positions of each type of collocates in causal relations, we analyzed the preceding clauses and the following clauses of connectives separately. Most of the perspective markers as collocates of *suoyi* appeared in both the clauses preceding the connective and the clauses following it, except for communication verbs. The collocates of *kejian*, however, differed in the position they appeared in contexts. Expressions of expectation appeared as important collocates of *kejian* in the preceding clause, which makes sense because in a subjective relation with an argument-claim

structure, expressions of expectation such as *jingran* ‘surprisingly’ in example (11) mark the speaker’s surprisal – either about the propositional content of the clause preceding the connective, or about the fact as an argument-claim relation as a whole. Both exclamatory adverbials and expressions of importance tended to appear with *kejian* in the clause following the connective. These perspective markers served as expressions of the speaker’s attitude towards the claim presented in the second segment of the relation.

Communication verbs exhibited very different collocation patterns depending on the clause they occurred in. In the preceding clause, formal communication verbs did not surface as important collocates of *suoyi*, but rather patterned with *kejian* more often (example (11)). In the following clause, the tendency was reversed – communication verbs only surfaced as collocates of *suoyi* such as in example (6). As example (11) shows, the communication verb *yue* ‘say’ co-occurred with *kejian* mainly in very formal texts, in which *kejian* was more often found. Therefore, such collocation pattern could be attributed to an effect of formality. On the basis of the current explorative study, we cannot draw a decisive conclusion on this issue. Further studies on the use of communication verbs in different contexts are needed, especially to find out whether our ideas about the formality and about the objective versus the subjective use of communication verbs can be corroborated.

(12)

(临济祖师) 答曰：如果一口气不来，这肉体还有情感吗？可见情感不在肉体上，而在灵性上。

(Master Linji, a Buddhism master) da **yue**: ruguo yi kou qi bu lai, zhe routi haiyou ganqing ma? **Kejian** qinggan buzai routi shang, er zai lingxing shang.
(Master Linji) answer **say**: if one CL breath NEG come, this body have emotion? **CONJ** emotion NEG at body, but at spirituality on.

(Master Linji) **said** in response: if one doesn’t breath anymore, does the body still have emotions? **So** emotion is not in the body, but rather in the spirit.

3.4.3 Future studies and conclusion

In conclusion, it is important to stress three limitations of this study. First, the sentences in causal relations were retrieved directly from the corpus without any manual annotations of the relation type, which would take more time and effort. *Kejian* is mainly used for subjective relations, while *suoyi* is generic (Li et al., 2013). This means that the contexts of *suoyi* contained both subjective and objective relations, while the contexts of *kejian* mainly consisted of subjective relations. The unbalanced distribution of relations in the contexts of the two connectives may be a confounding

factor. For instance, the fact that pronouns were distinctive for *suoyi* may be a feature of objective relations, because the descriptions of events and acts in objective relations may involve the use of pronouns. Nonetheless, the major findings such as the fact that modal verbs and cognition verbs are important collocates of *suoyi* are not characteristic of objective relations at all. We would expect stronger distinctive collocation patterns of these expressions with *suoyi* if we had limited the scope of the investigation to subjective relations only. More fine-grained analyses are expected to shed a clearer light on this issue.

Second, the collocational analysis only provides indications on the rough tendency of word use in the context of a target word. It cannot support any decisive inferences, such as the predictability of one word given the other word. To further investigate the relation between connectives and their collocates in the contexts, one could refer to regression analyses to look at whether the presence of certain words in the context correlates with more occurrences of a connective, or to experimental research to investigate the effects of perspective markers on the processing of connectives.

The third limitation is that we made no distinction between argumentative and informative genres due to practical reasons. These two genres have certain features in common in which they differ from narratives. For instance, both argumentative and informative genres have the author as the illocutionary force in most of the cases, while in narrative texts other characters are also frequently involved as the illocutionary force. However, argumentative genres also differ from informative genres in several respects – the use of communication verbs, for instance, may be different between argumentative texts and informative texts. Separate analyses of the argumentative genres and the informative genres can provide a more refined picture.

Despite these limitations, the collocational study, as an explorative approach, produced informative and valuable outcomes concerning the way subjectivity is expressed in discourse. What is more, this study illustrated the relation between connectives and perspective markers, which is useful for further psycholinguistic research.

Chapter 4. Causal connectives as indicators of source information: Evidence from the visual world paradigm

4.1 Introduction

The processing of discourse involves constructing mental representations of the input (Graesser, Millis, & Zwaan, 1997; Zwaan & Rapp, 2006). These mental representations of characters, actions, events, states etc. in discourse are known as situation models (Bower, 1989; Glenberg, Meyer, & Lindem, 1987; Johnson-Laird, 1983; Kintsch, 1988; Morrow, Greenspan, & Bower, 1987; van Dijk & Kintsch, 1983; Zwaan, Magliano, & Graesser, 1995; Zwaan & Radvansky, 1998). In a situation model, we keep track of all kinds of information of a story, such as temporal/spatial links among entities and causal relations between events, and also of the source of that information (Graesser, Bowers, Olde, & Pomeroy, 1999; Graesser et al., 1997; Zwaan et al., 1995; Zwaan & Rapp, 2006). The source of information overlaps with the subject/characters/agents who “are capable of speaking, perceiving and knowing” in discourse (Graesser et al., 1997: 172). In linguistic theories, the degree of involvement of such a locutionary agent or someone’s intentional mind is termed “subjectivity” (Finegan, 1995; Lyons, 1977; Sanders, Sanders, & Sweetser, 2009). In this paper, we investigate the construction of subjectivity in mental representations, specifically in the context of causal relations.

Information in a text can vary in its degree of subjectivity. It can be a description of a situation in the real world, as in (1a). In that case, we speak of an objective utterance. Alternatively, information can be presented as an opinion or belief presented from the point-of-view of an author or speaker, as in (1b). In that case, the utterance is subjective. In the situation model constructed for subjective cases, the speaker is represented in the mental representation of the addressee.

- (1)
- a. There is a tree in the garden.
 - b. I think there is a tree in the garden.

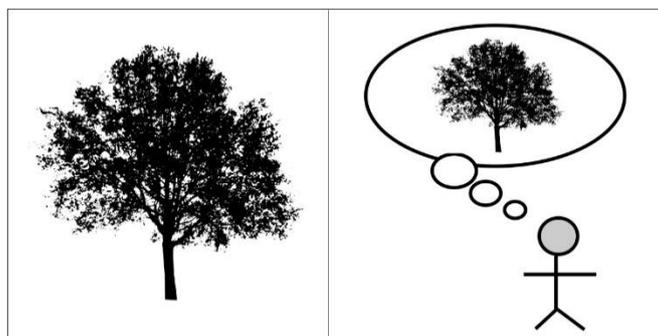


Figure 1. Representations of objective information and subjective information by addressees.⁷

The left panel in Figure 1 represents the situation model that is created when objective information is processed; in (1a), the statement that “there is a tree in the garden” is presented as rooted in the objective world. The picture on the right represents the mental state that results from subjective information. For the example in (1b), the speaker presents the idea or claim that there is a tree in the garden, and the addressee will process the sentence with a source of information involved. During language processing, one of these situation models is created. The situation on the right is expected to require longer processing time, because the discourse model that is constructed is more complex.

Similarly to individual clauses, relations between clauses can be categorized as subjective or objective, too. For instance, causal relations between segments, which is one type of information represented in the situation model, can also be categorized as subjective or objective. The person who is responsible for the reasoning is often termed the *Subject of Consciousness* (SoC) (Pander Maat & Sanders, 2001; Sanders et al., 2009; Sanders & Spooren, 2015). A conceptual distinction is drawn between “causes that are crucially located in a *Subject of Consciousness* and those that are located in the inanimate, outside world” (Pander Maat & Sanders, 2001: 251; cf., Lyons, 1977; Sanders et al., 2009; Verhagen, 2005). The former type of relation is

⁷ The image of the tree was created by Chrisdesign (acquired from OCAL Website) with the title: *Illustration of a tree silhouette*. The file is licensed under a free license (public domain). Link to the source: <http://www.freestockphotos.biz/stockphoto/15118>.

The image of the speech bubble was created by MithrandirMage with the title: *Human stick figure with thought bubble*. The file is licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license. Link to the source: https://commons.wikimedia.org/wiki/File:Thought_bubble.svg.

termed *subjective relation* (example 2b), and the latter one is termed *objective relation* (example 2a).

- (2)
- a. The factory has been polluting the water, *so* the local water supply is contaminated.
 - b. The factory has been polluting the water, *so* it has a very irresponsible owner.

In the situation model built for subjective relations such as (2b), an SoC is responsible for the reasoning. In the case of (2b), the SoC is the speaker. In the objective relation (2a), however, the speaker is not responsible for the relation. Therefore, a higher degree of subjectivity is expressed in (2b) compared to (2a). The degree of subjectivity of the relation can be encoded by linguistic cues that function as processing instructions on how to construct the situation model (Gernsbacher, 1990; Givón, 1992; Kintsch, 1992; Zwaan & Radvansky, 1998; Zwaan & Rapp, 2006). These linguistic cues can be explicit references to the SoC, or the remaining content of the utterance. For instance, in (3a), readers know from the linguistic cue *I think* the relation should be constructed as subjective. However, in (3b), the degree of subjectivity of the relation is not clear until a later point in the second clause, probably at the modal verb *may*. In such cases, readers rely on the content of the second clause to construct a subjective mental representation for (3b).

- (3)
- a. The intercity express is delayed for more than one hour, so *I think* the railway system has encountered a problem.
 - b. The intercity express is delayed for more than one hour, so the railway system *may* have encountered a problem.
 - c. The intercity express is delayed for more than one hour, *dus/kejian* ‘so’ the railway system has encountered a problem.

The degree of subjectivity of a relation can also be marked by connectives. Some connectives are prototypically used for objective relations, such as French *parce que* ‘because’, German *weil* ‘because’, Chinese *yin’er* ‘as a result’, Dutch *omdat* ‘because’, *doordat* ‘because of the fact that’ and *daardoor* ‘as a result’. Other connectives are prototypical markers of subjective relations, such as French *puisque* ‘because’ and *car* ‘because’, German *denn* ‘because’, Chinese *kejian* ‘so’, Dutch *want* ‘because’ and *dus* ‘so’ (Degand & Pander Maat, 2003; Li, Evers-Vermeul, & Sanders, 2013; Pander Maat & Sanders, 2000; Pit, 2003; Stukker & Sanders, 2012; Zufferey, 2012). The information encoded in connectives helps the reader to interpret the subjectivity information as in (3c) (Canestrelli, Mak, & Sanders, 2013; Li, Mak, Evers-Vermeul,

& Sanders, 2017). Finally, there are connectives that are underspecified in terms of the degree of subjectivity, such as English *because* and *so* and Chinese *yinwei* ‘because’ and *suoyi* ‘so’, i.e. they can be used to mark subjective as well as objective relations (Andersson, 2016; Li et al., 2013). If clauses are connected by underspecified connectives, readers/hearers can only rely on other elements in the utterances to establish the degree of subjectivity of the relations, as in (3b).

The processing effects of different connectives marking causal relations have been examined in on-line processing experiments. In an on-line reading study by Canestrelli et al. (2013), it was shown that the Dutch subjective connective *want* ‘because’ leads to an immediate processing delay in the region directly after the connective as in example (4b), in comparison to the objective connective *omdat* ‘because’ as in example (4a). According to Canestrelli et al., the processing delay after *want* can be attributed to the effect of subjectivity because the subjective connective *want* triggers a subjective interpretation of the relation, which requires longer processing time compared to the specific objective connective *omdat*. According to the authors, the subjective connective triggers “the representation of someone’s belief or opinion, be it from the author, speaker, or other person whose reasoning is presented in the text” (Canestrelli et al., 2013: 1410). In accordance with that interpretation, the processing delay is cancelled by the presence of an indicator of the SoC, such as *volgens Peter* ‘according to Peter’ in (4c) (see also Traxler, Sanford, Aked, & Moxey, 1997).

(4)

a.

Hanneke was buiten adem, **omdat** ze vier trappen was afgerend om de post te halen.

Hanneke was out of breath, **because** she ran down four stairs to get the mail.

b.

Hanneke had haast, **want** ze was vier trappen afgerend om de post te halen.

Hanneke was in a hurry, **because** she ran down four stairs to get the mail.

c.

Volgens Peter had Hanneke haast, **want** ze was vier trappen afgerend om de post te halen.

According to Peter, Hanneke was in a hurry, **because** she ran down four stairs to get the mail.

(Adapted from Canestrelli et al., 2013: 1403)

Thus, there are longer reading times after the subjective connective *want* than after the objective connective *omdat*. How can this extra processing time be explained? We hypothesize that this extra processing time is due to the construction of subjectivity in situation models, as in the right panel of Figure 1. This process requires, first of all,

recognizing the SoC. Thus, readers are expected to focus more on the SoC in the situation model when the linguistic input indicates that the utterance is subjective than when it is objective. In order to examine the attention devoted to the SoC in processing in detail, we conducted two Visual World Paradigm (VWP) eye-tracking experiments in which we measured the proportion of looks at the speaker in the picture. Evidence from the VWP may reveal the process of focusing on the SoC in situation models.

In VWP eye-tracking studies, people exhibit a tendency to look at an object or person when they hear a reference to this object or person. For instance, previous VWP studies on pronoun resolution have shown that pronouns direct attention to the subject that is preferred on the basis of the context (Cozijn, Noordman, & Vonk, 2011; Järvikivi, Van Gompel, Hyönä, & Bertram, 2005; Knoeferle, Crocker, Pickering, & Scheepers, 2005; Pyykkönen, & Järvikivi, 2010). Pronouns refer to entities in the discourse model, and as a result of hearing the pronoun, people look at the picture of the entity that is referred to. Connectives do not refer to entities in the discourse model the way pronouns do. However, Koring, Mak and Reuland (2012) have shown that looks at a picture do not only occur based on direct references. In their experiment, verbs induced an increase in eye gazes to pictures that were closely related to the subject of the verb (e.g. verb: *fell*; subject: *wood*; target in the picture: a saw). The reactivation of the subject, as the argument of the verb, is considered to be “the result of integrating the verb and its argument into one representation” (Koring et al., 2012: 361). In the same vein, we expect subjective connectives to activate the process of focusing on the SoC, which is essential in the processing of subjectivity. Hence, we expect more looks at the SoC when listeners hear a subjective connective than when they hear an objective connective.

4.2 Experiment 1: Subjectivity in Dutch connectives

In Experiment 1, we compared the processing of the Dutch specific objective connective *daardoor* ‘as a result’ and the specific subjective connective *dus* ‘so’ – based on theoretical and corpus-based work underpinning these semantic-pragmatic profiles (Pander Maat & Degand, 2001; Pander Maat & Sanders, 2000, 2001). The participants heard auditory linguistic input while they were presented with two contrastive scenes that were presented on the same screen. The participants saw a scene with an SoC and a scene without an SoC (see Figure 2). They heard Dutch sentences either connected by the subjective connective *dus* ‘so’ (subjective condition) or by the objective connective *daardoor* ‘as a result’ (objective condition). We measured the changes in fixation proportions on the SoCs caused by the introduction of the connectives.

We predicted an increase in looks at the SoC when the participants heard the subjective connective, compared to when they heard the objective connective. This

prediction is based on the assumption that when people hear the subjective connective *duś*, they will automatically infer that the causal relation arises from someone's mind instead of reality, and will represent this information in the situation model. In the case of the connective *daardoor*, which indicates that the causal relation can be observed in the outside world, this process does not take place.

4.2.1 Method

Participants. Twenty native Dutch speakers participated in the experiment (age range: 18-26; average age: 23; education level: college level or above). All participants were recruited from an adult participant database of the UiL-OTS lab, Utrecht University. Their vision was normal or corrected to normal and no hearing problems were reported. Participants were paid five euro for their participation. Informed consent was obtained from participants.

Materials. Twenty items were used in the experiment. Each item consisted of a display with two scenes about the same event: a scene with an SoC on one side and a scene without SoC on the other side. An example item is provided in Figure 2. An example of auditory input that came along with the visual stimulus is presented in Table 1.

The left picture (without-SoC scene) depicts a scene of an objective situation. In the right picture (with-SoC scene), however, the salient part is a speaker making a statement. The objective scene presented in the with-SoC scene stays in a small speech bubble to suggest that the depicted person is speaking about that scene. The positions of the with-SoC scene and the without-SoC scene were counterbalanced: half of the pictures had the with-SoC scene on the left and without-SoC scene on the right, for the other half it was the other way around.



Figure 2. Example of visual stimulus. Left: without-SoC scene;⁸ Right: with-SoC scene.⁹

Along with the 20 visual items, we created 20 auditory items as the linguistic input, which started one to two seconds after the appearance of the visual items. Two versions of each auditory item were made with a manipulation of connectives. The sentences of both versions were composed of two clauses connected by either the subjective Dutch connective *dus* 'so' (subjective condition) or the objective connective *daardoor* 'as a result' (objective condition). The first clauses of the input sentences were kept identical in both conditions. The first clause was a description of the event depicted in the without-SoC scene. Table 1 shows an example of the auditory

⁸ This image was created by Frank J. Aleksandrowicz with the title: *Harshaw chemical company discharges waste water into the Cuyahoga river*, and recorded by the Environmental Protection Agency. The image is licensed under a free license (public domain). Link to the source: https://commons.wikimedia.org/wiki/File:HARSHAW_CHEMICAL_COMPANY_DISCHARGES_WASTE_WATER_INTO_THE_CUYAHOGA_RIVER_-_NARA_-_550193.jpg

⁹ This image was created by Alfred Pertl with the title: *Interviews für ORF Seitenblicke - Armin Assinger bei Buchpräsentation bei Thalia in Wien*. The file is licensed under the Creative Commons Attribution-Share Alike 4.0 International license. The size of the picture had been adjusted to fit the computer screen for the purpose of the experiment. Link to the source: https://commons.wikimedia.org/wiki/File:Interviews_f%C3%BCr_ORF_Seitenblicke_-_Armin_Assinger_bei_Buchpr%C3%A4sentation_bei_Thalia_in_Wien.jpg

input corresponding to Figure 2 (see Appendix 4 for more examples).

Table 1

Example Sentences in Dutch

<i>subjective condition connected by dus</i>	
Dutch	Het bedrijf heeft het water vervuild, dus / heeft het / onverantwoordelijke eigenaar.
English translation	‘The factory has been polluting the water, so it has an irresponsible owner.’
 <i>objective condition connected by daardoor</i>	
Dutch	Het bedrijf heeft het water vervuild, daardoor / heeft het / een milieuramp veroorzaakt.
English translation	‘The factory has been polluting the water, as a result it has caused an environmental disaster.’

The 40 combinations of images and sentences (20 per condition) were divided into two lists according to a Latin square design. Each list contained one version of a sentence, and ten items from each condition. Participants were assigned to read one of the two lists in the experiment. The presentation of items in each list followed a pseudo-randomization order, which was fixed for each participant.

The Dutch sentences were uttered by a native speaker and recorded in a soundproof room. The sound files were manipulated in such a way that the connective started 3.5 seconds after the onset of the trial. Post-conjunctive silence was added after the connectives so that the second clause started 1 second after the onset of the connective. The inserted silence time did not affect the naturalness of the speech because pauses at clause boundaries (i.e. in this case around the connectives) are common in spoken language (Hawkins, 1971; Schilperoord, 1996; Swerts, 1998). By adding the silence after the connectives, we ensured that the critical region – the processing periods of connectives did not differ between the two conditions. Two words following the connectives were kept identical across conditions.

We expected participants to exhibit different responses to the two conditions at the connective region, and that such responses would be reflected in their fixations on the displayed scenes (Figure 2). For each item, participants under different conditions heard the same utterance except the connective before the diverging point in the second clause. During the connective region, the content of the second clause is not known yet. Therefore, in terms of the longitudinal changes in fixations on the SoC, any differences between the two conditions before the diverging point should be due to the effect of the connective, instead of any other linguistic elements because these were identical across conditions up till that time point.

Apparatus. The experiment was conducted on an EyeLink-1000 eye tracker (SR Research), sampling at 500 Hz (every 2 ms). A high-speed camera was affixed to a Desktop Mount to measure the eye movements. The items were presented on a 36.4 * 27.2 cm (screen size) monitor via a host computer with the *real time Linux* system. The experiment was controlled by the software *ZEP* (version 1.6.3, Veenker, 2013).

Procedure. The experiment was performed in a sound-treated lab booth. Participants first received an instruction on the procedure of the eye-tracking experiment, the time it would take to finish the experiment, and that their eye movements were recorded by the camera during the experiment. They did not have any task apart from looking at the pictures while listening to the sentences. The participants were seated on a medical chair with about 40-70 cm distance to the display computer screen. The experiment started with a calibration procedure and a validation of the calibration. When both the calibration and validation were successful, the experimenter left the booth and the experiment began. The experiment was machine-paced, i.e. the participants were directed to the next test item automatically after a set period (5 seconds after the sentences ended). The whole experiment took about 10 minutes.

Analysis. In the original dataset, the position of the eye was captured every 2 ms. From this record, we sampled the position of the eye with 20 ms steps relative to connective onset. The eye-tracking data were analyzed by a multilevel logistic regression model (Goldstein, 2003; Mirman, Dixon, & Magnuson, 2008) in R (R Core Team, 2015) with the lme4 package (Bates, Maechler, Bolker, & Walker, 2015). With this analytical model, the eye-tracking data were treated as longitudinal data with time as a predictor. The time variable was centered for each time frame to ease the effort for R to converge the models. The dependent variable in the analysis was whether the participant looked at the picture of the SoC or not. The multilevel modeling approach we applied took into account the random intercepts of item and subject.

The probability of fixation on the SoC was modeled as a function of two factors: *Time* and *Connective* (subjective condition vs objective condition). The part in the picture with an SoC was taken as the area of interest. The mean proportion of looks at the SoC area was analyzed during the critical region. The critical region for analysis was from the onset of the connective (3.5 seconds after the beginning of the sentence) to 200 ms after the onset of the second clause (4.7 seconds after the beginning of the sentence). The extra 200 ms was included because that is approximately the time period needed for initiating and computing a saccade in reaction to the input (Matin, Shao, & Boff, 1993). Thus, by the time point 4.7s, participants were supposed to be still processing the information before the second clause. During that time frame, the only difference between the conditions that could influence the proportion of looks at

the picture was the connective itself.

4.2.2 Results

Figure 3 and 4 show the growth of the proportion of fixations on the SoC over time under the two conditions. The proportion of looks started from a similar level in both conditions. Then there was an increase of looks at the SoC under both conditions. This increase of looks was probably due to the fact that a picture of a person naturally attracts more attention. After the initial sharp increase of fixations on the scene with an SoC, both lines returned to a lower level before the connective was introduced. After the introduction of the connective, both lines returned to a lower level before the connective was introduced.

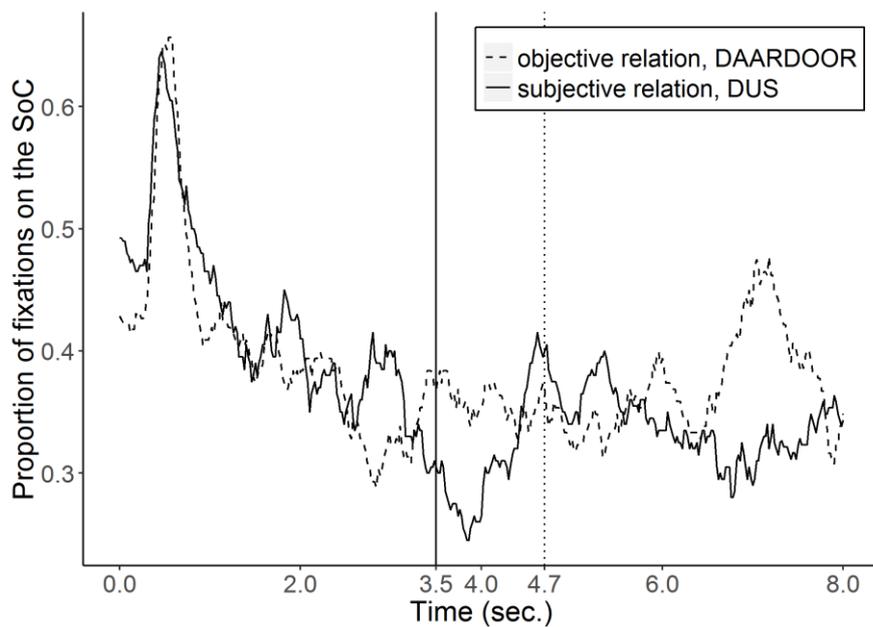


Figure 3. Proportion of fixations on the SoC throughout the trial.

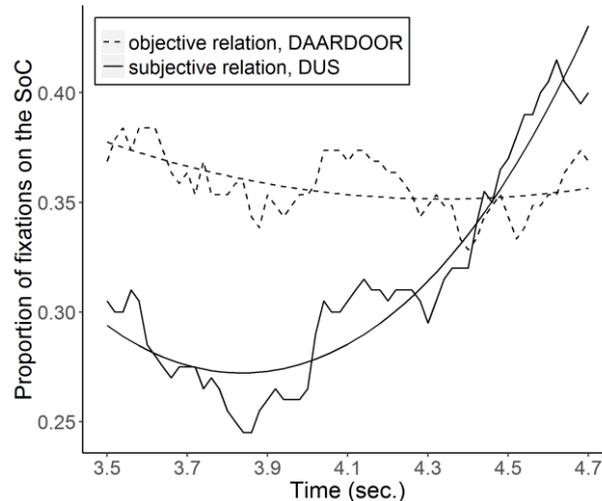


Figure 4. Proportion of fixations on the SoC during the connective time frame (3.5-4.7s, with regression lines estimated by the model) – subjective relation + *dus* vs objective relation + *daardoor*

The proportion of fixations on the SoC was modeled for the critical region (3.5s – 4.7s). We started with a base model with the random effects of items and subjects on the intercept, and the main effects of *Time*, $Time^2$ and *Connective*. By including *Time* as a predictor, the change of fixation proportions over time was captured. $Time^2$ was included because we anticipated a curved trajectory in the development of fixation proportions over time (Mirman et al., 2008). The fit of the model improved when the interaction effect of *Time* and *Connective* was added ($\chi^2(1) = 63.041$, $p < .001$). Moreover, adding the interaction of $Time^2$ and *Connective* also significantly increased the fit of the model ($\chi^2(1) = 11.465$, $p = .001$). Thus, the model with the random effects of subject and item, the main effects of *Time*, *Connective* and $Time^2$, the interaction effects of *Time* and *Connective*, and the interaction effects of $Time^2$ and *Connective* was used as the final model.

A summary of parameter estimates is presented in Table 2. The main effect of *Connective* shows the influence of connective on the average proportion of fixations on the SoC over the entire region (Table 2, ID.4). The significant main effect means that during the entire critical region, the subjective condition had a lower mean proportion of looks at the SoC than the objective condition. There was no effect of *Time* or $Time^2$ in the objective condition (Table 2, ID.2, ID.3), indicating a more or less straight flat line for the objective condition. The proportion of looks at the SoC remained stable after the objective connective *daardoor* was introduced.

However, there were also significant interaction effects of *Connective* and *Time* (Table 2, ID.5), as well as *Connective* and $Time^2$ (Table 2, ID.6). A significant increase

in the proportion of looks at the SoC under the subjective condition compared to the objective condition was observed. The significant interaction effect of *Time*² and *Connective* indicated that the subjective connective led to an increasing concave upward curve.

Table 2

Parameter Estimates for the Best-Fitting Multilevel Logistic Regression (reference level=objective connective)

<i>ID</i>	<i>Parameters</i>	<i>Estimate</i>	<i>SE</i>	<i>z</i>	<i>p</i>
<i>Fixed factors</i>					
1	Intercept	-.668	.171	-3.895	< .001
2	Time	-.084	.057	-1.487	.137
3	Time ²	.180	.180	.998	.318
4	Subjective <i>dus</i>	-.370	.044	-8.466	< .001
5	Subjective <i>dus</i> * Time	.640	.081	7.865	< .001
6	Subjective <i>dus</i> * Time ²	.877	.259	3.386	.001
<i>Random factors</i>					
1	Subject	.297	.545		
2	Item	.272	.522		

In sum, the statistical tests showed a different development in the proportion of looks at the SoC under two connective conditions (as shown in Figure 4): during the critical region, attention to the SoC stayed around the same level under the objective condition. In the subjective condition right after the introduction of the connective *dus* ‘so’, there was an increase of looks at the SoC in comparison to the objective connective *daardoor* ‘as a result’.

4.2.3 Discussion

In previous Dutch online reading experiments, a processing delay was observed immediately after the Dutch subjective connective *want* ‘because’, compared to objective *omdat* ‘because’ (Canestrelli et al., 2013). According to Canestrelli et al., this processing delay is related to the fact that an SoC is added to the mental representation of the discourse. After all, the interpretation of subjective relations is associated with somebody who is thinking, reasoning or arguing – the SoC. Therefore, we expected subjective connectives to draw attention to the SoC.

The results for the critical time region in the current study has confirmed this prediction about the effects of connectives in the processing of subjectivity information. During the 1.2 seconds after the onset of the connectives, there is a

difference in the change of looks at the SoC over time between the two connective conditions. In comparison to the objective connective *daardoor*, an increase in the proportion of looks at the SoC is found after the subjective connective *dus*. Apart from the interaction effects of *Connective* and *Time*, there is a main effect of *Connective* suggesting a lower mean proportion of looks at the SoC in the subjective condition compared to the objective one. This main effect is probably due to the fact that the looks at the SoC have not fallen to a stable level in the Dutch experiment within 3.5 seconds. These results need to be validated in an experiment with a longer time frame after the sentence onset.

The findings are consistent with prior studies in at least two respects. First, at the level of the cognitive representations, this VWP processing study gives credibility to previous theoretical hypotheses that the interpretation of a subjective relation involves the process of focusing on an SoC in the situation model. Second, the experiment shows that the process of focusing on the SoC can be instructed by the subjective connective *dus*. In other words, a subjective connective itself can function as a processing instruction which increases language users' attention to the responsible subject in the situation, i.e. the SoC, in comparison to an objective connective. The effect of the subjective connective *dus* 'so' found here is in line with the processing delay detected right after the Dutch subjective connective *want* 'because' compared to the objective connective *omdat* 'because' (Canestrelli et al., 2013). Both the processing delay and the increased attention to the SoC are associated with the processing of subjectivity.

Two questions remain unanswered in the Dutch experiment. First, although we have observed a difference between the subjective connective *dus* and the specific objective connective *daardoor* in directing people's attention to the SoC, it is not clear whether this difference should be attributed to the characteristics of the subjective connective or those of the objective connective. That is, the difference can either be due to an effect of the subjective connective in directing attention to the SoC, or an effect of the objective connective of guiding attention away from the SoC, or both. Without a neutral connective condition as a reference, we cannot disentangle the effects of connectives marking different degrees of subjectivity. A comparison to a neutral connective can also shed more light on what we are measuring precisely with this method.

Second, the growth patterns of fixations in the later period (i.e. after the critical region) were not comparable under the two conditions, because the contents of the second clause in the two conditions were partially different. In Dutch, the subjective connective *dus* is prototypically used for subjective causal relations, while the objective connective *daardoor* is used to express objective ones. Hence, it is impossible to create conditions with exactly the same second clause following *dus* and *daardoor*. In order to investigate the processing effect of the subjective connective at a later stage, we need a causal connective underspecified in the degree of subjectivity,

similar to English *because*. However, such underspecified connectives are not available in Dutch. Experiment 2 was therefore run in Chinese, which has both underspecified and specific causal connectives.

Mandarin Chinese has an underspecified causal connective *suoyi* ‘so’, which can be used in both subjective and objective relations, as well as an objective connective *yin’er* ‘as a result’ and a subjective connective *kejian* ‘so’ (Li et al., 2013). The rich profile of Chinese causal connectives is beneficial to our research in two aspects. First, the underspecified connective *suoyi* can serve as a reference level for the two specific types of connectives. By comparing the two specific connectives to the underspecified connective as the neutral level, we can identify the exact processing effects of each type of connective. Second, both for the objective and the subjective conditions, items can be created that differ only in the specificity of the connective. Thus, the effect of connectives in the later processing stages can be examined properly.

4.3 Experiment 2: Subjectivity in Chinese connectives

The two Dutch connectives in Experiment 1 are both marked for the level of subjectivity, either marking an objective relation or a subjective relation. In order to directly see the effect of the marking of the degree of subjectivity by connectives, the processing of specific connectives should be compared to the processing of underspecified connectives. Li et al. (2017) tested this on Chinese in a reading experiment. In clauses with a subjective causal relation (5a), readers slowed down at the end of the second segment in the condition with the underspecified connective *suoyi* ‘so’ compared to the condition with the specific subjective connective *kejian* ‘so/therefore’. In objective relations, such as (5b), Li et al. did not find a late difference between relations marked with the specific connective *yin’er* ‘as a result’, compared to relations marked with the underspecified connective *suoyi*.

- (5)
- a. Meng Na na tiao kuzi xianzai xiande hen fei, *suoyi/kejian* ta bi yiqian shou le bu shao.
That (old) pair of trousers now look very baggy on Meng Na, **so/therefore** she has become much thinner now than before.
 - b. Meng Na yi nian lai baoshou weibing de zhemo, *suoyi/yin’er* ta bi yiqian shou le bu shao. Ta shi liang ge haizi de muqin.
For a year Meng Na has been suffering from stomach trouble, **so/as a result** she has become much thinner now than before.

(Li et al., 2017: 51)

Li et al. (2017) attributed the processing cost to the cognitive complexity of subjectivity. To be specific, in the *suoyi* condition, the subjectivity information is not explicitly marked as it is in the *kejian* condition. Therefore, people have to process the subjectivity at a later stage, after they find out that the relation is subjective on the basis of the propositional content. Compared to the *kejian* condition, this process leads to an increase in reading times in the *suoyi* condition. After the specific subjective connective *kejian*, readers do not need to process subjectivity at a later processing stage as the subjectivity is already encoded by the connective, and hence processed at an earlier stage. In an experiment using the Visual World Paradigm, this late effect should be reflected in an increase in looks at the target in subjective relations with *suoyi* at the point where readers can infer the subjectivity from the content of the sentence, compared to the same relations marked with *kejian*.

In Experiment 2, the three causal connectives with different subjectivity profiles illustrated in (5a) and (5b) were used. There were four conditions: the two relation types (objective vs subjective) were expressed by either the underspecified connective *suoyi* or the specific connectives of the particular types of relation – *yin'er* or *kejian*.

4.3.1 Method

Participants. 46 native Chinese speakers participated in the experiment (age range: from 20 to 31; average age: 27; education: college level or above). All participants were recruited in Utrecht. Their vision was normal or corrected to normal and no hearing problems were reported. Participants received five euro for their participation. Informed consent was obtained from participants. The eye-tracking data from six participants were dropped because of poor data quality.

Materials. Twenty items were used in the experiment. Each of the items was composed of a picture (presented on the screen) paired with a spoken sentence. In this Chinese experiment, we adopted the same picture setting as in the Dutch experiment (Figure 2). In addition to the experimental items, there were sixteen fillers. Filler sentences were temporal relations marked by temporal connectives. Filler pictures had the same properties as the experimental items.

Four conditions of sentences as auditory input were created according to a two-by-two design: relation type (subjective vs objective) and connective type (specific vs underspecified). Examples of the four conditions are presented in Table 3 (see Appendix 4 for a full list of the experimental items). Subjective relations were expressed either by the specific subjective connective *kejian* or the underspecified connective *suoyi*. Objective relations were expressed either by the specific objective connective *yin'er* or the underspecified connective *suoyi*. The onset of the connective was at six seconds from the item onset.

The 80 combinations of images and sentences (20 per condition) were divided

into four lists according to a Latin square design. Each list contained one version of a sentence, and ten items from each condition. Participants were assigned to read one of the four lists in the experiment. The presentation of items in each list followed a pseudo-randomization order, which was fixed for each participant.

In the two subjective conditions, a modal verb (either *keneng* ‘may’ or *yiding* ‘must’) was inserted about 3-4 characters after the connective. The onset time of the modal verb (8.3 second from the item onset) was the same across items. By including the modal verbs, it was possible to measure the effects of *kejian* and *suoyi* on the proportion of looks at the SoC later in the sentence. The modal verb *keneng* ‘may’ or *yiding* ‘must’ provides information to the participant that the relation is subjective since these modal verbs introduce an epistemic stance. In the *kejian* condition, hearers are supposed to know the relation is subjective at the connective, while in the *suoyi* condition, this degree of subjectivity is not clear until the modal verb is heard.

Table 3

Example Sentences in Chinese

Subjective relation + *suoyi* (underspecified):

Zhe jia siying de huagongchang yizhi zai paifang wushui, **suoyi** ta de changzhu **keneng** bing bu guanxin huanjing baohu.

Subjective relation + *kejian* (specific):

Zhe jia siying de huagongchang yizhi zai paifang wushui, **kejian** ta de changzhu **keneng** bing bu guanxin huanjing baohu.

English translation

‘The private chemistry factory has been polluting the water, **so** its owner **may** not care about environment protection.’

Objective relation + *suoyi* (underspecified)

Zhe jia siying de huagongchang yizhi zai paifang wushui, **suoyi** fujin heliu li de yulei da mianji siwang.

Objective relation + *yin'er* (specific)

Zhe jia siying de huagongchang yizhi zai paifang wushui, **yin'er** fujin heliu li de yulei da mianji siwang.

English translation

‘The private chemistry factory has been polluting the water, **so** fishes in the rivers nearby are dying at a large scale.’

The Chinese sentences were uttered by a native Mandarin Chinese speaker and recorded in a soundproof room. Silence was added to sentences in such a way that the connective started 6 seconds after the onset of each trial, and the second clause started 7.2 seconds after the onset of the trial. In the subjective conditions, the onset of modal

verbs was at 8.3 seconds from the onset of the item.

In each item, the first clause was identical in all four conditions. Therefore, in terms of the longitudinal changes in fixations on the SoC, any differences among the four conditions before the onset of the second clause should be attributed to the effect of connective instead of any other linguistic elements. Moreover, the second clauses were the same in the two subjective conditions, and so were the second clauses in the two objective conditions. Thus, during the processing of the second clause, any differences in fixation patterns between two conditions of the same relation type were due to the processing of the connectives, which varied in terms of whether they marked the level of subjectivity, or due to the effects of modal verbs in subjective relations.

Apparatus & procedure. The Chinese experiment was conducted with the same apparatus and the same procedure as the Dutch one. The Chinese experiment took 15 minutes for each subject.

Analysis. A multilevel logistic regression analysis was performed (cf. section 2.1). We measured the proportion of fixations in two time intervals: the *Connective time frame* - from the onset of the connective (6.0s) to 200 ms after the beginning of the second clause (7.4s); and in the subjective condition, the *Modal verb time frame* - from the onset of the modal verb (8.3s) till 1.2s after the onset of the modal verb (9.5s). As in Experiment 1, in the former time frame, 200 ms were added because it takes approximately that time to initiate a saccade in response to an external input. Therefore, the fixation patterns during the time period 6.0s - 7.4s reflected the reaction to the connectives. In the later time frame we measured the effect of the modals on the processing of the subjective relations.

4.3.2 Results

Just as in the Dutch experiment, the fixations on the SoC increased to a high percentage at the very beginning. The picture with a speaking person attracted most of the attention. Then the proportion of fixations in the four conditions declined gradually to a low level, until the onset of the connectives (6s). The proportion of looks at the SoC diverged after different connectives were introduced (as shown in Figure 5, 6 and 7). At a later stage of the two subjective conditions, the modal verb time frame, the specific connective *kejian* and underspecified connective *suoyi* also showed different effects on the fixation proportion (see Figure 8).

Connective time frame (6s – 7.4s)

For the connective time frame, we modeled the proportion of fixations on the SoC as a function of four factors: *Connective type* (specified vs underspecified), *Relation type* (objective vs subjective), *Time* and *Time*². The interactions were also taken into account. We started with a base model with subjects and items included as random factors, and *Connective type*, *Relation type*, *Time* and *Time*² as fixed factors. Then we added the two-way interactions of *Connective type* and *Time*, *Relation type* and *Time*, *Connective type* and *Relation type*, and the three-way interaction of *Connective type*, *Relation type* and *Time* to the model. The model fit significantly improved by adding the interaction effects ($\chi^2(4) = 51.37, p < .001$).

We also explored models with quadratic components: the two-way interactions of *Connective type* and *Time*², and *Relation type* and *Time*², and the three-way interaction of *Connective type*, *Relation type* and *Time*². Adding these quadratic components improved the model fit significantly ($\chi^2(3) = 26.534, p < .001$). Therefore, the final model contained the three-way interaction of *Connective type*, *Relation type* and *Time*, the three-way interaction of *Connective type*, *Relation type* and *Time*², as well as all the two-way interactions. Table 4 shows the parameter estimates of the final model.

There was a significant three-way interaction of *Connective type*, *Relation type* and *Time* (Table 4, ID.11). To disentangle this three-way interaction, a series of pairwise comparisons were made between different conditions. First, we tested whether the effect in Experiment 1 was replicated by comparing the subjective relation marked by the specific connective *kejian* with the objective relation marked by the specific connective *yin'er*. Second, the effect of objective marking was tested by comparing the objective relation marked by the underspecified connective *suoyi* with the objective relation marked by the specific connective *yin'er*. Finally, the effect of subjective marking was tested by comparing the subjective relation marked by the underspecified connective *suoyi* with the subjective relation marked by the specific connective *kejian*.

Table 4
*Parameter Estimates for the Best-Fitting Multilevel Logistic Regression – Relation Type * Connective Type * Time*

<i>ID</i>	<i>Parameters</i>	<i>Estimate</i>	<i>SE</i>	<i>z</i>	<i>p</i>
<i>Fixed factors</i>					
1	Intercept	-1.641	.173	-9.473	< .001
2	Time	.071	.054	1.315	.189
3	Time ²	.008	.148	.051	.959
4	Connective type (specific)	.001	.047	.024	.981
5	Relation type (subjective)	-.039	.047	-.823	.410
6	Connective type (specific) * Relation type (subjective)	-.035	.067	-.520	.603
7	Connective type (specific) * Time	-.260	.078	-3.340	.001
8	Relation type (subjective) * Time	.025	.076	.333	.739
9	Connective type (specific) * Time ²	-.374	.212	-1.763	.078
10	Relation type (subjective) * Time ²	.608	.207	2.939	.003
11	Connective type (specific) * Relation type (subjective) * Time	.528	.108	4.869	< .001
12	Connective type (specific) * Relation type (subjective) * Time ²	.202	.296	.684	.494
<i>Random factors</i>					
1	Subject	.680	.825		
2	Item	.237	.486		

Comparison 1: *subjective relation specific connective ‘kejian’ versus objective relation specific connective ‘yin’er’.*

Figure 5 shows the proportion of looks over time in the two specific connective conditions. The model with the interaction effect of *Connective* and *Time* increased the fit of the model compared to the base model with only the random effects and the main effects of *Connective*, *Time* and *Time*² ($\chi^2(1) = 54.081$, $p < .001$). Adding the interaction between *Connective* and *Time*² also improved the model fit significantly ($\chi^2(1) = 15.329$, $p < .001$). Therefore, the model including all the main effects, the interaction effect of *Connective* and *Time*, the quadratic component of *Time*² and its interaction effect with *Connective* was taken as the final model. The parameter estimates of this model are presented in Table 5.

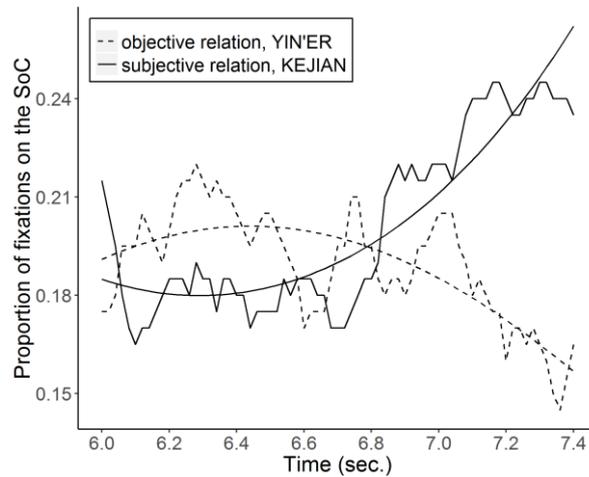


Figure 5. Proportion of fixations on the SoC during the connective time frame (6.0s – 7.4s, with regression lines estimated by the model) – subjective relation + *kejian* vs objective relation + *yin'er*

Table 5

Parameter Estimates for the Best-Fitting Multilevel Logistic Regression – Subjective Relation + *kejian* vs Objective Relation + *yin'er*

ID	Parameters	Estimate	SE	z	p
Fixed factors					
1	Intercept	-1.834	.245	-7.490	< .001
2	Time	-.198	.057	-3.454	.001
3	Time ²	-.383	.156	-2.464	.014
4	Subjective <i>kejian</i>	-.061	.049	-1.245	.213
5	Subjective <i>kejian</i> * Time	.579	.079	7.288	< .001
6	Subjective <i>kejian</i> * Time ²	.848	.217	3.918	< .001
Random factors					
1	Subject	1.769	1.330		
2	Item	.275	.524		

The effect of *Time* in the reference condition (*yin'er* condition) showed that the proportion of fixations on the SoC declined after *yin'er* (specific connective for objective relations) over time (Table 5, ID.2). A negative effect of *Time*² (Table 5, ID.3) was observed for the *yin'er* condition, i.e. a downward curve for the development of the proportion of looks at the SoC after *yin'er*.

The interaction effect of *Time* and *Connective* shows different tendencies under the two conditions. The growth of the proportion of looks in the *subjective relation + kejian* condition over time diverged considerably from the *objective relation + yin'er* condition (Table 5, ID.5). The interaction of $Time^2$ and *Connective* was also significant (Table 5, ID.6), which implied an opposite trajectory of the proportion of looks over time after the two specific connectives expressing different types of relations.

By releveling the factor of *Connective* (with *kejian* as the reference level), we found a significant effect of time for the *kejian* condition. The proportion of looks increased in the *subjective relation + kejian* condition ($\beta = .382$, $SE = .055$, $z = 6.923$, $p < .001$). The *subjective relation + kejian* condition also had a positive quadratic effect of $Time^2$, i.e. an upward curve for the development of the proportion of looks at the SoC after *kejian* ($\beta = .465$, $SE = .151$, $z = 3.090$, $p = .002$).

Comparison 2: *objective relation specific connective 'yin'er' versus objective relation underspecified connective 'suoyi'.*

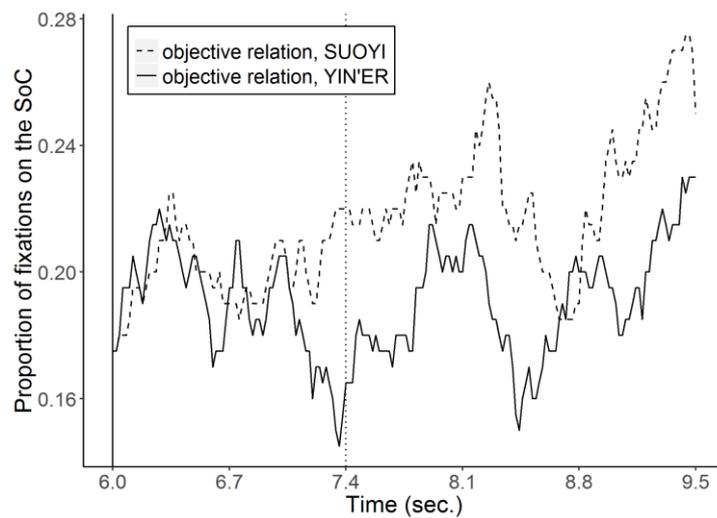


Figure 6-1. Proportion of fixations on the SoC during the time frame 6.0s – 9.5s – objective relation + *suoyi* vs objective relation + *yin'er*.

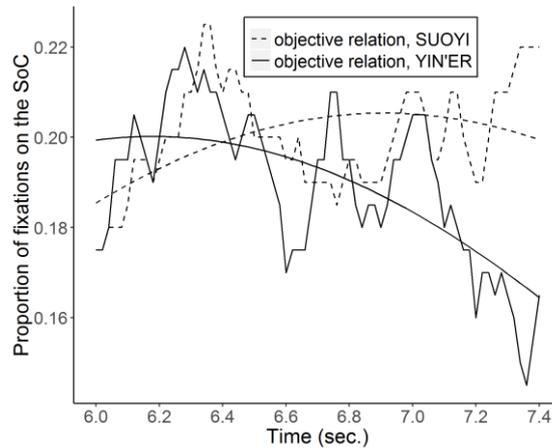


Figure 6-2. Proportion of fixations on the SoC during the connective time frame (6.0s – 7.4s, with regression lines estimated by the model) – objective relation + *suoyi* vs objective relation + *yin'er*.

The development of the fixation proportions in the two objective relation conditions is presented in Figure 6. The model including the interaction effect of *Connective* and *Time* increased the fit of the model compared to the base model with the random effects and main fixed effects ($\chi^2(1) = 11.782, p < .001$). Adding the interaction between *Connective* and *Time*², however, did not improve the model fit significantly ($\chi^2(1) = 3.238, p = .072$). Therefore, the final model included the main effects of *Connective*, *Time* and *Time*², and the interaction effect of *Connective* and *Time* (see Table 6).

There was a significant interaction effect of *Connective* and *Time* (Table 6, ID.5). Under the *suoyi* condition, there was no development in the looks at the SoC over time (Table 6, ID.2 and ID.3). However, as revealed earlier in *Comparison 1*, the proportion of looks at the SoC decreased over time after *yin'er* (Table 5, ID.2).

Table 6
Parameter Estimates for the Best-Fitting Multilevel Logistic Regression – Objective Relation + yin'er vs Subjective Relation + suoyi

<i>ID</i>	<i>Parameters</i>	<i>Estimate</i>	<i>SE</i>	<i>z</i>	<i>p</i>
<i>Fixed factors</i>					
1	Intercept	-1.913	.286	-6.696	< .001
2	Time	.080	.058	1.395	.163
3	Time ²	-.191	.111	-1.723	.085
4	Objective <i>yin'er</i>	.038	.035	1.095	.274
5	Objective <i>yin'er</i> * Time	-.279	.081	-3.436	.001
<i>Random factors</i>					
1	Subject	1.967	1.402		
2	Item	.622	.789		

Comparison 3: subjective relation specific connective 'kejian' versus subjective relation underspecified connective 'suoyi'.

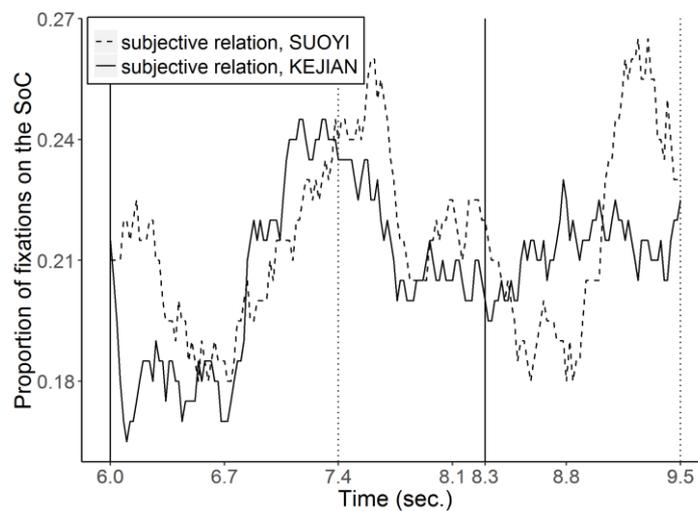


Figure 7-1. Proportion of fixations on the SoC during the time frame 6.0s – 9.5s – subjective relation + *suoyi* vs subjective relation + *kejian*.

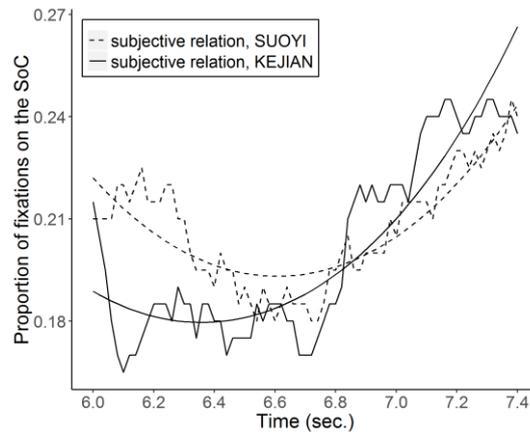


Figure 7-2. Proportion of fixations on the SoC during the connective time frame (6.0s – 7.4s, with regression lines estimated by the model) – subjective relation + *suoyi* vs subjective relation + *kejian*.

The data of the two subjective relation conditions are presented in Figure 7. The model including the interaction effect of *Connective* and *Time* increased the fit of the model compared to the base model with only the main effects ($\chi^2(1) = 12.931, p < .001$). Adding the interaction between *Connective* and *Time*² did not significantly improve the model fit ($\chi^2(1) = .675, p = .411$). Therefore, the model including the main effects of *Connective*, *Time* and *Time*², and the interaction effect of *Connective* and *Time* was taken as the final model. The parameter estimates of this model are presented in Table 7.

Table 7

Parameter Estimates for the Best-Fitting Multilevel Logistic Regression – Subjective Relation + *kejian* vs Subjective Relation + *suoyi*

ID	Parameters	Estimate	SE	z	p
Fixed factors					
1	Intercept	-1.765	.214	-8.257	< .001
2	Time	.100	.053	1.863	.062
3	Time ²	.547	.105	5.226	< .001
4	Subjective <i>kejian</i>	-.057	.032	-1.791	.073
5	Subjective <i>kejian</i> * Time	.275	.076	3.599	< .001
Random factors					
1	Subject	1.056	1.028		
2	Item	.366	.605		

Table 7 shows no significant main effect of *Connective*: the mean fixation proportions on the SoC during the whole time frame were the same after the two different connectives (Table 7, ID.4). There was a change over time in *suoyi* condition as the quadratic effect of *Time*² suggests: an increase following a decrease at the beginning (Table 7, ID.3). From *Comparison 1*, we observed a main effect of *Time* in the *kejian* condition ($\beta = .382$, $SE = .055$, $z = 6.923$, $p < .001$), and *Time*², i.e., an increase of fixation proportions after *kejian* over time as well ($\beta = .465$, $SE = .151$, $z = 3.090$, $p = .002$).

However, the interaction effect of *Time* and *Connective* (Table 7, ID.5) indicated a difference in the form of changes under the *kejian* condition compared to the *suoyi* condition. What is this difference? Figure 7-2 suggests a similar rise in both conditions, preceded by different patterns in the beginning.

To test this, the time frame was split into two periods: 6.0s - 6.62s (the average duration time of the connectives), and 6.62s - 7.4s. In the former time frame, the average fixation proportion was lower under the *kejian* condition in comparison to the *suoyi* condition ($\beta = -.181$, $SE = .049$, $z = -3.696$, $p < .001$). In the latter time frame, the average fixation proportions did not differ across conditions ($\beta = .060$, $SE = .043$, $z = 1.390$, $p = .165$). Main effects of *Time* were found for both the *suoyi* condition and the *kejian* condition: the proportion of looks at the SoC increased over time after *kejian* ($\beta = .655$, $SE = .094$, $z = 7.002$, $p < .001$), and also increased after *suoyi* ($\beta = .655$, $SE = .094$, $z = 7.002$, $p < .001$). The absence of interaction effect between *Connective* and *Time* ($\chi^2(1) = 1.013$, $p = .314$) indicated that the proportion of looks under two conditions increased equally over time during the latter period. Therefore, the difference in fixation proportions observed at the connective time frame was mainly due to the differences at the connective itself. During the silence period after the average offset of the connectives (6.62s) and before the onset of the second clauses (7.4), the growth of fixation proportions was the same for the *kejian* condition and the *suoyi* condition.

Modal verb time frame (8.3s – 9.5s)

For the connective time frame, *comparison 3* shows no difference in the re-activation of the SoC after the underspecified connective *suoyi* compared to the specific subjective connective *kejian*. The question then is whether there were no processing differences at all between these two conditions. In order to investigate this, we looked at the processing of the modal verb, which gives explicit information on the

subjectivity of the utterance.

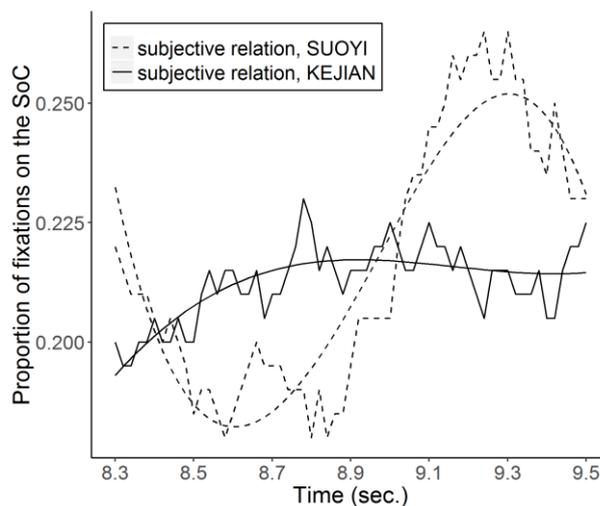


Figure 8. Proportion of fixations on the SoC during the modal verb time frame (8.3s – 9.5s, with regression lines estimated by the model) – subjective relation + *kejian* vs subjective relation + *suoyi*.

The proportion of looks at the SoC was measured for the two subjective conditions during the modal verb time region. In Figure 7-1 and Figure 8, an S-curve with two bends can be observed under the *suoyi* condition. Therefore, the cube of *Time* was added to the base model. The base model with the main effects (*Connective*, *Time*, $Time^2$, $Time^3$) was improved by adding the interaction effect of *Time* and *Connective* ($\chi^2 = 1.098$, $df = 1$, $p = .001$). Adding the interaction effect of $Time^2$ and *Connective* also increased the model fit significantly ($\chi^2 = 5.740$, $df = 1$, $p = .017$), and so did the interaction of $Time^3$ and *Connective* ($\chi^2 = 1.856$, $df = 1$, $p = .001$). The base model with the main effects and all of the interaction effects (linear, quadratic and cubic) is presented in Table 8.

Table 8

Parameter Estimates for the Best-Fitting Multilevel Logistic Regression – Subjective Relation + ‘kejian’ vs Subjective Relation + ‘suoyi’ (modal verb frame)

<i>ID</i>	<i>Parameters</i>	<i>Estimate</i>	<i>SE</i>	<i>z</i>	<i>p</i>
<i>Fixed factors</i>					
1	Intercept	-1.737	.289	-6.019	<.001
2	Time	.011	.177	.064	.949
3	Time ²	-.284	.224	-1.266	.206
4	Time ³	.351	.73	.48	.631
5	Subjective <i>suoyi</i>	-.076	.053	-1.442	.149
6	Subjective <i>suoyi</i> * Time	1.061	.25	4.24	<.001
7	Subjective <i>suoyi</i> * Time ²	.772	.314	2.459	.014
8	Subjective <i>suoyi</i> * Time ³	-3.354	1.028	-3.264	.001
<i>Random factors</i>					
1	Subject	.893	.945		
2	Item	1.184	1.088		

No significant *Time* effect was observed for the *kejian* condition during the modal verb time frame (linear: ID.2; quadratic: ID.3; cubic: ID.4). Those non-significant results indicated a more or less horizontal line for the *kejian* condition, i.e. the proportion of looks in the *kejian* condition stayed at the same level during this time frame after the modal verb was presented.

However, the interaction effects of *Connective* with the *Time* factors indicated diverging trajectories of the fixation proportions over time under the *suoyi* condition compared to the reference level *kejian*. Compared to the *kejian* condition, there was a temporary increase in the proportion of fixations at the SoC in the *suoyi* condition after the modal verb was presented.

4.3.3 Discussion

The Chinese experiment replicated the main result of the Dutch experiment: the proportion of looks at the SoC increased over time after the specific subjective connective *kejian* ‘so’ compared to the condition with the objective connective *yin’er* ‘as a result’, just as there was an increase after *dus* ‘so’ compared to *daardoor* ‘as a result’. These results confirm that the processing of subjectivity is associated with an increased focus on the SoC. Connectives encoding a high degree of subjectivity lead to relatively more attention to the SoC compared to those encoding a low degree of subjectivity.

Moreover, the Chinese experiment provides answers to the two questions left open by the Dutch experiment. The first question was whether the difference in the fixation proportions is due to the effect of the subjective connective, the objective connective, or both. By comparing the specific subjective connective and the specific objective connective to the underspecified connective *suoyi*, we found that the increase of focus on the SoC is not unique for the subjective connective. The underspecified connective *suoyi* also leads to an increase in the proportion of fixations on the SoC. By contrast, the objective connective *yin'er* differs from the subjective connective and the underspecified connective by guiding attention away from the SoC. There is an increase in focus on the SoC after *suoyi* compared to *yin'er* that is similar to the increase in focus found after the subjective connective *kejian*.

The question then is whether this implies that readers in the *suoyi* condition also process the information as subjective. The answer to this question can be found in the pattern of fixations in the remainder of the second clause of the subjective relations. In the *suoyi* condition, the increase in the proportion of looks at the SoC in the modal verb frame indicates that readers at this point still need to reactivate the source of information in their mental representation. In the *kejian* condition they do not have to do this anymore, which is why there is no increase in looks during the modal verb frame. From this difference we derive that readers have already reactivated the SoC at the connective region in the *kejian* condition, but not in the *suoyi* condition. This pattern of results supports the hypothesis that the subjective connective *kejian*, but not the connective *suoyi*, instructs the listener to incorporate the SoC in the situation model of the input. The underspecified connective *suoyi* only marks the causal nature of the relation, and hence hearers need to incorporate the SoC in the situation model when they hear the modal verb. The question then remains what the increase in looks at the SoC after the underspecified connective *suoyi* means. Or, to put it in more general terms, what exactly are we measuring with this VWP method? These questions will be discussed in the general discussion.

The current VWP evidence in combination with the previous on-line reading results gives insight in the mental representations of subjective relations: in order to interpret subjective relations, people need to keep track of an SoC, who has to be reactivated when information is interpreted as originating from this subject/agent. This tracking and reactivation process is reflected in increased attention to the SoC in the VWP, and in longer reading times in reading experiments.

4.4 General discussion

The current visual world paradigm experiments set out to investigate the way in which people process subjectivity in causal coherence relations. Previous processing studies used online reading times to measure effects of subjectivity (Canestrelli et al., 2013;

Li et al., 2017). Canestrelli et al (2013) reported longer processing times immediately after subjective connectives compared to objective connectives. According to Canestrelli et al., this difference in on-line reading times should be attributed to the fact that the interpretation of subjective relations requires the construction of a mental representation in which an SoC is involved. In other words, the reader or hearer needs to incorporate the source of information into the mental representation of subjective relations. This takes more time and effort compared to the processing of objective relations, in which no such SoC is involved.

In this study, we tested whether this explanation is on the right track, by looking at the way in which subjective and objective connectives affect the amount of attention devoted to the SoC. In the utterances in our study, this was always the speaker. The two VWP experiments provide evidence for the interpretation of the reading time effects in terms of the construction of a mental model in which an SoC is involved. A significant difference in attention to the speaker is found between the Dutch subjective connective *dus* ‘so’ and the objective connective *daardoor* ‘as a result’, as well as for the Chinese subjective connective *kejian* ‘so’, compared to the objective connective *yin’er* ‘so’. As hypothesized, the subjective connectives, which encode a higher degree of subjectivity, lead to an increased attention to the SoC compared to the objective connectives. This effect is found in both languages, immediately at or after the connective, the region that corresponds to the region in which a processing delay was found in online reading experiments (Canestrelli et al., 2013). This indicates that subjective connectives indeed trigger a growing focus on the narrator as the SoC in the situation models, whereas objective connectives function as a linguistic cue for hearers to pay less attention to the SoC.

To further address the exact effect of connectives in directing attention to the SoC, we compared the Chinese underspecified connective *suoyi* ‘so’ to the subjective connective *kejian* and the objective connective *yin’er*. The change in attention to the SoC triggered by the underspecified connective *suoyi* patterns with the effects of the subjective connective *kejian* – both *suoyi* and *kejian* lead to an increased attention to the SoC. The objective connective *yin’er*, however, differs from these two connectives in that it guides people’s attention away from instead of to the SoC. It explicitly specifies the relation to be an objective one, and more specifically a non-volitional content relation. This is a causal type of relation between events that do not involve any SoC (Sanders et al., 2009). From the VWP data we can derive that *yin’er* immediately signals that the interpretation of the relation does not involve an SoC.

The absence of a processing difference between the underspecified connective *suoyi* and the subjective connective *kejian* is unexpected. The question is whether this implies that the same mental model is constructed in both cases. For the answer to this question it is crucial to look at the behavior of the participants in response to the modal verb that occurs later on in the utterance. This modal verb is an unambiguous signal that the relation is to be interpreted as subjective. Only in the *suoyi* condition the

modal verb led to an increase in looks at the SoC; in the *kejian* condition no change in the proportion of looks at the SoC was found at the modal verb region. Again, this difference can be related to the construction of a mental representation that involves an SoC. The underspecified connective *suoyi* leaves the hearer/reader uninformed about the degree of subjectivity, and therefore requires activation of the SoC at a later stage than in the *kejian* condition, which already specifies high subjectivity at the connective. And again, this VWP difference in attention to the SoC patterns with the results of previous online reading studies: in the predicate region of the connective clause – the region comparable to the modal verb region in our study – Li et al. (2017) found longer reading times in causal sentences connected by the underspecified connective *suoyi* compared to those connected by *kejian*.

Similar to the increase in looks at the SoC immediately after the subjective connective *kejian*, the activation of the SoC at the modal verb region in sentences with *suoyi* can be related to the tracking of source information. However, this leaves the question what the earlier increase in looks – at the region containing *suoyi* – exactly reflects, as this underspecified connective itself does not provide information that the relation involves an SoC. In order to understand this increase in looks, it is important to remember that the SoC in our experimental sentences was always the speaker. In other words, the picture containing the SoC always represented the person who reported his or her conclusion on the basis of an argument. One might argue that the speaker is somehow involved in the representation of all types of information: we often remember who has told us something, even if it is an objective fact. In addition, speakers usually do not convey information randomly, but present information they think is relevant or interesting, and they structure it in a specific way. As we might see connectives as processing instructions provided by the speaker on how to structure information, connectives in general may put the processing instructor, the speaker, in focus. Still, the degree to which they do this may vary with their degree of subjectivity. If this explanation is correct, all connectives may lead to a focus on the speaker to some extent, because this speaker – in his/her role of narrator – is involved in structuring the discourse and presenting links between utterances. This explanation implies that, in our experiment, there are at least two triggers for an increase or decrease in looks at the person depicted as the speaker in the Visual World Paradigm. All connectives – underspecified, objective and subjective ones – activate the speaker in his role of narrator. On top of that, certain connectives activate the speaker in his/her role of SoC, as in the case of subjective connectives, or result in deactivation of the speaker, as in the case of objective connectives. Objective connectives indicate the speaker, but not in his/her role as SoC, but ‘just’ as the narrator. That is why the attention to the person in the picture is less than in the case of subjective connectives, which activate a speaker as both the narrator and the SoC.

If focusing on the speaker after connectives is a general tendency, we would expect to find a similar increase in looks at the speaker picture after other types of

connectives. A first indication that this idea is on the right track comes from the processing effects of the connectives in the filler items in our experiment. In these fillers, sentences were connected by either of two temporal connectives: *erhou* ‘and then’ and *ranhou* ‘then’. An analysis of these items indeed confirms the general tendency of connectives to generate an increase in looks at the speaker.¹⁰ A follow-up question is how connectives marking other types of relations, and linguistic cues that provide other links to the speaker (e.g. perspective markers) may influence the attention to the speaker. More research is needed in order to further investigate how the attention to the speaker – either in his role of the narrator or in his role of SoC – comes and goes while people process language.

This paper connects visual attention in response to linguistic cues marking subjectivity with the effects of subjectivity on processing times in reading. We have constructed a method with which we can explore the nature of the extra processing time in subjective relations. This measure of looks at the SoC gives us insight into the development of situation models during processing: an intentional mind (the SoC) is attended to when a connective indicates the involvement of a narrator making subjective opinions or coherence connections between segments. However, if the connective specifies the relation to be objective, peoples’ attention is guided away from the SoC and the objective scene gets more attention. The results support the proposal by Graesser et al. (1997, 1999; see also Sparks & Rapp, 2011; Strømsø, Bråten, & Britt, 2010) that comprehenders keep track of the source of information (*who said what*) in the situation models they build for comprehension. For causal relations, we have shown that the subjectivity of the information is tracked and immediately updated on the basis of linguistic cues such as connectives and modal verbs.

In our exploration of differences between connectives with varying degrees of subjectivity, we only selected one type of objective connectives. However, objective connectives and the relations they express can vary in terms of volitionality. It has repeatedly been argued (Pander Maat & Degand, 2001; Pander Maat & Sanders, 2000, 2001; Sanders et al., 2009) that non-volitional content relations such as (6a) – the objective type of relation we used in this study – differ from volitional content relations such as (6b) in terms of the involvement of an SoC.

¹⁰ In the analysis of fillers with the connectives *erhou* and *ranhou*, we observed main effects of *Time* ($\beta = 0.141$, $SE = 0.030$, $z = 4.762$, $p < 0.001$) and *Time*² ($\beta = 0.203$, $SE = 0.081$, $z = 2.502$, $p = 0.012$) on the proportion of looks at the picture containing the speaker during the connective time frame (6s-7.4s). For these temporal connectives, the proportion of looks showed a curvilinear increase over time with a decline at the beginning.

- (6)
- a. The factory has been polluting the water, *daardoor/yin'er/ as a result* the local water supply is contaminated.
 - b. The factory has been polluting the water, *daarom/yushi/that is why* the residents nearby decided to file a complaint to the local government.

The non-volitional content relation involves no SoC at all, while the volitional content relation involves a so-called character SoC, *the residents* in this case. Different from the speaker type of SoC (recall example (2a)) that is responsible for the subjective reasoning, a character SoC is the person responsible for a volitional act. In some languages, this distinction in volitionality is encoded by connectives as well. For instance, Dutch *omdat* ‘because’ and *daarom* ‘that’s why’ as well as Chinese *yushi* ‘that’s why’ explicitly mark causal relations as volitional content (Degand & Pander Maat, 2003; Li et al., 2013; Pit, 2003; Sanders et al., 2009), whereas Dutch *daardoor* ‘as a result’ and Chinese *yin'er* ‘so’ indicate non-volitionality. Future research could be done to investigate the role of connectives in the processing of volitional content relations and non-volitional content relations with the current experimental set-up in VWP. The current study has shown that the degree of subjectivity encoded in connectives functions as a specific processing instruction for building mental representations, in two typologically totally different languages. Future research is needed to investigate how different types of SoC are constructed in situation models, and how connectives of other types may influence online reading times and the construction of situation models.

Chapter 5. The role of connectives and perspective markers in the on-line processing of subjective relations

5.1 Introduction

Language users frequently express their attitudes, judgements and (un)certainly. Various linguistic cues are employed to differentiate such subjective expressions from mere descriptions of objective real-world facts. Within a clause, adverbials express attitudes towards certain propositional elements, for instance, on the size of a ferry (example 1). Some linguistic cues have a larger scope, over the whole clause or sentence. For instance, *perhaps* and *may* mark the content of the clause as the hesitant judgement of the author (examples 2 and 3). Other cues attribute the content to another person instead of the author, such as *John said* in (4).

- (1) The passenger capacity of this ferry is *ridiculously* large.
- (2) *Perhaps*, the passenger capacity of this ferry is large.
- (3) The passenger capacity of this ferry *may* be large
- (4) *John said* the passenger capacity of this ferry is large.

In linguistic theories, the involvement of a speaker whose opinion is conveyed is called subjectivity (Finegan, 1995; Langacker, 1990; Lyons, 1977). In terms of coherence relations in discourse, a basic distinction is drawn between subjective and objective relations (Sanders & Sweetser, 2009). For instance, (5) has a higher degree of subjectivity compared to example (6), because it expresses a claim-argument relation instead of a consequence-cause relation.

- (5) The passenger capacity of this long-distance ferry must be large, *because* it needs one hour before everyone is boarded.
- (6) Passengers of the long-distance ferry have arrived early at the gate, *because* it needs one hour before everyone is boarded.

Readers are sensitive to such differences in the degree of subjectivity. Traxler, Bybee and Pickering (1997; cf. also Traxler, Sanford, Aked, & Moxey, 1997) found that subjective relations marked by *because*, such as (7b), led to longer processing times at the second clause compared to objective relations marked by the same connective, such as (7a). The processing delays were observed at the pre-final region of the second clause (*first prize* in (7)) when it becomes clear that the second clause is an argument for the claim in the first clause. The eye-tracking study provided evidence for the incremental interpretation of sentences – as soon as readers are informed on the

subjectivity of the relation, they react to the complexity by slowing down the reading speed.

(7)

- a. Heidi felt very proud and happy, because/ she won/ first prize/ at the art show.
- b. Heidi could imagine and create things, because/ she won/ first prize/ at the art show.

(Traxler et al., 1997a: 485)

The question is: why do subjective relations lead to longer processing times? One possible reason is that establishing subjectivity involves tracking the sources of information. To process subjective relations, readers need to attribute the information to someone's personal belief, which may require more cognitive load than merely interpreting objective relations based on real-world facts.

In English, both objective relations as in (6), and subjective relations as in (5) can be linguistically expressed by *because*, a causal connective that is underspecified for subjectivity. In other languages, connectives mark the degree of subjectivity on top of their function in explicating a specific type of coherence relation between clauses. For instance, the Dutch counterpart of the subjective relation in (5) is preferably marked by the subjective connective *want* 'because', instead of the objective connective *omdat* 'because', as is illustrated in (8). The Dutch connectives *want* 'because' (example (8a)) and *dus* 'so' (example (8b)) and the Chinese connective *kejian* 'so' indicate the reasoning of a speaker, while the objective connectives *daardoor* 'as a result', *daarom* 'that's why', *omdat* 'because', and *yin'er* 'as a result' mark the relation between two clauses as real-world fact (example (9a) and (9b); Degand & Pander Maat, 2003; Li, Evers-Vermeul, & Sanders, 2013; Pander Maat & Sanders, 2000; Pit, 2003, 2006; Stukker & Sanders, 2012; Zufferey, 2012).

(8)

- a. De passagierscapaciteit van deze veerboot is groot, **want** het duurt een uur voordat iedereen aan boord is.
The passenger capacity of this ferry is large, **because** it takes one hour before everyone is boarded.
- b. Het duurt een uur voordat iedereen aan boord is, **dus** de passagierscapaciteit van deze veerboot moet groot zijn.
It takes one hour before everyone is boarded on the ferry, **so** the passenger capacity of this ferry must be large.

(9)

- a. Passagiers van de veerboot zijn vroeg bij de poort aangekomen, *omdat* het een uur duurt voordat iedereen aan boord is.
Passengers of the ferry have arrived early at the gate, *because* it takes one hour before everyone is boarded.
- b. Het duurt een uur voordat iedereen aan boord van de veerboot is, *daarom* zijn de passagiers vroeg bij de poort aangekomen.
It takes an hour before everyone is boarded on the ferry, *so the* passengers have arrived early at the gate.

Thus, apart from marking the causal nature of the relation between two discourse segments, some connectives also mark the degree of subjectivity. The connective *because* only has the function of marking causality, while *want* ‘because’, as a specific subjective connective has two functions: first, it marks the fact that the two segments of the sentence are connected in a causal relation; second, it indicates that the relation between the two segments can be attributed to a source – a person making the claim: the causal relation is marked to be subjective. These functions of connectives affect processing: the subjective connective *want* ‘because’ leads to an immediate processing delay compared to the objective connective *omdat* ‘because’, well before the content of the second segment makes clear that the relation is subjective (Canestrelli, Mak, & Sanders, 2013). Thus the effect in Dutch arises earlier than the effect that was found in English in sentences with the underspecified connective *because* (Traxler 1997b). However, is this difference in processing between *because* in English and *want* ‘because’ in Dutch really due to the difference between underspecified connectives and specified connectives in terms of subjectivity, or is it simply due to language differences between Dutch and English or differences between experiments? Mandarin Chinese provides a chance to examine this comparison with one language and in one experiment.

The three types of connectives illustrated in (7), (8) and (9) are available in Chinese: an underspecified causal connective *suoyi* ‘so’, a specific subjective connective *kejian* ‘so’, and a specific objective connective *yin'er* ‘as a result’. An experimental study on the processing of Chinese connectives has shown that the underspecified connective *suoyi* ‘so’, compared to the specific subjective connective *kejian* ‘so’, led to longer processing time at the end of subjective sentences (Li, Mak, Evers-Vermeul, & Sanders, 2017). In sentences with the underspecified *suoyi* ‘so’ such as (10a), readers can only find out the relation is subjective on the basis of the propositional content of the second clause. This finding is comparable to the results of the experiments with the English connective *because*. When readers are informed by a specific subjective connective at an earlier processing stage that the relation is subjective as in (10b), their reading time on the final region was faster than in the sentences marked by an underspecified connective such as (10a).

(10)

- a. Changtu youlun xuyao tiqian yi ge xiaoshi zhunbei dengchuan, **suoyi** youlun yiding zaikeliang feichang keguan.
The long-distance ferry needs one hour in advance for boarding, **so** the passenger capacity of the ferry must be considerable.
- b. Changtu youlun xuyao tiqian yi ge xiaoshi zhunbei dengchuan, **kejian** youlun yiding zaikeliang feichang keguan.
The long-distance ferry needs one hour in advance for boarding, **so** the passenger capacity of the ferry must be considerable.

The effects of connectives on reading times in on-line reading studies of different languages can be explained in terms of the tracking of the source of information in processing subjective relations. On the one hand, when the connective does not indicate that the relation is subjective (in the case of *suoyi* ‘so’ and *because*), the subjectivity of the relation, and hence the need to look for the source that is responsible for the subjective relation, is revealed by the propositional content of the sentence. This results in longer reading times in the second clause (Li et al., 2017; Traxler et al., 1997a). On the other hand, when the connective does encode the subjectivity of the relation (as in the case of *want* ‘because’) the search for the source of information starts immediately, which is visible in an immediate processing delay after the connective (Canestrelli et al., 2013).

Thus, previous reading experiments have shown that the reading time for subjective relations is longer than that of objective relations. Moreover, readers make use of linguistic cues such as connectives while processing the relation’s degree of subjectivity. However, what we do not know from these studies is when and how readers make use of other linguistic cues of subjectivity in the clause preceding and/or following the connective. How do these cues interact with the signal provided by connectives?

In Section 2, we will discuss other types of linguistic cues marking subjectivity in discourse. In Section 3, we will present the methodological considerations underlying our processing study on Chinese, in which we use eye-tracking to examine the role of stance markers, modal verbs and connectives during the processing of subjective causal relations. In Section 4, the results of this experiment will be discussed, and in Section 5 we will relate these to theoretical accounts of subjectivity.

5.2 Collocation patterns and processing effects of subjectivity markers

Connectives are not the only cues to express subjectivity in discourse relations. Modal verbs/adverbials such as *may* in example (3) and cue phrases such as *John said* in (4) also indicate the source of information. These cues can be used in combination with connectives to express subjectivity. How linguistic cues of subjectivity distribute in discourse and how they are processed by readers may shed light on the influence of linguistic marking on the comprehension of subjectivity.

5.2.1 Subjectivity markers in discourse

The corpus-based collocational study in Chapter 3 examined the contextual collocates of the underspecified connective *suoyi* and of the specific subjective *kejian*, especially the expression of subjectivity in the context. Some linguistic cues appeared more often in the context of *suoyi* compared to that of *kejian*, such as the modal verbs *keneng* ‘may’ and *yinggai* ‘should’, and the cognition verbs *zhidao* ‘know’, *xiang* ‘think’, and *renwei* ‘consider’. By contrast, other linguistic cues were identified as the collocates of *kejian*, such as the expressions of surprisal *jurán* ‘unexpectedly’ and *jingran* ‘surprisingly’, and the indicators of importance *zhongyao* ‘important’. From the perspective of Horn’s speaker/hearer economy (1984), the collocation patterns can be explained as a tendency to keep the uttered information both sufficient for hearers and efficient for speakers. Modal verbs and cognition verbs supplement the underspecified information of *suoyi*, in that they are indicators of the degree of subjectivity. However, such supplementary linguistic cues of subjectivity are not needed when the degree of subjectivity has already been encoded by *kejian*.

The linguistic expressions found in the collocational study can be well classified in parallel to different types of evaluations (Bednarek, 2009), perspective markers (Dancygier & Sweetser, 2005), or stance markers (Conrad & Biber, 2000). Stance markers express subjectivity along different dimensions. Conrad and Biber (2000: 57) suggest three sub-types of stance markers (see Bednarek 2006, 2009 and Thompson & Hunston, 2000 for similar classifications):

- epistemic stance, which indicates how certain the speaker or writer is, or where the information comes from (e.g. *probably*, *according to the President*);
- attitudinal stance, which indicates feelings or judgements about what is said or written (e.g. *surprisingly*, *unfortunately*);
- style stance, which indicates how something is said or written (e.g. *honestly*, *briefly*).

In sum, the collocational study in Chapter 3 shows that modal verbs and stance markers occur differently in the context of the connectives *suoyi* and *kejian*. Modal verbs and epistemic stance markers patterned with *suoyi*, while attitudinal stance markers patterned with *kejian*. The current paper provides a perspective to investigate the interplay between connectives and other linguistic cues in on-line processing, taking into account the collocational patterns found in authentic data on language use. In the following sections we review the results of earlier processing experiments investigating the effect of these cues on processing.

5.2.2 Modal verbs as subjectivity markers

In the collocational study we observed that modal verbs appear more in the context of *suoyi* than in the context of *kejian*. This may be due to the fact that modal verbs provide information about subjectivity, in the absence of such information at the connective *suoyi*.

The role of modal verbs as expressions of subjectivity is supported by processing evidence. The visual world paradigm (VWP) eye-tracking experiment in Chapter 4 shows the influence of modal verbs on people's visual attention, and the interaction between the effect of modal verbs and that of connectives. In the VWP-experiment, participants heard subjective relations containing a modal verb *yiding* 'must' or *keneng* 'may' either marked by the underspecified connective *suoyi* 'so' as in (10a), or by the subjective connective *kejian* 'so' as in (10b). Meanwhile, they were presented with pictures containing two scenes: an objective scene as described by the first clause of (10) on one side and a picture of a speaker on the other side with a speech bubble containing the same objective scene. The proportion of looks on the speaker was measured as a reflection of people's attention to the speaker. When the connective was underspecified, an increased attention to the speaker was observed at the modal verb in the second clause. By contrast, when the connective indicated a subjective relation, as in the case of *kejian*, the modal verb did not have such an effect in directing attention to the speaker. The VWP findings showed that modal verbs function as processing instructions: they instruct readers to pay attention to the source of information – the speaker in the picture. When this information is encoded by the subjective connective, the modal verb does not have this effect anymore since the subjectivity information encoded by the modal verb is redundant. Therefore, it is reasonable to assume that the modal verb is the point at which readers realize that the relation is subjective.

In this paper, we use modal verbs as a means to measure the effect of subjectivity in the absence of or in combination with other cues of subjectivity. With a specific subjective connective marking the subjective relation, we expect a delay at or immediately after *kejian*, when subjectivity information is presented for the first time

(cf. Canestrelli et al., 2013). By contrast, when the relation is connected by an underspecified connective *suoyi*, the processing delay is expected at a later point in the sentence where the relation is unfolded as subjective by the modal verb (cf. Li et al., 2017).

5.2.3 Epistemic stance markers as subjectivity markers in the preceding clause

Epistemic stance markers have also been shown to influence the processing of subjective relations. Traxler et al. (1997a) found that epistemic stance markers such as *John said/thought* and *perhaps* facilitate the processing of subjective relations with *because*. The extra processing time of subjective relations compared to objective ones was cancelled out by the presence of stance markers. In Canestrelli et al.'s (2013) reading studies, epistemic stance markers such as *according to Peter* in (11c) cancelled out the processing asymmetry between clauses containing the objective connective *omdat* in (11a) and clauses containing the subjective connective *want* in (11b) (see Section 1). In other words, with an epistemic stance marker in the preceding context, the reading times of the region after the connective *want* in (11c) were comparable to the ones of (11a).

(11)

a.

Hanneke was buiten adem, *omdat* ze vier trappen was afgerend om de post te halen.

Hanneke was out of breath, *because* she ran down four stairs to get the mail.

b.

Hanneke had haast, *want* ze was vier trappen afgerend om de post te halen.

Hanneke was in a hurry, *because* she ran down four stairs to get the mail.

c.

Volgens Peter had Hanneke haast, *want* ze was vier trappen afgerend om de post te halen.

According to Peter, Hanneke was in a hurry, *because* she ran down four stairs to get the mail.

(Adapted from Canestrelli et al., 2013: 1403)

In both the English and the Dutch study, the epistemic stance markers help readers establish a source of information by explicitly attributing the content of the first clause to a character (*John* or *Peter*) or indicating the presence of the speaker who is not sure (*perhaps*) about the following content. Therefore, at a later region – when subjectivity is unfolded by the specific connective *want* or by the propositional content in the second clause, readers have already processed the information in the first segment as

subjective, and hence the reading times in sentences with epistemic stance markers are shorter than those in sentences without epistemic stance markers.

The facilitation effect of epistemic stance markers on the processing of the subjective connective *want* in (11c) demonstrates an overlap in epistemic meanings between the subjective connective *want* and the epistemic stance markers. This overlap is consistent with collocation patterns in language use from the corpus data discussed in Section 2.1 – epistemic stance markers and modal verbs co-occurred more with the underspecified connective *suoyi* compared to the specific subjective connective *kejian*. The corpus-based study and previous on-line reading studies provide evidence to Horn’s theory from the perspectives of both the speaker economy and the reader economy. From the perspective of saving the speaker’s production efforts – not providing more information than necessary, with a specific subjective connective *kejian*, fewer epistemic stance markers were used in the context compared to the underspecified connective *suoyi*. For readers, the co-occurrence of epistemic stance markers and subjective connectives, for instance *according to Peter* and *want* ‘so’, provides sufficient linguistic cues to help readers establish a subjective relation, and that is why a reading facilitation was found.

Stance markers such as *John said/thought* and *according to Peter* influenced the processing of subjective relations marked by *want* ‘because’ in (11c) and *because* in (7b). However, subjective relations can also be formulated in an argument-claim structure as in (8b) and (10). In an argument-claim relation, subjective connectives such as Dutch *dus* ‘so’ and Chinese *kejian* ‘so’ mark the second segment as a claim and the relation as subjective. In these cases, epistemic stance markers such as *according to Peter* in the preceding context of the relation may not directly mark the second segment as a claim. In the argument-claim type of relation illustrated in (12), the stance marker *according to Peter* marks the first segment as an argument made by *Peter*. But the claim that *she must have been in a hurry* is not necessarily the opinion of *Peter*. This raises the following question: will stance markers play a similar role in the interpretation of the argument-claim type of subjective relations as they do for the claim-argument relations? If they do, will the effect surface at the connective and/or at the modal verb?

- (12) According to Peter, Hanneke ran down four stairs to get the mail, so she must have been in a hurry.

Our first research question is:

How does the presence of perspective markers affect the processing patterns at the connective region and at later regions in subjective relations with an argument-claim structure?

As predicted by previous reading experiments, interpreting subjectivity is an accumulative process – once readers have already established a source of information in the previous context, by *according to Peter* for instance, it does not take extra efforts to attribute additive contents to this source later on. In line with this prediction, we expect the stance markers to cancel out the processing asymmetry of connective *kejian* and *suoyi* at the connective region. Moreover, at the modal verb region, the longer reading times under the *suoyi* condition in comparison to the *kejian* condition should also be eliminated by the presence of stance markers in the preceding context.

5.2.4 Attitudinal stance markers as subjectivity markers in the preceding clause

Previous processing studies in this area revealed the role of epistemic stance markers in the processing of subjective relations and connectives. But what about the effects of other types of stance markers? In a follow-up study, Canestrelli, Mak and Sanders (2016) examined the processing difference between subjective and objective connectives in subjective contexts containing attitudinal markers such as *ridiculously* in (1): *The passenger capacity of this ferry is ridiculously large*. Although such attitudinal stance markers also express subjectivity, they did not affect the processing of subjectivity as much as epistemic stance markers like *according to Peter* did. In other words, the reading disadvantage of *want* in comparison to *omdat* was not cancelled out by stance markers like *ridiculously* that mark the attitudinal stance of the speaker. This lack of effects may be due to the fact that the evaluative marker *ridiculously* only modified elements within the clause instead of the entire clause. Thus, we do not know yet whether attitudinal stance markers are different from epistemic stance makers in influencing the processing the degree of subjectivity of the relation. This is critical to the understanding of subjectivity. By definition, all types of stance markers imply a source of information, but they indicate different ways of how a source is involved: epistemic stance markers indicate the speaker's degree of certainty about the expressed opinion, whereas attitudinal stance markers indicate the speaker's attitude/feelings.

Our second research question is:

Do the effects of epistemic stance markers on the processing of subjective relations also hold for attitudinal stance markers?

In terms of the comprehension of subjectivity, this question concerns whether the influence of stance markers on the processing asymmetry is due to the general subjectivity expressed by all stance markers – the involvement of a source of information – or by particular dimension(s) of subjectivity expressed by specific types of stance markers expressing different ways in which an information source is

involved. If the extra time to process subjectivity is due to the cognitive load of establishing a source of information in general, regardless of the way in which the source is involved, all types of stance markers should have the effect of alleviating the processing load. In this sense, all stance markers should cancel out the processing delay associated with subjective connectives. If the processing load of subjectivity is not due to establishing the source of information in general, but related to how this source is involved, different stance markers should have different effects. Accordingly, the processing asymmetry between subjective connectives and underspecified connectives may not be cancelled out by the presence of other types of stance markers.

5.3 Method

Participants

65 participants took part in this experiment (44 female, mean age 26, age range 18 – 36 years). All participants were native speakers of Mandarin Chinese and were paid for their participation. Informed consent was obtained from all participants.

Materials

The materials consisted of 48 sets of test items and 48 fillers. All test items were subjective sentences, and each item was composed of an S1 (introductory clause of factual events) and an S2 (judgement/conclusion) connected by a connective, and an S3 as a spill-over sentence (full list in Appendix 5). We manipulated the use of connectives (*suoyi*, *kejian*) and stance markers in S1 (no stance marker, epistemic stance marker and attitudinal stance marker). In all sentences, we inserted a modal verb in the second clause of subjective relations right after the subject.

We systematically investigated the processing effects of different connectives and stance markers with six experimental conditions. Two no stance marker conditions with either *suoyi* or *kejian* were created to set a baseline of how Chinese speakers process subjective relations marked by different connectives without the influence of stance markers (see Table 1, condition 1 & 2). Two epistemic stance marker conditions (epistemic stance marker + *suoyi* versus epistemic stance marker + *kejian*) were used to test the effects of epistemic stance markers (see Table 1, condition 3 & 4). In order to compare the effects of attitudinal stance markers and epistemic stance markers in affecting the processing of subjective relations, we added two attitudinal stance marker conditions (attitudinal stance marker + *suoyi* versus attitudinal stance marker + *kejian*) (see Table 1, condition 5 & 6).

Thus, each set of test items consisted of six versions of a sentence (Table 1). The 48 sets were divided into six lists according to a Latin square design. Each list contained one version of a sentence, and eight items in each condition. Participants were assigned to read one of the six lists in the experiment. The presentation of items

in each list followed a pseudo-randomization order, which was fixed for each participants.

One third of the sentences were followed by a verification statement. The verification statements were about the content of single clauses and not about the relation between clauses. Participants were informed to judge the statements by pressing a “True” or “False” button.

Table 1

Example test items

Condition 1 & 2 No stance markers

*Zhejia xiaoqu cheku zai baitian shijian bei dao, **suoyi/kejian** xiaoqu anbao **keneng** zuo de bu daowei. Yixie jumin yaoqiu zengjia xiaoqu shexiangtou.*

This living:district garage during day time PASS rob, **CONJ** district security **may** do PRT NEG enough. Some residents request increase living:district camera.

The garage of this district was robbed during daytime, **so** the security of this district **may** not have done enough (work). Some residents requested to increase the number of cameras in this district.

Condition 3 & 4 Epistemic stance markers

*Zhejia xiaoqu cheku **tingshuo** zai baitian shijian bei dao, **suoyi/kejian** xiaoqu anbao **keneng** zuo de bu daowei. Yixie jumin yaoqiu zengjia xiaoqu shexiangtou.*

This living:district garage **hearsay** during day time PASS rob, **CONJ** district security **may** do PRT NEG enough. Some residents request increase living:district camera.

As reported, the garage of this district was robbed during daytime, **suoyi/kejian** ‘**so**’ the security of this district **may** not have done enough (work). Some residents requested to increase the number of cameras in this district.

Condition 5 & 6 Attitudinal stance markers

*Zhejia xiaoqu cheku **jingran** zai baitian shijian bei dao, **suoyi/kejian** xiaoqu anbao **keneng** zuo de bu daowei. Yixie jumin yaoqiu zengjia xiaoqu shexiangtou.*

This living:district garage **surprisingly** during day time PASS rob, **CONJ** district security **may** do PRT NEG enough. Some residents request increase living:district camera.

Surprisingly the garage of this district was robbed during daytime, **suoyi/kejian** ‘**so**’ the security of this district **may** not have done enough (work). Some residents requested to increase the number of cameras in this district.

Apparatus

The experiment was conducted on an EyeLink-1000 eye tracker (SR Research), sampling at 500 Hz (every 2 ms). The experiment was controlled by the software *ZEP* (version 1.6.3, Veenker, 2013). A high-speed camera was affixed to a Desktop Mount

to measure the eye movements during reading.

Procedure

Participants were tested individually in a sound-treated lab booth. They first received an instruction on the procedure of the experiment and tasks, which included reading sentences on the screen and judging the verification statements randomly following the sentences. Before starting the test, the experimenter adjusted the seat and distance to make sure the participant sat in a comfortable way while their eyes could be measured properly. The practice trial started with a calibration test followed by a validation of the calibration. This calibration procedure was repeated before the real experiment. The practice trial (with three items) acquainted participants with the experiment procedures. After the practice, the experiment began. The experiment took about 30 minutes. Automatic drift checks were performed before each item.

5.4 Results

Example (13) demonstrates the regions we analyzed within the second clause (S2). The first region was the connective region. The subject region contained the words between the connective and the modal verb – usually the subject of S2. The modal verb region consisted of the modal verb and three or four characters (one or two words) after the modal verb, and the final region contained the final words of S2.

(13)

[region1 **Suoyi/kejian**] [region2 *xiaoqu anbao*] [region3 **keneng zuo de**] [region4 *bu daowei*].

[region1 **CONJ**] [region2 *district security*] [region3 **may do PRT**] [region4 **NEG enough**].

[region1 **So**] [region2 *the security of this district*] [region3 **may not**] [region4 *have done enough (work)*].

We included data from 60 participants; the data from the five remaining participants were excluded because of poor data quality (e.g., severe drifts, too many blinks). The participants whose data were included in the analysis all had an accuracy rate of above 84% (27 out of 32) in judging the verification statements. Blinks and missing observations due to skipped regions were excluded from the data. We computed four measures of the eye-tracking data: *first pass first gaze* (fg), *first pass total gaze* (tg), *first pass regression path* (rp), and *total fixation durations* (totfix). *First pass first gaze* is the time between the onset of the first pass first fixation in a region and the end time of the last fixation before leaving the region in any direction; *first pass total gaze* is the sum of the durations of fixations that fall within the coded region before moving

progressively; *first pass regression path* is the time between the onset of the first pass first fixation in a certain region and the end time of the last fixation before leaving the region in a forward direction; *total fixation durations* is the total reading time of a region. Observations that were two standard deviations above or below the item mean and the subject mean were kept out from the analyses. These outliers were less than 0.5% of the total number of observations (fg: 0.2%; tg: 0.2%; rp: 0.5%; totfix: 0.3%). Table 2 shows the means and standard deviations in all four measures.

For all four measures, a Linear Mixed Effects Regression (lmer) analysis (Baayen, Davidson, & Bates, 2008) was performed on the reading time of each region. We started with a full model including the main effects of *Connective* and *Stance* and the interaction between *Connective* and *Stance*. The no stance marker condition was taken as the baseline of the factor *Stance*, and the *suoyi* condition as the baseline of the factor *Connective*. The intercepts of items and subjects were included as random factors. We then tested whether excluding the interaction effects from a model would decrease the model fit significantly. The interaction effects whose exclusion did not cause a significant decrease of the model fit were dropped from the final model. Since the main research questions of this study concern the influence of different connectives and stance markers on reading, we did not drop any main effects from the model. In the following reports of findings, we focus on the significant results and the important comparisons of reading times.

Region 1: connective region

Dropping the interaction between *Connective* and *Stance* did not reduce the model fit significantly in any of the four measures (fg: $\chi^2(2)=1.044$, $p=.593$; tg: $\chi^2(2)=1.407$, $p=.495$; rp: $\chi^2(2)=0.514$, $p=.774$; totfix: $\chi^2(2)=1.370$, $p=.504$). Therefore, the final analytical models only included the factors of *Connective* and *Stance* and the random factors.

A main effect of *Connective* at this region showed that the reading times at the subjective connective *kejian* were longer than the reading times of the underspecified connective *suoyi* across all *Stance* conditions (fg: $\beta=0.052$, $SE=0.018$, $t=2.856$, $p=.004$; tg: $\beta=0.046$, $SE=0.018$, $t=2.499$, $p=.013$; rp: $\beta=0.057$, $SE=0.023$, $t=2.511$, $p=.012$). There was a main effect of *Stance*: the reading time of connectives under the attitudinal stance marker condition was longer compared to the no stance marker condition (rp: $\beta=0.067$, $SE=0.028$, $t=2.406$, $p=.016$). No other significant effects were observed in this region.

Table 2

Means aggregated by participants measured by first pass first gaze (fg), first pass total gaze (tg), first pass regression path (rp), and total fixation durations (totfix). Standard deviations are provided in parentheses.

	Connective region	Subject region	Modal verb region	Final region
<i>first pass first gaze (fg)</i>				
No stance marker + <i>suoyi</i>	243 (50)	344 (111)	444 (161)	396 (169)
No stance marker + <i>kejian</i>	257 (70)	330 (106)	426 (136)	376 (148)
Epistemic stance marker + <i>suoyi</i>	244 (61)	361 (137)	406 (154)	387 (127)
Epistemic stance marker + <i>kejian</i>	270 (73)	345 (126)	415 (174)	393 (148)
Attitudinal stance marker + <i>suoyi</i>	256 (64)	363 (125)	407 (157)	393 (148)
Attitudinal stance marker + <i>kejian</i>	273 (80)	336 (106)	427 (166)	371 (116)
<i>first pass total gaze (tg)</i>				
No stance marker + <i>suoyi</i>	254 (62)	393 (133)	471 (180)	433 (168)
No stance marker + <i>kejian</i>	260 (71)	373 (130)	470 (174)	435 (171)
Epistemic stance marker + <i>suoyi</i>	252 (62)	414 (175)	434 (169)	445 (194)
Epistemic stance marker + <i>kejian</i>	280 (80)	383 (150)	450 (190)	446 (174)
Attitudinal stance marker + <i>suoyi</i>	265 (72)	409 (156)	450 (202)	454 (168)
Attitudinal stance marker + <i>kejian</i>	279 (96)	389 (134)	457 (177)	410 (129)
<i>first pass regression path (rp)</i>				
No stance marker + <i>suoyi</i>	263 (71)	435 (173)	504 (204)	553 (225)
No stance marker + <i>kejian</i>	287 (130)	422 (184)	510 (216)	560 (256)
Epistemic stance marker + <i>suoyi</i>	274 (81)	477 (248)	466 (192)	571 (255)
Epistemic stance marker + <i>kejian</i>	303 (106)	432 (182)	494 (212)	575 (233)
Attitudinal stance marker + <i>suoyi</i>	338 (187)	439 (193)	488 (228)	631 (250)
Attitudinal stance marker + <i>kejian</i>	322 (173)	440 (186)	500 (224)	513 (205)
<i>total fixation durations (totfix)</i>				
No stance marker + <i>suoyi</i>	317 (103)	551 (197)	636 (241)	499 (198)
No stance marker + <i>kejian</i>	329 (97)	530 (191)	598 (225)	505 (225)
Epistemic stance marker + <i>suoyi</i>	314 (88)	572 (224)	605 (254)	538 (262)
Epistemic stance marker + <i>kejian</i>	336 (88)	520 (191)	601 (261)	511 (216)
Attitudinal stance marker + <i>suoyi</i>	321 (125)	543 (219)	614 (265)	521 (232)
Attitudinal stance marker + <i>kejian</i>	325 (122)	510 (190)	605 (251)	482 (149)

Region 2: subject region

Excluding the interaction between *Connective* and *Stance* did not make the model significantly worse (fg: $\chi^2(2)=0.354, p=.838$; tg: $\chi^2(2)=0.611, p=.737$; rp: $\chi^2(2)=2.052, p=.358$; totfix: $\chi^2(2)=1.038, p=.595$). Therefore, the final models for this region only included the factors of *Connective* and *Stance* and the random factors.

The connective *kejian* led to a general facilitation compared to the connective *suoyi* across all *Stance* conditions in reading the subject region (fg: $\beta=-0.040, SE=0.020, t=-2.063, p=.039$; tg: $\beta=-0.042, SE=0.019, t=-2.185, p=.029$; totfix: $\beta=-0.068, SE=0.023, t=-2.960, p=.003$). The influence of *Stance* was not significant in this region by any of the four measures: the epistemic stance marker condition did not differ from the no stance marker condition (fg: $\beta=0.005, SE=0.024, t=0.188, p=.851$; tg: $\beta=-0.001, SE=0.024, t=-0.049, p=.961$; rp: $\beta=0.012, SE=0.027, t=0.462, p=.644$; totfix: $\beta=-0.017, SE=0.028, t=-0.608, p=.543$); the attitudinal stance marker condition was not different from the no stance marker condition either (fg: $\beta=0.030, SE=0.024, t=1.265, p=.206$; tg: $\beta=0.031, SE=0.024, t=1.292, p=.196$; rp: $\beta=0.021, SE=0.027, t=0.778, p=.437$; totfix: $\beta=-0.020, SE=0.028, t=-.721, p=.471$).

Region 3: modal verb region

At the modal verb region, excluding the interaction between *Connective* and *Stance* did not affect the model fit significantly in any of the four measures (fg: $\chi^2(2)=4.317, p=.116$; tg: $\chi^2(2)=0.804, p=.669$; rp: $\chi^2(2)=1.111, p=.574$; totfix: $\chi^2(2)=1.252, p=.535$). The final models for this region only had the factors of *Connective* and *Stance* and the random factors.

We found significant main effects of *Stance*: the reading time of the modal verb region was shorter under the epistemic stance condition compared to the no stance marker condition (tg: $\beta=-0.057, SE=0.026, t=-2.250, p=.025$; rp: $\beta=-0.060, SE=0.028, t=-2.134, p=.033$). Such a facilitation effect was also found for the attitudinal stance marker condition compared to the no stance marker condition (tg: $\beta=-0.052, SE=0.026, t=-2.019, p=.044$). When taking the epistemic stance marker condition as the baseline, we found no significant difference between the attitudinal stance marker condition and the epistemic stance marker condition (fg: $\beta=0.010, SE=0.030, t=-0.471, p=.638$; tg: $\beta=-0.009, SE=0.025, t=-0.346, p=.729$; rp: $\beta=0.007, SE=0.031, t=0.225, p=.822$; totfix: $\beta=-0.009, SE=0.029, t=-0.313, p=.754$). No significant main effects of *Connective* were observed in this region (fg: $\beta=-0.001, SE=0.021, t=-0.057, p=.955$; tg: $\beta=0.009, SE=0.021, t=0.448, p=.654$; rp: $\beta=0.011, SE=0.023, t=0.493, p=.622$; totfix: $\beta=-0.039, SE=0.023, t=-1.727, p=.084$).

Region 4: final region

For the final region, excluding the interaction between *Connective* and *Stance* had a significant effect on model fit in measures tg ($\chi^2(2)=6.288, p=.043$) and rp ($\chi^2(2)=9.284, p=.010$), but not in g ($\chi^2(2)=2.365, p=.307$) or totfix ($\chi^2(2)=0.351$,

$p=.839$). Hence, the final analytical model for rp and tg measures included the factor of *Connective* and *Stance*, as well as their interaction. The models for fg and totfix measures only included *Connective* and *Stance*.

The effects of *Connective* under the no stance marker condition were not significantly different from those under the attitudinal stance condition (tg: $\beta=-0.070$, $SE=0.047$, $t=-1.497$, $p=.134$; rp: $\beta=-0.118$, $SE=0.062$, $t=-1.919$, $p=.055$), or under the epistemic stance condition (tg: $\beta=0.048$, $SE=0.047$, $t=1.008$, $p=.313$; rp: $\beta=0.068$, $SE=0.062$, $t=1.108$, $p=.268$). In other words, the interaction effects of *Connective* and *Stance* neither lay in the contrast between the no stance marker condition versus the attitudinal stance marker condition, nor in the contrast between no stance versus epistemic stance. However, the interesting interaction effect of *Connective* and *Stance* on the model fit suggests an interaction elsewhere. Therefore, we relevelled the model by using the attitudinal stance marker condition as the baseline, so that we could directly compare the attitudinal stance marker and the epistemic stance marker condition.

When the connective was *suoyi*, the baseline of *Connective*, compared to the attitudinal stance marker condition, epistemic stance markers led to shorter processing times in general (rp: $\beta=-0.092$, $SE=0.044$, $t=-2.108$, $p=.035$). However, the interaction effect showed a different direction of reading time difference between *kejian* and *suoyi* under the attitudinal stance condition compared to that under the epistemic stance condition (tg: $\beta=0.118$, $SE=0.047$, $t=2.490$, $p=.013$; rp: $\beta=0.187$, $SE=0.062$, $t=3.009$, $p=.003$). When the stance marking was attitudinal, compared to *suoyi*, *kejian* led to a reading facilitation at the final region (tg: $\beta=-0.091$, $SE=0.033$, $t=-2.736$, $p=.006$; rp: $\beta=-0.148$, $SE=0.044$, $t=-3.396$, $p=.001$); while when the stance marker was epistemic, there was no difference between *suoyi* and *kejian* in reading times (as illustrated in Figure 1-1 and Figure 1-2).

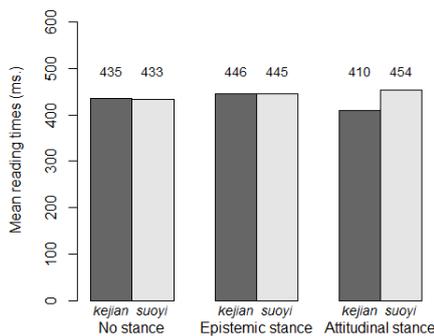


Figure 1-1. Mean reading times of the final region (tg)

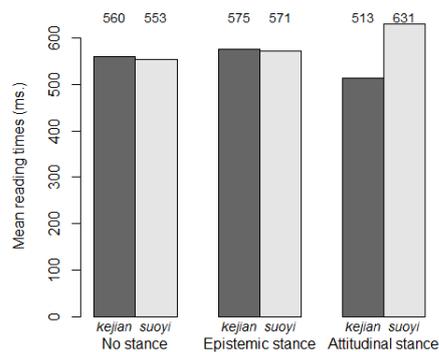


Figure 1-2. Mean reading times of the final region (rp)

5.5 Discussion

This reading study addressed the processing of subjectivity and the role of different markers of subjectivity in sentence processing. We hypothesized that the processing of subjectivity involves a tracking of the source of information and that linguistic cues such as connectives, perspective markers and modal verbs function as processing instructions that help the reader determine what the source of the information is.

5.5.1 Effects of connectives

As a starting point, we tested the effects of connectives at three regions: connective, subject and modal verb when no stance markers were available in the previous context. Our data confirmed the effect of subjective connectives in argument-claim causal relations with Chinese – the subjective connective *kejian* ‘so’ led to an immediate processing delay compared to the underspecified connective *suoyi* at the connective region. This result is comparable to the findings by Canestrelli et al. (2013) with *want* ‘because’ expressing claim-argument causal relations. The processing delay supported previous claims that subjective relations are more difficult to process. Subjective connectives triggered the establishment of subjective relations, which requires extra cognitive efforts: this was found in both Dutch and Chinese, and in both the claim-argument relation and the argument-claim relation.

We found a facilitation effect of the subjective connective *kejian* compared to the underspecified connective *suoyi* at the subject region. In previous reading experiments, when the connective was underspecified, an effect of subjectivity was found at a later region – the pre-final region in Traxler et al. (1997a) and the final region in Li et al. (2017). In the present experiment, we found an effect of the specific subjective connective in the same direction, but at a position where readers could not yet know from the content of the sentence that the relation was subjective. One explanation is that the effect is due to parafoveal processing of the modal verb: the shorter reading times under the *kejian* condition, which have been found across languages, can be due to the facilitation effect of subjective connectives in providing processing cues of subjectivity in advance. There is, however, an alternative explanation: the processing asymmetry may be due to a delaying effect of the underspecified connective – *suoyi* allows both subjective and objective interpretations, while *kejian* only allows a subjective interpretation. Under the *suoyi* condition, readers need to maintain multiple interpretations when processing the propositional content, which may increase cognitive processing load. In other words, it may not be the subjective connective *kejian* that facilitates processing, but rather the underspecified connective *suoyi* slows down processing.

At the modal verb region, readers should be able to realize the second clause is

a claim based on the argument made in the first clause, i.e. the relation is subjective. Hypothetically, the processing time of the modal verb region should be shorter under the *kejian* condition compared to the *suoyi* condition because *kejian* established a subjective source of information at the start of the second clause. However, we did not find any effects of connective choice at the modal verb region under the no stance marker conditions – the subjective connective *kejian* did not make the processing of the later modal verb region easier in comparison to the *suoyi* condition. One explanation from the view of parafoveal processing might be that the readers processed the modal verb already at the subject position. If this was the case, the processing asymmetry at the subject region demonstrated the effect of connectives in processing subjective relations.

5.5.2 Effects of stance markers

Given the collocation patterns found in the corpus-based study, we predicted an interplay between connectives and different markers of subjectivity in processing subjectivity in causal relations. The first research question was:

How does the presence of perspective markers affect the processing patterns at the connective region and at later regions in subjective relations with an argument-claim structure?

At the first two regions of S2 – the connective and the subject, we find very limited effects of *Stance*. The type of stance marking hardly influences the processing of these regions. Moreover, there is no interaction effect of *Stance* and *Connective* at this stage. The subjective connective requires longer reading time at the connective region regardless of whether a stance marker is available or not, and regardless of which type of stance marker is used. The facilitation effect of a subjective connective on the subject region after the connective also holds for all three stance marking conditions. The lack of interaction effects between stance and connectives may be due to the argument-claim structures in the current experiment. The processing delay associated with subjective connectives compared to objective/underspecified connectives is about establishing a source for the subjective claim in the second segment. Epistemic stance markers influenced the processing asymmetry between subjective connectives and objective connectives in the claim-argument type of subjective relations (Canestrelli et al., 2013), because epistemic stance markers explicitly mark the first segment as a claim made by an intentional mind. However, in the current study, the stance markers only had scope over the first segment – the argument, and did not have scope over the second clause. Therefore, the stance markers did not mark the second segment as a claim. This might explain why their influence on the processing of the connective region and the subject region is limited. It would be interesting to find out how stance markers would affect Chinese claim-argument relations marked with

connectives differing in subjectivity.

At the modal verb region there was a main effect of stance markers: the reading time of the modal verb region was significantly reduced by the presence of epistemic stance markers as well as attitudinal stance markers in the preceding clause, in comparison to the no stance marker condition. This finding can be explained in terms of the representation of subjectivity. With either an epistemic stance marker or an attitudinal stance marker in the preceding context, a source of information can be established at the beginning of the first clause, and the subjectivity information encoded by the modal verb can be linked to this pre-constructed source of information (see Section 1). Accordingly, no extra effort to establish a new source of information is needed at the modal verb for either the attitudinal stance marker condition or the epistemic stance marker condition. Even though stance markers did not have scope over the second segment of the sentence, they put an extra emphasis on the source of information. Such emphasis may have made the source of information more accessible compared to the no stance condition, thereby alleviating the efforts to attribute the second segment to the same source of information.

Do these findings mean that the two types of stance markers are the same in influencing the processing of subjectivity? In the second research question, we asked:

Do the effects of epistemic stance markers on the processing of subjective relations also hold for attitudinal stance markers?

For this research question, we expected a difference between epistemic and attitudinal stance markers in their processing effects and also in their influence on the processing of *kejian* and *suoyi*, because the two stance markers have different collocation patterns with connectives in language use: epistemic stance markers patterned with *suoyi* while attitudinal stance markers patterned with *kejian*.

The effect of attitudinal stance markers did not differ from that of epistemic stance markers at the modal verb region. Both types of markers facilitated the processing of the modal verb in subjective relations regardless of the connective type, which indicated that they both helped readers establish subjectivity in discourse relations. However, the effect of attitudinal stance markers differed from that of the epistemic stance markers at the final region, which partially reflected the differences between stance markers in collocation patterns with connectives. The interaction effects of *Stance* and *Connective* exhibited an influence of stance marker type on the processing asymmetry of the two connectives. When there was no stance marker in the context, the reading times of the final region were not significantly different between the *suoyi* condition and the *kejian* condition. The similarity in processing times between the two conditions makes sense because at this final region, readers are supposed to know that the relation is subjective – even in the condition with the smallest amount of linguistic instruction on subjectivity (no stance marker +

underspecified *suoyi* condition), since the modal verb provides the information of subjectivity. By the final region, readers have been well informed that the second clause is a claim and that the relation is formulated in an argument-claim structure. Therefore, no extra processing time is needed under the *suoyi* condition to process the final region.

At the final region, a difference was observed between epistemic stance markers and attitudinal stance markers in influencing the processing effects of different connectives. At this final stage of sentence processing, readers have to understand the structure of the relation – how the argument supports the subjective claim, for instance. The differences between the two types of stance markers can be explained from the perspective of the factuality of the argument. With an epistemic stance marker in the context such as (14a), clauses connected by *suoyi* and those connected by *kejian* were also read at a similar speed at the final region. *Jushuo* ‘it is said’ marks the first segment as information from a source – someone reporting what he/she has heard about the *one-month vacation of Liu Yushan*. But *jushuo* ‘it is said’ does not mark the first segment as a fact. On the contrary, the factuality of the first segment in (14a) is put in doubt by *jushuo*, because it is presented as an event described by someone. If the factuality of the first segment is put in doubt, the conclusions drawn on the basis of that information are also doubtful, and that is probably why readers did not benefit from the presence of epistemic stance markers at this stage of processing.

By contrast, there was a processing advantage of the subjective connective *kejian* compared to *suoyi* in combination with attitudinal stance markers such as in (14b). Attitudinal stance markers introduce the source of information by expressing the speaker’s attitude towards the content. With an attitudinal stance marker such as *surprisingly* in (14b) in the first segment, the content of the first segment is assumed to be factual. In an argument-claim relation, the factuality of the argument makes the claim stronger. This might explain why when attitudinal stance markers were used in combination with *kejian*, the relation was probably most accessible to readers.

(14)

- a. Liu Yishan **jushuo** meinian you yi zhengge yue de daixin nianjia, **suoyi/kejian** ta de gongsi yiding gei yuangong de fuli bucuo.
It is said (that) Liu Yishan has one whole month vacation with salary every year, **suoyi/kejian ‘so’** her company must provide good welfare for its employees.
- b. Liu Yishan **jingran** meinian you yi zhengge yue de daixin nianjia, **suoyi/kejian** ta de gongsi yiding gei yuangong de fuli bucuo.
Surprisingly Liu Yishan has one whole month vacation with salary every year, **suoyi/kejian ‘so’** her company must provide good welfare for its employees.

The facilitation effect of attitudinal stance markers in combination with *kejian* is in line with the collocation patterns in the corpus-based study presented in Chapter 3, which demonstrated that *kejian* co-occurs more often with attitudinal stance markers than *suoyi*.

5.5.3 Incremental sentence processing

The on-line reading of subjective relations marked by connectives, modal verbs and stance markers exhibits an incremental process of sentence comprehension. One of the important pieces of information that readers need to process to comprehend subjective relations is the source of information. To establish the source of the information and keep track of this source, readers utilize different linguistic cues throughout the sentence. Connectives set up a relation between segments. Some connectives also specify the degree of subjectivity of the relation by attributing the relation to a locutionary agent and emphasize a claim-argument or argument-claim structure of the subjective relation, such as *want* in (11b) and *kejian* in (14). Modal verbs such as *yiding* ‘must’ and *keneng* ‘may’ explicitly mark the second segment as a claim that is attributed to the source of information. Stance markers not only express *that* a source of information is involved, but also *how* the source of information is involved: the source can be involved in expressing opinions as in (14a), or attitudes or feelings as in (14b).

These linguistic cues function as processing indicators by incrementally providing cues on the source of information. Subjective connectives immediately triggered the construction of subjective relations, which requires more cognitive efforts because it requires the reader to establish a source of information. The modal verbs explicitly inform readers on the subjectivity of the relation independent of the choice of connectives. This is the point where the stance markers in the context exhibited an effect: the processing of modal verbs was facilitated by both the epistemic and attitudinal stance markers irrespective of the connective type. This finding showed that, as long as a source of information is established in the preceding context, no matter how this source of information is involved – epistemically or attitudinally – readers can benefit from the previously established source of information in processing a subjective relation. At the end of the sentence, readers wrapped up all relevant information incrementally provided in the sentence, including the information of how a source of information is involved. This information affected sentence processing because it instructs readers how well the argument supports the claim in a subjective relation. If the speaker/author uses an epistemic marker to express uncertainty about the proposition in the first clause, this also weakens the strength of the argument-claim relation as a whole.

The collocation patterns in language use have shown that the speaker/author’s pragmatic strategy is in line with Horn’s *Relation* and *Quality* principle. If the speaker

uses a certain combination of expressions (i.e. collocations) for the purpose of the hearer economy, hearers/readers should benefit from these expressions, as should be reflected in reading times. Parts of the patterns in language use were indeed reflected in on-line processing. For instance, at the end of the sentence, readers were sensitive to the co-occurrence tendency of *kejian* and attitudinal stance markers. However, at other points the collocation patterns were not reflected in the reading times. People did not process modal verbs differently in response to different connectives in the context, which was predicted on the basis of the distribution patterns of modal verbs and connectives. Explorations on collocation patterns are useful to obtain general distributional information on the use of linguistic cues expressing subjectivity. But experimental research is needed to investigate how the distributional information is utilized by comprehenders in processing, and to what extent they are sensitive to such information.

The current study shows that readers make use of the information provided by different markers of subjectivity to construct a discourse representation that encodes subjectivity: they are sensitive to information on not only *whether* a source of information is involved, but also *how* this source is involved. Readers incrementally process subjectivity on the basis of the information provided by these linguistic cues – they keep track of the source of information. Meanwhile, by evaluating the factuality of the event described in the first clause, readers also evaluate how well the argument functions as support for the claim formulated in the subjective relation. These processes in the mental representation of subjective relations are reflected by on-line reading times.

Chapter 6. General discussion and conclusion

6.1 Subjectivity in causal relations

Subjectivity is an important notion in both language use and language comprehension. Speakers/authors express their opinions, beliefs and feelings, and hearers/readers interpret these pieces of information as belonging to a locutionary agent. Subjectivity is often defined as the involvement of a speaker in the utterance (Finegan, 1995). In this thesis, subjectivity in coherence relations has been studied. In terms of causal discourse relations, the degree of subjectivity defines two basic types of relations: objective relations connect real-world events, while subjective relations connect arguments with claims or speech acts made by a speaker, the author, a discourse character or some other locutionary agent (Sanders, 1997; Sweetser, 1990). To interpret subjectivity, comprehenders need to take the presence of such a *Subject of Consciousness* (SoC) into account, which results in a different processing pattern of subjective relations such as (1a) compared to objective ones such as (1b) (Traxler, Sanford, Aked, & Moxey, 1997).

- (1)
- a. Susan lost her money and credit cards because she left her purse at the bus stop.
 - b. Susan was careless with money and credit cards because she left her purse at the bus stop.

(Traxler et al., 1997b: 91)

Previous studies have mainly treated subjectivity and perspective as separate notions. In this dissertation, I have proposed to use the notion of perspective as a tool to operationalize subjectivity: while subjectivity indicates *that* someone's opinion or attitude about the information is involved, perspective concerns *who* that 'someone' is, and *how* he/she is involved. When the content of an utterance is from a *Subject of Consciousness* (SoC), a higher degree of subjectivity is represented than when no such SoC is involved. SoCs can be involved along the dimension of epistemic stance – if they are expressing their personal opinions and judgments, or along the dimension of attitudinal stance – if their attitudes and feelings are expressed.

The information on the involvement of a speaker or character can be expressed by a large range of linguistic cues (Traugott, 2003). For example, in some languages connectives encode the degree of subjectivity of the relation, such as the Dutch

connectives *dus* ‘so’ and *want* ‘because’ (Pander Maat & Degand, 2001; Pander Maat & Sanders, 2000; Stukker and Sanders, 2012), and the Chinese connective *kejian* ‘so’ (Li, Evers-Vermeul and Sanders, 2013). These linguistic cues directly influence the processing of subjectivity in discourse. According to Canestrelli, Mak and Sanders (2013), the subjective connective *want* ‘because’ in (2a) leads to an immediate processing delay compared to the objective connective *omdat* ‘because’ (2b).

(2)

- a. Jeff werd boos op zijn burens, **omdat** ze altijd hun stereo hard hadden aanstaan tot diep in de nacht.
Jeff got angry with his neighbors, **because** they always had their stereo turned on till deep in the night.
- b. Jeff had asociale burens, **want** ze hadden altijd hun stereo hard aanstaan tot diep in de nacht.
Jeff had annoying neighbors, **because** they always had their stereo turned on till deep in the night.

The processing asymmetry between the subjective connective and the objective connective can be influenced by perspective markers. For instance, in (2c), when the perspective marker *Volgens Lieke* ‘according to Lieke’ marked the perspective of Lieke in the clause preceding the connective, the extra processing time of the subjective connective *want* compared to the objective connective *omdat* was cancelled out (Canestrelli et al., 2013).

(2)

- c. **Volgens Lieke** had Jeff asociale burens, **want** ze hadden altijd hun stereo hard aanstaan tot diep in de nacht.
According to Lieke, Jeff had annoying neighbors, **because** they always had their stereo turned on till deep in the night.

(Canestrelli et al., 2013: 170)

The facilitative effects of perspective markers on the processing of the subjective relation marked by the connective *want* show that the extra processing times following *want* were related to the involvement of someone’s perspective. Readers/hearers took extra cognitive efforts in processing when they needed to interpret someone else’s opinion.

However, several questions on the construction of subjectivity and the roles of linguistic cues remain unanswered. For instance, how are different linguistic cues of subjectivity distributed in language? And how do they contribute to the comprehension of subjective relations? What does subjectivity mean for the

construction of a mental representation of a sentence? The current study investigated the conceptual representations of subjectivity as well as the linguistic expressions of it.

In this dissertation, I focused on subjectivity in causal relations in Mandarin Chinese, which has a connective system with both a connective underspecified in terms of subjectivity – *suoyi* ‘so’ – and specific connectives to express subjective relations and objective relations respectively – *kejian* ‘so’ and *yin’er* ‘as a result’. This system allowed me to investigate the contextual features and the processing effects of connectives encoding different degrees of subjectivity.

6.2 Summary of main findings

Three empirical studies were performed to investigate the encoding and processing of subjectivity in discourse relations: a corpus-based collocational study, a visual world paradigm eye-tracking study and an on-line reading study. These different methods allowed me to obtain insights into this topic from various perspectives.

6.2.1 Corpus-based collocational study

Previous corpus-based studies have focused on the use of connectives in discourse and the relations they express. For instance, Degand and Pander Maat (2003) found that the Dutch connective *want* ‘because’ patterned with subjective relations while *omdat* ‘because’ was more often used to express objective relations (cf. Pit, 2003; Sanders & Spooren, 2015). A similar distinction in terms of the relation type preferred by connectives was reported by Li et al. (2013) for Chinese: *kejian* ‘so’ is typically used for subjective relations, while *yin’er* ‘so’ has a preference for objective relations; a third connective, *suoyi* ‘so’, can be used to express both types of relations. However, this unidimensional approach of focusing on connectives alone does not provide information on how different linguistic cues are used to express subjectivity collaboratively.

We started an explorative collocational study to investigate the distribution of linguistic cues expressing subjectivity. We directly compared the contexts of two Chinese connectives expressing different degrees of subjectivity – the specified subjective connective *kejian* ‘so’ and the generic, underspecified connective *suoyi* ‘so’. The degree of subjectivity they encode was expected to correlate with the contextual features of the two connectives. With this distinctive collocational analysis, we aimed to address the following research questions:

- RQ1. Do connectives of different subjectivity degrees differ in their types of collocates?

- RQ2. More specifically, do connectives differ in the types of perspective markers they co-occur with?

By directly comparing the context of *kejian* with that of *suoyi*, we found pronouns, modal verbs (e.g. *keneng* ‘may’) and cognition verbs (e.g. *renwei* ‘consider’, *zhidao* ‘know’) consistently patterning with the underspecified connective *suoyi* across genres. Expressions of surprisal (e.g. *jingran* ‘surprisingly’) and importance (e.g. *zhongyao* ‘important’, *jiazhi* ‘value’) were found to be the distinctive collocates of the subjective connective *kejian*. Connectives encoding different degrees of subjectivity differ in the type of collocates – specifically in the distribution of perspective/stance markers (Conrad & Biber, 2000). *Suoyi*, the underspecified connective, patterned with epistemic stance markers, which express the source of evidence and the (un)certainly of the speaker. However, the collocates of *kejian* were indicators of attitudinal stance, which express the speaker’s attitudes and emotions. The influence of different perspective markers on the processing of subjective relations was tested in a follow-up reading experiment.

6.2.2 Visual world paradigm eye-tracking study

From a cognitive point of view, it has been argued that an SoC is involved in the representation of subjective causal relations constructed by comprehenders (Sanders, Sanders, & Sweetser, 2012). In terms of situation models, the question is: does the relative complexity of subjectivity involve the tracking of a locutionary agent as the source of information? Previous reading studies have found a delay associated with the processing of subjective relations (Canestrelli et al., 2013; Traxler et al., 1997a; Traxler et al., 1997b). A visual world paradigm (VWP) eye-tracking experiment was run to answer the question whether the longer reading times in processing subjectivity are related to the tracking of the information source in the situation model, which may take extra cognitive efforts. The study aimed to explore the mental representation behind the processing of subjectivity, on the basis of the following research questions:

- RQ3. Does the processing of subjectivity involve tracking of the sources of information?
- RQ4. What is the role of connectives and modal verbs in tracking the information source in situation models?

The hypothesis that the processing of subjectivity involves a locutionary agent or SoC as the source of information was tested in two VWP studies on Dutch and Chinese. The results show that people had a tendency to look at a visual display of a speaker, the relevant SoC in this experiment, when they were processing subjectivity in causal relations, and that this tendency was guided by linguistic cues. In both languages,

there was increased attention to the picture with a speaker after subjective connectives (*dus* ‘so’ and *kejian* ‘so’) in comparison to objective connectives (*daardoor* ‘as a result’ and *yin’er* ‘as a result’). When the connective was underspecified (*suoyi* ‘so’) in terms of subjectivity, a modal verb (*yiding* ‘must’/*keneng* ‘may’) in the second clause guided more attention to the speaker than it did under the condition with a specific subjective connective. In other words, the modal verb took the role of guiding attention to the speaker if the connective did not provide sufficient information on the degree of subjectivity beforehand.

These VWP results showed that the processing of subjectivity involves the tracking of the source of information in the mental representation. The identification and tracking of the source of information corresponded to the longer processing times associated with subjectivity that were found in the aforementioned reading studies. The findings thus confirmed previous explanations of the difficulty of subjective relations in terms of tracking the source of information: more attention was paid to the source of information when processing subjective information. Moreover, in this tracking process, linguistic cues expressing subjectivity functioned as processing instructions in guiding people’s attention to the SoC.

6.2.3 On-line reading study

A further question was whether in language comprehension readers are sensitive to the different combinations of connectives and some of the other linguistic markers of subjectivity that surfaced in the collocational study. We performed a follow-up reading experiment to investigate the role of perspective markers and connectives in on-line processing, investigating the following two research questions:

- RQ5. How does the presence of perspective markers affect the processing patterns at the connective region and at later regions in subjective relations with an argument-claim structure?
- RQ6. Do the effects of epistemic stance markers on the processing of subjective relations also hold for attitudinal stance markers?

In a reading experiment, subjective sentences were constructed, in which the type of stance markers (no stance marker/ epistemic stance marker/ attitudinal stance marker) and the type of connectives (*kejian*/ *suoyi*) were manipulated. All sentences contained a modal verb in the second clause to make sure that this second clause was interpreted as a claim, and the causal relation was a subjective one.

Consistent with the predictions, the connective encoding a higher degree of subjectivity resulted in an immediate processing delay. Longer reading times were found at the subjective connective *kejian* ‘so’ compared to the underspecified connective *suoyi* ‘so’, which was in line with Canestrelli et al.’s (2013) findings on

the Dutch subjective connective *want* ‘because’. This delaying effect of the subjective connective could be attributed to its role in triggering a process of tracking a source of information. At the subject region of the second clause, however, *kejian* showed a facilitation effect compared to *suoyi*. This facilitation may be due to the effect of *kejian* in saving readers the efforts of maintaining multiple interpretations of the relation.

The presence of perspective markers did not change the processing patterns at the connective, but they reduced the reading time of the modal verb in the second clause, compared to the no stance marker condition. At the modal verb region, when readers got an explicit cue that the relation should be subjective, they benefited from the presence of perspective markers at the beginning of the first clause. This finding was important to our understanding of the comprehension of subjectivity: perspective markers either referring to the opinion of a source of information or implying the source by involving his/her attitude can function as processing instructions – speeding up the processing of modal verbs at the second clause of the subjective relation.

Attitudinal stance markers and epistemic stance markers did not differ in influencing the processing of modal verbs. However, an interaction effect of type of perspective marker and connective was found at the final region of the second clause: the reading times of this region under the *kejian*-condition and the *suoyi*-condition were the same when the stance marker was epistemic, while they were much shorter under the *kejian*-condition than under the *suoyi*-condition when the stance marker was attitudinal. In other words, the combination of *kejian* and attitudinal stance markers facilitated reading the most.

This interaction effect of type of perspective marker and connective at the end of the second clause was consistent with the collocational patterns in the corpus data. As illustrated in example (3), the attitudinal stance marker *juran* ‘unexpectedly’ co-occurred frequently with *kejian* in natural language data, and exactly this combination of use led to faster reading times at the end of the second clause compared to other combinations.

(3)

Chengji tekua *Ø/tingshuo/juran* wudian le yige duo xiaoshi, *suoyi/kejian* tielu xitong *keneng* chu le shigu. Zhezhong qingkuang bing bu changjian. Intercity express *Ø/it is said/unexpectedly* is:delayed ASP(PFV) one more hour, *suoyi/kejian (CONJ)* railway system *may* happen ASP(PFV) accident. This:kind situation EMP NEG common.
Ø/ It is said (that)/ unexpectedly the intercity express is delayed for more than one hour, *suoyi/kejian* the railway system *may* have encountered an accident. This kind of situation is not common.

Previous reading studies have shown that the involvement of a source of information (subjectivity) affects the comprehension of discourse relations. This study contributed to the field by showing that the way in which a source of information is involved (i.e. the specific dimension of subjectivity) also influenced readers when they are processing subjective relations.

6.3 Theoretical issues and future studies

Findings from three different methodologies (i.e. corpus-based research, visual world paradigm processing and on-line reading) depict a refined picture of subjectivity. This informs us about several theoretical issues such as the mental representation of subjectivity, the linguistic expression of subjectivity in language use and the roles of these linguistic expressions in language processing.

6.3.1 Subjectivity and perspective

The empirical evidence from the VWP supported the hypothesis that a tracking of the source of information is involved in the mental representation of subjectivity. The tracking of the source of information requires comprehenders to ground the information in an utterance in the perspective of a speaker or another character who is responsible for the subjective relation.

The collocational patterns and the on-line reading times also brought new insights into the relation between subjectivity and perspective. Perspective can be used as a tool to operationalize subjectivity in discourse. Perspective markers in discourse provided a viewpoint to investigate how subjectivity is expressed by authors/narrators and how readers/hearers comprehend subjectivity. First, whether a locutionary agent is involved determines the degree of subjectivity. If a locutionary agent is involved, for example by making a claim or an evaluation based on some arguments, the relation is more subjective than in sentences merely describing real-world events. The involvement of a subjective perspective can be indicated by linguistic cues such as perspective markers, modal verbs and subjective connectives. In language use data, we found perspective markers and modal verbs appearing in collocations with the underspecified connective. This finding could be interpreted as the tendency of the speaker to provide sufficient information when expressing the subjectivity of relations. In language comprehension, readers/hearers made use of these different linguistic cues to interpret subjectivity.

Second, perspective specifies different dimensions of subjectivity, the way in which the locutionary agent is involved. In the traditional definitions of subjectivity, the notion of subjectivity covers both subjective beliefs and viewpoints (Finegan, 1995; Lyons, 1977), and subjective attitudes (Lyons, 1977; Traugott, 1995). These

aspects of subjectivity can be expressed by linguistic expressions such as *John thinks* (viewpoints), *importantly* (attitudes), *fortunately* (emotions), *probably* (certainty), etc. In these definitions, subjectivity has been treated as one monolithic concept. However, we propose to decompose the notion of subjectivity along several dimensions that up till now have only been distinguished in the literature on perspective marking: epistemic stance, attitudinal stance and style stance (Bednarek, 2006; Conrad & Biber, 2000). The epistemic stance indicates evidentiality and speaker's (un)certainly; the attitudinal stance expresses the speaker's attitudes and emotions; and the style stance indicates how something is expressed. The on-line reading study has shown how at least two of these dimensions of subjectivity influence processing, and the collocational analyses indicated different patterns in language use that could be accounted for in terms of these dimensions. Our empirical findings supported a point of view that differentiates these dimensions of subjectivity. For instance, when an attitudinal stance is involved such as in example (4a), with the combination of an attitudinal stance marker and *kejian* – a collocation pattern in language use, readers processed the sentences faster compared to (4b).

- (4)
- a. Fortunately, this fire didn't cause any injuries or deaths, so the fire department must have organized the rescue in time.
 - b. Truly, this fire didn't cause any injuries or deaths, so the fire department must have organized the rescue in time.

Style stance markers, which indicate how something is spoken or written also express subjectivity, did not surface in the collocational results, but still need to be further studied and compared with the other two dimensions.

6.3.2 Pragmatic principles in the use of connectives and perspective markers

The use of connectives and perspective markers in our corpus-based study was consistent with Horn's *Quality* and *Relation* principle, as well as the *Uniform Information Density Theory* (UID). From the perspective of the *Relation principle* (speaker economy: reducing the speaker's production effort), when the subjective connective *kejian* was used to express the epistemic dimension of subjectivity, epistemic stance markers in the context of the connective were redundant, i.e. not efficient from the speaker economy account. When an underspecified connective like *suoyi* was used, the epistemic dimension of subjectivity was not specified by the connective. In this case, epistemic stance markers in the context provided sufficient information for readers/hearers to interpret subjectivity – the *Quality principle* (hearer economy: reducing the hearer's comprehension effort) could be observed.

Accordingly, the comprehension process of hearers/readers' comprehension should be facilitated.

Similar to the epistemic stance markers and the subjective connective *kejian*, attitudinal stance markers (e.g. *jingran* 'surprisingly' and *zhongyao* 'important'), also indicate that a locutionary agent is involved. However, attitudinal stance markers express subjectivity in a different dimension – they express the attitude or feelings of the locutionary agent towards the information. In other words, attitudinal stance markers do not overlap with *kejian* in the epistemic dimension of subjectivity. This explains why the combined use of attitudinal stance markers and *kejian* does not go against Horn's *Relation* principle.

The collocation results can also be well explained by the UID theory, which suggests that the information density of an utterance should be kept uniform – without too much overlap in meaning (Frank & Jaeger, 2008; Jaeger, 2010; Levy & Jaeger, 2007). In the context with epistemic stance markers (e.g. cognition verbs, modal verbs), the content on the epistemic dimension of subjectivity is highly expectable (high probability and low information). Therefore, it is more likely to have an underspecified connective, *suoyi*. By contrast, utterances with fewer occurrences of epistemic stance markers make the content conveyed by the context unexpected (low probability and high information). In this case, the use of a specific connective *kejian* is preferred. The prevalence of epistemic stance markers in the context of *suoyi* and their lower co-occurrence with *kejian* kept the uniform information density throughout the sentence in terms of subjectivity.

6.3.3 Incremental processing of subjectivity

In the literature, there are competing hypotheses for the on-line processing of discourse relations. The delayed-integration hypothesis suggests that interpretation of relations is delayed till the end of the second clause when the representations of the two clauses are integrated (Kintsch & van Dijk, 1978; Millis & Just, 1994). However, the recent insights into language processing supported an incremental processing hypothesis, which proposes that comprehenders immediately compute the relation as the content is unfolded (Canestrelli et al., 2013; Cozijn, Noordman, & Vonk, 2011; Koornneef & Sanders, 2013; Traxler et al., 1997a; van Silfhout, Evers-Vermeul, Mak, & Sanders, 2014). The question is how linguistic markers of subjectivity earlier in the sentence influence processing effects of subjectivity markers later on in the sentence.

Just like previous studies, our on-line reading study confirmed the incremental hypothesis of sentence processing: the processing of subjectivity was not delayed till the end of the sentence, but rather achieved incrementally as the sentence unfolded. The processing was guided by linguistic cues such as connectives and modal verbs. Every cue of subjectivity throughout the sentence immediately contributed to the mental representation and hence to the processing of new incoming material. For

instance, in example (3), when the subjective relation is expressed by *kejian*, the specific subjective connective, an immediate processing delay followed compared to the *suoyi*-condition. The processing delay was attributed to the interpretation of subjectivity information in the relation: as soon as readers knew the relation is subjective, they started constructing a mental representation that includes a source of information.

Importantly, we found empirical evidence for the incremental processing of subjectivity of causal relations in the visual world paradigm experiment. The subjective connectives in both Dutch and Chinese immediately triggered a process of tracking the source of information. The modal verbs in Chinese exhibited a similar function of directing attention to a visual display of the speaker when they were the first cues expressing subjectivity in the sentence. In other words, as soon as the comprehenders were clearly informed that the relation was subjective, either by the connective or by the modal verb, they started to construct the source of information in these mental representations in order to process subjectivity.

In addition, our studies on both on-line reading and VWP showed that the processing of different linguistic markers of subjectivity was influenced by other markers. In the processing experiments, we systematically examined the situations where different linguistic markers provided similar information of subjectivity. It was clearly shown that when the subjectivity information encoded in a linguistic element was provided by other elements in the preceding context and had already been processed, the processing of this element as a consecutive marker was affected. For instance, when there was a marker of subjectivity in the context – a perspective marker such as *jushuo* ‘it is said’ or a subjective connective *kejian*, the processing of modal verb in the following context was affected, in both on-line reading and VWP.

With no perspective markers in the context, the modal verbs *keneng* ‘may’ later on in the second clause functioned as processing cues that immediately triggered the interpretation of subjectivity information: longer reading times were taken to process the modal verb compared to the situations with perspective markers.

6.3.4 Insights from research on subjectivity in Chinese

From a methodological point of view, converging evidence from different methods (collocational analysis, on-line reading and VWP) and the study of Chinese as the object enabled me to address several important issues in discourse representations of subjectivity and causality. The corpus-based collocational analysis showed how linguistic cues expressed subjectivity in discourse relations in combination with connectives. The on-line reading study examined how these collocation patterns found in language use influenced the processing of subjective relations. Reading times in the on-line study reflected the cognitive efforts people spent on processing subjectivity. And with the VWP study, it was possible to explore what is actually going on in

people's mind while they are processing information, which helped explain differences in reading times between conditions.

Studying the lexicon of Chinese causal connectives has provided us with a more precise view on the encoding and processing of subjectivity. Previous studies have looked at subjective relation versus objective relations or at subjective connectives versus objective connectives. In the current study of Chinese, which has a connective that is underspecified for subjectivity as well as specific subjective/objective connectives in the lexicon, the effects of specific connectives can be compared to those of the underspecified connective.

By directly comparing the context of the specific subjective connective and the context of the underspecified connective in the corpus, a list of perspective markers that express subjectivity in combination with connectives surfaced in the collocational study, including epistemic stance markers (e.g. modal verbs and cognition verbs) and attitudinal stance markers (e.g. expressions of surprisal and importance). These types of perspective markers in combination with different connectives were then examined in an on-line reading study, which manifested the incremental nature of the processing of subjectivity.

Comparing the processing effects of specific subjective and objective connectives with the underspecified connective in processing provided a clear picture on the roles of different connectives: the subjective connective *kejian* initiated immediate processing delay and led to increased attention to the SoC compared to the underspecified connective; the objective connective *yin'er* 'as a result' guided away comprehenders' attention from the SoC. The VWP study tested the theoretical hypothesis on the comprehension of subjectivity in discourse: the processing of subjectivity involved the tracking of the source of information. This finding underscores the importance of subjectivity in the mental representation of discourse relations.

The investigation on the use and processing of connectives can be extended in several aspects for future research. More studies are needed to investigate the relationship between different types of connectives and perspective markers, and how they are used in discourse. The collocational study in Chapter 3 showed that when a connective did not provide cues of subjectivity in the epistemic dimension, other perspective markers were likely to supplement the lack of information. For instance, it is an interesting topic for further research what the collocates are of connectives marking other types of information (e.g. volitionality, temporality, etc.).

In the collocational analysis, as an explorative study, the *suoyi* sentences expressing a subjective relation were not separated from *suoyi* sentences expressing an objective relation. Therefore, the results may have been confounded by the features of objective relations. That is to say, some of the contextual features of *suoyi* compared to *kejian* could be due to the characteristics of objective relations expressed by *suoyi*. To get more refined results, we need future studies with an encoding of the exact types

of relations *suoyi* expresses.

In the Chinese reading experiment, only non-volitional causal sentences as a prototypical type of objective relations were included. However, objective relations expressed by *suoyi* in the corpus can also be volitional, which was not examined in our experiments. Different from the non-volitional causal relation, which does not involve any intentions, a volitional causal relation involves a subject that performs actions. The relations with such an actor may be expressed and processed differently in discourse compared to the non-volitional relations. Future experimental studies may better address the distinctions in processing between the two types of objective relations.

6.4 Conclusion

In this dissertation, we have first identified the characteristics of Chinese connectives and perspective markers as linguistic cues of subjectivity. The corpus-based collocational analysis illustrated how linguistic cues were used in combination to express subjectivity in language use data. In the processing studies, we have shown the role of these linguistic cues as processing instructions in both on-line reading and VWP. These results contribute to the overall picture that cannot be built without the converging evidence from different methods. Applying this cognitive approach of using converging evidence has brought fruitful insights in this field of research. Moreover, from the abovementioned studies, many questions have been identified that can be the topic of future research in Chinese and other languages regarding subjectivity and causality in discourse.

References

- Andersson, M. (2016). *The architecture of result relations: Corpus and experimental approaches to result coherence relations in English* (Unpublished doctoral dissertation). Stockholm University, Sweden.
- Anthony, L. (2016). AntConc (Version 3.4.4.0) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from <http://www.laurenceanthony.net/>
- Asr, F. T., & Demberg, V. (2015). Uniform information density at the level of discourse relations: Negation markers and discourse connective omission. In *Proceedings of the 11th International Conference on Computational Semantics* (pp. 118–128). London: Association for Computational Linguistics.
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412.
- Bartsch, S. (2004). *Structural and functional properties of collocations in English: A corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Tübingen: Gunter Narr Verlag.
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.
- Bednarek, M. (2006). *Evaluation in media discourse: Analysis of a newspaper corpus*. London: Continuum.
- Bednarek, M. (2009). Dimensions of evaluation: Cognitive and linguistic perspectives. *Pragmatics & Cognition*, 17(1), 146–175.
- Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243–257.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.
- Boogaart, R., Colleman, T., & Rutten, G. (2014). Constructions all the way everywhere: Four new directions in constructionist research. In R. Boogaart, T. Colleman, & G. Rutten, (Eds.), *Extending the scope of construction grammar* (pp. 1–14). Berlin: Mouton de Gruyter.
- Bower, G. H. (1989). Mental models in text understanding. In F. A. Bennett & K. M. McConkey (Eds.), *Cognition in individual and social contexts* (pp. 129–144). Amsterdam: Elsevier.
- Britton, B. K. (1994). Understanding expository text: Building mental structures to induce insights. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (1st ed., pp. 641–674). San Diego, CA: Academic Press.
- Cain, K., & Nash, H. M. (2011). The influence of connectives on young readers'

- processing and comprehension of text. *Journal of Educational Psychology*, 103(2), 429–441.
- Canestrelli, A. R., Mak, W. M., & Sanders, T. J. M. (2013). Causal connectives in discourse processing: How differences in subjectivity are reflected in eye-movements. *Language and Cognitive Processes*, 28(9), 1394–1413.
- Canestrelli, A. R., Mak, W. M., & Sanders, T. J. M. (2016). The influence of genre on the processing of objective and subjective causal relations: Evidence from eye-tracking. In N. Stukker, W. P. M. S. Spooren, & G. Steen (Eds.), *Genre in language, discourse and cognition* (pp. 51–73). Berlin: De Gruyter Mouton.
- Carlson, L., & Marcu, D. (2001). *Discourse tagging reference manual* (ISI technical report. ISI-TR-545). Retrieved from: <http://www.isi.edu/~marcu/discourse/>
- Center for Chinese Linguistics, Peking University (2003). *CCL corpus* (web version). Retrieved from: http://ccl.pku.edu.cn:8080/ccl_corpus/index.jsp?dir=xiandai
- Church, K., Gale, W., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In U. Zernik (Ed.), *Lexical acquisition: Exploiting on-line resources to build up a lexicon* (pp. 115–164). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), 22–29.
- Conrad, S., & Biber, D. (2000). Adverbial marking of stance in speech and writing. In S. Hunston & G. Thompson (Eds.), *Evaluation in text: Authorial stance and the construction of discourse* (pp. 56–73). Oxford: Cambridge University Press.
- Cozijn, R. (2000). *Integration and inference in understanding causal sentences*. Tilburg University, Tilburg. Retrieved from <https://pure.uvt.nl/portal/en/publications/integration-and-inference-in-understanding-causal-sentences>
- Cozijn, R., Noordman, L. G. M., & Vonk, W. (2011). Propositional integration and world-knowledge inference: Processes in understanding *because* sentences. *Discourse Processes*, 48(7), 475–500.
- Dancygier, B., & Sweetser, E. E. (2005). *Mental spaces in grammar: Conditional constructions*. Cambridge: Cambridge University Press.
- Degand, L., & Pander Maat, H. L. W. (2003). A contrastive study of Dutch and French causal connectives on the speaker involvement scale. In A. Verhagen & J. van de Weijer (Eds.), *Usage based approaches to Dutch* (pp. 175–199). Utrecht: LOT.
- de Kok, D., & Brouwer, H. (2011). *Natural language processing for the working programmer*. Retrieved from: <http://freecomputerbooks.com/books/nlpwp.pdf>
- Deng, Y. (2007). Guobiao ‘yinci’ he ‘yin’er’ [Analysis of the usage of ‘yinci’ and ‘yin’er’]. *Journal of Guangzhou University (Social Science Edition)*, 6(8), 79–82.
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence.

- Computational Linguistics*, 19(1), 61-74.
- Eggins, S., & Slade, D. (1997). *Analysing casual conversation*. London: Cassell.
- Evert, S. (2005). *The statistics of word cooccurrences: Word pairs and collocations* (Doctoral dissertation, University of Stuttgart). Retrieved from: <https://elib.uni-stuttgart.de>
- Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 1212–1248). Berlin: Mouton de Gruyter.
- Eyelink 1000 [Apparatus and software]. Mississauga, Canada: SR Research.
- Feng, G., & Zheng, W. (2011). Guonei zhongwen zidong fenci jishu yanjiu zongshu [Review of Chinese automatic word segmentation]. *Library and Information Service*, 55(2), 41–45.
- Finegan, E. (1995). Subjectivity and subjectivisation: An introduction. In D. Stein & S. Wright (Eds.), *Subjectivity and subjectivisation: Linguistic perspectives* (pp. 1–15). Cambridge: Cambridge University Press.
- Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis* (pp. 1–32). Oxford: Blackwell.
- Frank, A. F., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 933–938). London: Cognitive Science Society.
- Gernsbacher, M.A. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Erlbaum.
- Givón, T. (1992). The grammar of referential coherence as mental processing instructions. *Linguistics*, 30, 5–55.
- Glenberg, A. M., Meyer, M., & Lindem, K. (1987). Mental models contribute to foregrounding during text comprehension. *Journal of Memory and Language*, 26(1), 69–83.
- Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Edward Arnold.
- Graesser, A. C., Bowers, C., Olde, B., & Pomeroy, V. (1999). Who said what? Source memory for narrator and character agents in literary short stories. *Journal of Educational Psychology*, 91(2), 284–300.
- Graesser, A., & McNamara, D. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, 3(2), 371–398.
- Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology*, 48(1), 163–189.
- Graesser, A. C., Olde, B., & Klettke, B. (2002). How does the mind construct and represent stories? In M. C. Green, J. J. Strange, & T. C. Brock (Eds.), *Narrative impact: Social and cognitive foundations* (pp. 229–262). Mahwah, NJ: Lawrence Erlbaum Associates.
- Gries, S. T. (2013). 50-something years of work on collocations. *International Journal*

- of Corpus Linguistics*, 18(1), 137–166.
- Gries, S. T., & Stefanowitsch, A. (2004). Extending collocation analysis: A corpus-based perspective on “alternations”. *International Journal of Corpus Linguistics*, 9(1), 97–129.
- Gries, S. T., & Stefanowitsch, A. (2010). Cluster analysis and the identification of collexeme classes. In S. Rice & J. Newman (Eds.), *Empirical and experimental methods in cognitive/functional research* (pp. 73–90). Stanford, CA: CSLI Publications.
- Gong, S., Ahrens, K., & Huang, C. (2007). Chinese sketch engine and mapping principles: A corpus-based study of conceptual metaphors using the building source domain. In *Proceedings of the 8th Chinese Lexical Semantics Workshop* (pp. 130–136). Vancouver: Association for Computational Linguistics. doi: 10.1142/S1793840608001755
- Grosz, B., & Sidner, C. (1986). Attention, intentions, and the structure of discourse. *Journal Computational Linguistics*, 12(3), 175–204.
- Guo, J. (2008). ‘yinwei...suoyi...’ ju he ‘jiran...name...’ ju de chayi [A contrastive analysis between sentences with ‘yinwei...suoyi...’ and ‘jiran...name...’]. *Chinese Language Learning*, 3, 22–29.
- Hawkins, P. R. (1971). The syntactic location of hesitation pauses. *Language and Speech*, 14(3), 277–288.
- Hobbs, J., Stickel, M., Martin, P., & Edwards, D. (1988). Interpretation as abduction. In J. Hobbs (Ed.), *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics* (pp. 95–103). Buffalo: Association for Computational Linguistics. doi: 10.3115/982023.982035
- Horn, L. R. (1984). Toward a new taxonomy for pragmatic inference. In D. Schiffrin (Ed.), *Meaning, form, and use in context: linguistic applications* (pp. 11–42). Washington, D.C.: Georgetown University Press.
- Huang, C., Kilgarriff, A., Wu, Y., Chiu, C., Smith, S., Rychlý, P., Bai, M., & Chen, K. (2005). Chinese sketch engine and the extraction of grammatical collocations. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing* (pp. 48–55). Jeju Island, Korea: Asian Federation of Natural Language Processing.
- Huang, C., Hong, J., Ma, W., & Šimon, P. (2015). From corpus to grammar: Automatic extraction of grammatical relations from annotated corpus. *Linguistic Corpus and Corpus Linguistics in the Chinese Context, Journal of Chinese Linguistics Monograph Series*, 25, 192–221.
- Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1), 23–62.
- Järvikivi, J., Van Gompel, R. P. G., Hyönä, J., & Bertram, R. (2005). Ambiguous pronoun resolution: Contrasting the first mention and subject preference accounts. *Psychological Science*, 16(4), 260–264.

- Jiang, T. (2010). 'Jiran' de yufahua guocheng [The grammaticalization of 'jiran']. *Journal of Language and Literature Studies* 13, 102–105.
- Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.
- Kamalski, J., Sanders, T., & Lentz, L. (2008). Coherence marking, prior knowledge, and comprehension of informative and persuasive texts: Sorting things out. *Discourse Processes*, 45(4–5), 323–345.
- Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The sketch engine: Ten years on. *Lexicography*, 1(1), 7–36.
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychological Review*, 95(2), 163–182.
- Kintsch, W. (1992). How readers construct situation models for stories: The role of syntactic cues and causal inferences. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From Learning Processes to Cognitive Processes*. Essays in honor of William K. Estes (Vol. 2, pp. 261–278). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85(5), 363–394.
- Knoeferle, P., Crocker, M., Pickering, M., & Scheepers, C. (2005). The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition*, 95(1), 95–127.
- Knott, A., & Dale, R. (1994). Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18, 35–62.
- Knott, A., & Sanders, T. J. M. (1998). The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics*, 30(2), 135–175.
- Knott, A., Sanders, T. J. M., & Oberlander, J. (2001). Levels of representation in discourse relations. *Cognitive Linguistics*, 12(3), 197–209.
- Koornneef, A. W., & Sanders, T. J. M. (2013). Establishing coherence relations in discourse: The influence of implicit causality and connectives on pronoun resolution. *Language and Cognitive Processes*, 28(8), 1169–1206.
- Koring, L., Mak, W. M., & Reuland, E. J. (2012). The time course of argument reactivation revealed: Using the visual world paradigm. *Cognition*, 123(3), 361–379.
- Langacker, R. W. (1990). Subjectification. *Cognitive Linguistics*, 1(1), 5–38.
- Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt, & T. Hofmann (Eds.), *Advances in neural information processing systems* (1st ed., pp. 849–856). Cambridge, MA: MIT Press.
- Li, F. (2014). *Subjectivity in Mandarin Chinese: The meaning and use of causal*

- connectives in written discourse*. (Doctoral dissertation, Utrecht University). Utrecht: LOT. Retrieved from <http://dspace.library.uu.nl/handle/1874/297588>
- Li, F., Evers-Vermeul, J., & Sanders, T. J. M. (2013). Subjectivity and result marking in Mandarin: A corpus-based investigation. *Chinese Language and Discourse*, 4(1), 74–119.
- Li, F., Mak, W. M., Evers-Vermeul, J., & Sanders, T. J. M. (2017). On the online effects of subjectivity encoded in causal connectives. *Review of Cognitive Linguistics*, 15(1), 34–57.
- Li, F., Sanders, T.J.M., & Evers-Vermeul, J. (2016). On the subjectivity of Mandarin reason connectives: Robust profiles or genre-sensitivity? In W. Spooren, G. Steen & N. Stukker (Eds.), *Genre in language, discourse and cognition* (pp. 15–49). Berlin: Mouton de Gruyter.
- Li, J. (2011). Lun ‘youyu’ yu ‘yinwei’ de chayi [On the differences between ‘youyu’ and ‘yinwei’]. *Chinese Teaching in the World*, 25(4), 490–496.
- Li, J., & Liu, Y. (2004). ‘Youyu’ yu ‘jiran’ de zhuguanxing chayi [The differences of ‘youyu’ and ‘jiran’ in subjectivity]. *Chinese Language*, 2, 123–128.
- Liu, J., & Wei, C. (2008). Zhongwen fenci suanfa yanjiu [Arithmetic research on Chinese segmentation]. *Microcomputer Applications*, 129(18), 11–16.
- Li, S. (2012). ‘Kejian’ de biaoji gongneng he yufahua guocheng [The function of ‘kejian’ in discourse marking and the process of its grammaticalization]. *Journal of Northwest University (Philosophy and Social Sciences Edition)*, 42(3), 177–179.
- Li, X. (2009). ‘Cong’er’, ‘yin’er’ de gongneng chayi jiqi lishi jieshi [The functional differences between ‘cong’er’ and ‘yin’er’]. *Chinese Language Learning*, 1, 50–56.
- Liu, Y., & Yao, X. (2011). ‘Kejian’ de qingtaihua yu guanlianhua [On the modalization and relevantization of ‘kejian’]. *Chinese Linguistics*, 4, 27–33.
- Long, S., Zhao, Z., & Tang, H. (2009). Zhongwen fenci suanfa gaishu [Overview on Chinese segmentation algorithm]. *Computer Knowledge and Technology*, 5(10), 2605–2607.
- Lorch, R. F., & Lorch, E. P. (1986). On-line processing of summary and importance signals in reading. *Discourse Processes*, 9(4), 489–496.
- Lü, S. 1999. *Xiandai hanyu babaici* [Eight hundred words of Chinese]. Beijing: Commercial Publishing House.
- Lyons, J. (1977). *Semantics*. Cambridge: Cambridge University Press.
- Mak, W. M., & Sanders, T. J. M. (2010). Incremental discourse processing: How coherence relations influence the resolution of pronouns. In M. Everaert, T. Lentz, H. de Mulder, & O. Nilsen (Eds.), *The linguistics enterprise: From knowledge of language to knowledge in linguistics* (pp. 167–182). Amsterdam: John Benjamins.
- Mann, W., & Thompson, S. (1988). Rhetorical structure theory: Towards a functional

- theory of text organization. *Text*, 8(3), 243–281.
- Manning, C., & Schütze, H. (2000). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.
- Martin, J. R. (2000). Beyond exchange: Appraisal systems in English. In S. Hunston, & G. Thompson (Eds.), *Evaluation in text: Authorial stance and the construction of discourse* (pp. 56–73). Oxford: Oxford University Press.
- Matin, E., Shao, K. C., & Boff, K. R. (1993). Information processing time with and without saccades. *Perception & Psychophysics*, 53(4), 372–380.
- McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. New York: Cambridge University Press.
- McEnery, T., & Wilson, A. (2001). *Corpus linguistics: An introduction*. Edinburgh: Edinburgh University Press.
- McEnery, T. & Z. Xiao. (2004). The Lancaster Corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, (pp. 1175-1178). Lisbon.
- Millis, K. K., & Just, M. A. (1994). The influence of connectives on sentence comprehension. *Journal of Memory and Language*, 33(1), 128–147.
- Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, 59(4), 475–494.
- Morrow, D. G., Greenspan, S. L., & Bower, G. H. (1987). Accessibility and situation models in narrative comprehension. *Journal of Memory and Language*, 26, 165–187.
- Mukherjee, J., & Gries, S. T. 2009. Collostructional nativisation in new Englishes: Verb-construction associations in the International Corpus of English. *English World-Wide*, 30(1), 27–51.
- Núñez, R. (2007). Inferential statistics in the context of empirical cognitive linguistics. In M. Gonzalez-Marquez, I. Mittelberg, S. Coulson, & M. Spivey (Eds.), *Methods in cognitive linguistics* (pp. 87–118). Amsterdam: John Benjamins.
- Nuyts, J. (2001). Subjectivity as an evidential dimension in epistemic modal expressions. *Journal of Pragmatics*, 33(3), 383–400.
- Pander Maat, H. L. W., & Degand, L. (2001). Scaling causal relations and connectives in terms of speaker involvement. *Cognitive Linguistics*, 12(3), 211–245.
- Pander Maat, H. L. W., & Sanders, T. J. M. (2000). Domains of use or subjectivity: The distribution of three Dutch causal connectives explained. In E. Couper-Kuhlen & B. Kortmann (Eds.), *Cause - condition - concession – contrast: Cognitive and discourse perspectives*. (pp. 57–82). Berlin/New York: Mouton de Gruyter.
- Pander Maat, H. L. W. & Sanders, T. J. M. (2001). Subjectivity in causal connectives: An empirical study of language in use. *Cognitive Linguistics*, 12(3), 247–273.

- Pecina, P. (2009). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1–2), 137–158.
- Pit, M. (2003). *How to express yourself with a causal connective: Subjectivity and causal connectives in Dutch, German and French* (Doctoral dissertation, Utrecht University). Amsterdam: Rodopi.
- Pit, M. (2006). Determining subjectivity in text: The case of backward causal connectives in Dutch. *Discourse Processes*, 41(2), 151–174.
- Pyykkönen, P., & Järvikivi, J. (2010). Activation and persistence of implicit causality information in spoken language comprehension. *Experimental Psychology*, 57(1), 5–16.
- Qi, C. (2007). Yuqi fuci yu jumo yuqi zhuci de gongxian guilü [On the co-occurrence rules of modal adverbs and sentence-ending auxiliary modal words]. *Journal of Yunan Normal University (Humanities and Social Sciences)*, 39(3), 125–130.
- R Core Team. (2015). R: A language and environment for statistical computing (Version 3.1.3) [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from: <http://www.R-project.org/>
- Sanders, T. J. M. (1997). Semantic and pragmatic sources of coherence: On the categorization of coherence relations in context. *Discourse Processes*, 24(1), 119–148.
- Sanders, T. J. M., & Noordman, L. G. M. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse Processes*, 29, 37–60.
- Sanders, J., & Redeker, G. (1996). Perspective and the representation of speech and thought in narrative discourse. In *Spaces, worlds, and grammar* (pp. 290–317). Chicago: University of Chicago Press.
- Sanders, J., Sanders, T. J. M., & Sweetser, E. E. (2012). Responsible subjects and discourse causality: How mental spaces and perspective help identifying subjectivity in Dutch backward causal connectives. *Journal of Pragmatics*, 44(2), 191–213.
- Sanders, T. J. M., Sanders, J., & Sweetser, E. E. (2009). Causality, cognition and communication: A mental space analysis of subjectivity in causal connectives. In T. Sanders & E. Sweetser (Eds.), *Causal categories in discourse and cognition* (pp. 19–60). Berlin: Mouton de Gruyter.
- Sanders, T. J. M., & Spooren, W. P. M. S. (2007). Discourse and text structure. In D. Geeraerts & J. Cuykens (Eds.), *Handbook of Cognitive Linguistics* (pp. 916–941). Oxford: Oxford University Press.
- Sanders, T. J. M., & Spooren, W. P. M. (2015). Causality and subjectivity in discourse: The meaning and use of causal connectives in spontaneous conversation, chat interactions and written text. *Linguistics*, 53(1), 53–92.
- Sanders, T. J. M., Spooren, W. P. M. S., & Noordman, L. (1992). Toward a taxonomy of coherence relations. *Discourse Processes*, 15(1), 1–35.
- Sanders, T. J. M., Spooren, W. P. M. S., & Noordman, L. G. M. (1993). Coherence

- relations in a cognitive theory of discourse representation. *Cognitive Linguistics*, 4(2), 93–133.
- Sanders, T. J. M., & Stukker, N. (2012). Causal connectives in discourse: A cross-linguistic perspective. *Journal of Pragmatics*, 44(2), 131–137.
- Sanders, T. J. M., & Sweetser, E. E. (2009). *Causal categories in discourse and cognition*. Berlin: Mouton de Gruyter.
- Schilperoord, J. (1996). *It's about time: Temporal aspects of cognitive processes in text production* (Doctoral dissertation, Utrecht University). Amsterdam: Rodopi.
- Schilperoord, J., & Verhagen, A. (1998). Conceptual dependency and the clausal structure of discourse. In J.-P. Koenig (Ed.), *Discourse and cognition: Bridging the gap* (pp. 141–163). Chicago: University of Chicago Press.
- Schmied, J. (1993). Qualitative and quantitative research approaches to English relative constructions. In S. Clive & E. Atwell (Eds.), *Corpus based computational linguistics* (pp. 85–96). Amsterdam: Rodopi.
- Shih, C., & Sproat, R. (1996). Issues in text-to-speech conversion for Mandarin. *Computational Linguistics and Chinese Language Processing*, 1(1), 37–86.
- Sparks, J.R., & Rapp, D.N. (2011). Readers' reliance on source credibility in the service of inference generation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(1), 230–247.
- Speelman, D. (to appear). *Mastering corpus linguistics methods: A practical introduction with Antconc and R*. Hoboken: Wiley-Blackwell.
- Speelman, D., & Geeraerts, D. (2009). Causes for causatives: The case of Dutch *doen* and *laten*. In T. J. M. Sanders & E. Sweetser (Eds.), *Causal categories in discourse and cognition* (pp. 173–204). Berlin: Mouton de Gruyter.
- Spooren, W. P. M. S., & Sanders, T. J. M. (2008). The acquisition order of coherence relations: On cognitive complexity in discourse. *Journal of Pragmatics*, 40(12), 2003–2026.
- Spooren, W. P. M. S., Sanders, T. J. M., Huiskes, M., & Degand, L. (2010). Subjectivity and causality: A corpus study of spoken language. In S. Rice & J. Newman (Eds.), *Empirical and experimental methods in cognitive/functional research* (pp. 241–255). Chicago: University of Chicago Press.
- Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, 8(2), 209–243.
- Stefanowitsch, A., & Gries, S. T. (2008). Channel and constructional meaning: A collostructional case study. In K. Gitte & R. Dirven (Eds.), *Cognitive sociolinguistics* (pp. 129–152). Berlin: Mouton de Gruyter.
- Strømsø, H., Bråten, I., & Britt, M. A. (2010). Reading multiple texts about climate change: The relationship between memory for sources and text comprehension. *Learning & Instruction*, 20(3), 192–204.
- Stukker, N., & Sanders, T. J. M. (2008). Another('s) perspective on subjectivity in

- causal connectives: A usage-based analysis of volitional causal relations. In W. Ramm & C. Hansen-Fabricsius (Eds.), *Linearisation and segmentation in discourse. Multidisciplinary approaches to discourse* (pp. 113–122). Oslo, Norway: Oslo University.
- Stukker, N., & Sanders, T. J. M. (2012). Subjectivity and prototype structure in causal connectives: A cross-linguistic perspective. *Journal of Pragmatics*, 44(2), 169–190.
- Sun, M., & Zou, J. (2001). Hanyu zidong fenci yanjiu pingshu [A critical appraisal of the research on Chinese word segmentation]. *Contemporary Linguistics*, 1, 22–32.
- Sweetser, E. E. (1990). *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure*. Cambridge: Cambridge University Press.
- Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics*, 30(4), 485–496.
- Tang, Y., & Zhu, Y. (2008). Hanyu beidong/chuzhi gongxian ju luelun [On the sentences with both passive construction and disposal construction]. *Journal of Sun Yat-Sen University (Social science edition)*, 48(1), 53–58.
- Tao, H. (2007). Subjectification and the development of special-verb existential/presentative constructions. *Language and Linguistics* 8(2), 575–602.
- Thompson, G., & Hunston, S. (2000). Evaluation: An introduction. In S. Hunston & G. Thompson (Eds.), *Evaluation in text: Authorial stance and the construction of discourse* (pp. 1–27). Oxford: Oxford University Press.
- Ting, K. (2010). Precision and recall. In C. Sammut & G. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 781). New York: Springer.
- Traugott, E. C. (1989). On the rise of epistemic meanings in English: An example of subjectification in semantic change. *Language*, 65(1), 31–55.
- Traugott, E. C. (1995). Subjectification in grammaticalisation. In S. Wright & D. Stein (Eds.), *Subjectivity and subjectivisation: Linguistic perspectives* (pp. 31–54). Cambridge, UK: Cambridge University Press.
- Traugott, E. C. (2003). From subjectification to intersubjectification. In R. Hickey (Ed.), *Motives for language change* (pp. 124–139). Cambridge: Cambridge University Press.
- Traxler, M. J., Bybee, M. D., & Pickering, M. J. (1997a). Influence of connectives on language comprehension: Eye tracking evidence for incremental interpretation. *The Quarterly Journal of Experimental Psychology Section*, 50A(3), 481–497.
- Traxler, M., Sanford, A. J., Aked, L. M., & Moxey, L. M. (1997b). Processing causals and diagnostics in discourse. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 23(1), 88–101.
- Tseng, H., Chang, P., Andrew, G., Jurafsky, D., & Manning, C. (2005). A conditional random field word segmenter for Sighan Bakeoff 2005. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing* (pp. 168–171).

- Jeju Island, Korea: Asian Federation for Natural Language Processing. doi: 10.1.1.69.3329
- van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.
- van Silfhout, G., Evers-Vermeul, J., Mak, W. M., & Sanders, T. J. M. (2014). Connectives and layout as processing signals: How textual features affect students' processing and text representation. *Journal of Educational Psychology*, 106(4), 1036–1048.
- van Silfhout, G., Evers-Vermeul, J., & Sanders, T. J. M. (2015). Connectives as processing signals : How students benefit in processing narrative and expository texts. *Discourse Processes*, 52(1), 47–76.
- Veenker, T.J.G. (2013). The Zep experiment control application (Version 1.6.3) [Computer software]. Utrecht, Netherlands: Utrecht Institute of Linguistics OTS, Utrecht University.
- Verhagen, A. (2005). *Constructions of intersubjectivity: Discourse, syntax, and cognition*. Oxford: Oxford University Press.
- Wang, C. (2006). Guanyu 'mei(you)' gen 'le' gongxian de wenti [On the co-occurrence of 'měi (yǒu)' and 'le']. *Chinese Teaching in the World*, 1, 41–50.
- Wang, W. (2002). 'Suoyi' wanquan biancheng lianci de shidai [When did *suoyi* thoroughly change into a conjunction]. *Research in Ancient Chinese Language*, 2, 30–33.
- Wang, X., & Guan, Y. (2005). *Jisuanji ziran yuyan chuli* [Natural language processing]. Beijing: Tsinghua University Press.
- Wei, Y., & Evers-Vermeul, J. (2014, June). *Causal implicatures of Chinese temporal conjunctions: A diachronic and corpus-based study of 'erhou', 'ranhou' and 'yushi'*. Paper presented at The Bloomsbury Round Table: 3rd International Symposium on Chinese Language and Discourse, London. Retrieved from: <http://www.bbk.ac.uk/linguistics/downloads/Conference%20Booklet.pdf>
- Wiechmann, D. (2008). On the computation of collocation strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory*, 4(2), 253–290.
- Xing, F. (2001). *Hanyu fujū yanjiu* [A study of Chinese complex sentences]. Beijing: Commercial Publishing House.
- Xing, F. (2002). 'Youyu' ju de yuyi pianxiang bian [On the semantic preference of the pattern introduced by 'youyu']. *Chinese Language*, 4, 337–342.
- Yin, H. (2011). *Foudingci yu fuci gongxian de jufa yuyi yanjiu* [Syntactic and semantic research on the co-occurrence of negatives and adverbs]. Shanghai: Foreign language teaching and research press.
- Zhang, H. (2011). Pindu fuci yu fouding fuci gongxian shi yuxu de bu duichen [Asymmetry of the word order in the collocation of frequency adverbs and negative adverbs]. *Journal of Capital Normal University (Social Sciences*

- Edition*), 1, 113–118.
- Zhang, H., Yu, H., Xiong, D., & Liu, Q. (2003). HHMM-based Chinese lexical analyzer NLP-ICTCLAS. In *Proceeding of the second SIGHAN workshop on Chinese Language Processing* (pp. 184–187). Sapporo, Japan: Association for Computational Linguistics. doi: 10.3115/1119250.1119280
- Zhang, Qi. (2012). ‘Kejian’ de yufahua [The grammaticalization of ‘kejian’]. *Journal of Hubei University of Education*, 29(5), 28–32.
- Zufferey, S. (2012). “Car, parce que, puisque” revisited: Three empirical studies on French causal connectives. *Journal of Pragmatics*, 44(2), 138–153.
- Zwaan, R. A., Magliano, J. P., & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(2), 386–397.
- Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2), 162–185.
- Zwaan, R. A., & Rapp, D. N. (2006). Discourse comprehension. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (2nd ed., pp. 725–764). San Diego: Academic Press.

Appendices

Appendix 1: Overview of segmentation tools for Chinese

1. *NLPIR-ICTCLAS* (Institute of Computing Technology, Chinese Lexical Analysis System) is an HHMM-based (Hierarchical Hidden Markov Model) framework for Chinese word segmentation and annotation. HHMM-based frameworks allow analysis from five levels: atom segmentation, simple and recursive unknown words recognition, class-based segmentation and POS (Part of Speech) tagging (Zhang, Yu, Xiong, & Liu, 2003). *NLPIR-ICTCLAS* provides word segmentation with an accuracy rate of 98.5% (see the evaluation on NLPIR-ICTCLAS-3 in Feng & Zheng, 2011).
2. *Hailiang intellectual word segmentation – research version* is a well-developed word segmentation system for Chinese. It has applied adequate algorithms to better settle the ambiguous segmentations and unknown word recognition. The reported accuracy rate of *Hailiang* in segmenting the closed corpus CCL (Center for Chinese Linguistics of Peking University) is 99.6% (Feng & Zheng, 2011).
3. The *Stanford word segmenter* is a Java implementation of the CRF-based (Conditional Random Field) Chinese word segmenter. It has achieved a high F-score in four Mandarin Chinese corpora: 0.947 for Academia Sinica Corpus, 0.943 for the corpus by University of Hong Kong, 0.950 for the corpus by Peking University, and 0.964 for the corpus by Microsoft Research Asia (Tseng, Chang, Andrew, Jurafsky, & Manning, 2005).
4. The *LTP-cloud* (Language Technology Platform) is an on-line system which provides word segmentation, part-of-speech tagging, syntactic analysis and annotations of semantic roles. Among those functions, the word segmentation module has achieved high accuracy in People's Daily newspaper data (development set: Precision = 0.973, Recall = 0.972, F-score = 0.973¹; test set: Precision = 0.972, Recall = 0.970, F-score = 0.972).
5. *SCWS* (Simple Chinese Word Segmentation) is an open source word segmentation engine based on inserted dictionaries. The tested accuracy is 95% and the recall is 91%.
6. Other tools: *ChineseTA*, supported by Silicon Valley Language Technologies; *Corpus Word Parser* offered by cncorpus.org, *Pan Gu Segment*; *MMSEG* (a word identification system for Mandarin Chinese).

¹*Precision* and *recall* are evaluation measures widely used in machine learning, natural language processing and information retrieval, etc. *Precision* equals the total number of relevant items retrieved divided by the total number of items that are retrieved; *recall* is the total number of relevant items retrieved divided by the total number of relevant items in the database (Ting, 2010). *F-score* is the weighted harmonic mean of *Precision* and *Recall*.

Appendix 2: Overview of Modern Chinese written corpora

Name	Size	Content	Period	Segmented	Accessibility		
					Text availability	Info on corpus size	Download of items
Academia Sinica Balanced Corpus of Modern Chinese 4.0 (现代汉语平衡语料库)	sentences: 1,396,133; word tokens: 11,245,932; word types: 239,598	Balanced: (report, biography, poem, letter, etc.), oral (scenario, conversation, conference); Style: narration; argumentation, exposition; describe; Medium: newspaper, general magazine, academic journal, textbook, audio/visual medium, conversation/interview, etc.; Topics: philosophy, natural science, social sciences, arts, general/leisure, literature	since 1996	yes	full text (charged) web search (free)	yes yes	d.n.a. no limit

Name	Size	Content	Period	Segmented	Accessibility		
					Text availability	Info on corpus size	Download of items
BCC Chinese corpus (北京语言大学现代汉语语料库)	characters: 1,500,000,000	Balanced: news, web literature, etc.	released in September 2014	yes	web search	yes	Max. 1000/10,000 items; randomization allowed
Chinese Gigaword Fifth Edition	n.a.	News	1990-2010	no	full text (charged, available from LDC ² , Catalog No. LDC2016T13)	no	d.n.a

Name	Size	Content	Period	Segmented	Accessibility		
					Text availability	Info on corpus size	Download of items
Chinese POS Tagged Corpus	words tokens: 5,000,000	Balanced: Newswire, fictions, proses, scripts, descriptive documents, letter, argumentative texts , biographies, conversations, essays.	developed 2002.10- 2003.10	yes	full text (charged, available from CLDC, No. CLDC-LAC- 2003-003)	yes	d.n.a
Chinese Treebank 9.0	sentences: 132,076; word tokens: 2,084,387; characters: 3,247,331	Newswire, magazine and articles government documents, chat messages and transcribed conversational telephone speech.	1994-2006	yes	full text (charged, available from LDC, Catalog No. LDC2011T13)	yes	d.n.a

Name	Size	Content	Period	Segmented	Accessibility		
					Text availability	Info on corpus size	Download of items
Cncorpus (国家语委现代汉语通用平衡语料库)	characters: 19,455,328; word tokens: 12,842,116 (incl. punctuations)	Balanced: Textbooks; Humanities and social science (history, economics, literature, arts, etc.); science (agriculture, engineering and technology, etc.), newspapers and magazines, practical writing documents (official documents, letters, advertisements, etc.)	1919-2002	yes	full text (charged, raw corpus available from CLDC ³ No. CLDC-LAC-2006-001)	yes	d.n.a
HIT-CIR Chinese Dependency Treebank	without relations: 50,000 sentences; with relations: 10,000 sentences	Newswire	n.a.	yes	full text	yes	Max. 5000 items; randomization not allowed

³ Chinese Linguistic Data Consortium(中文语言资源联盟)

Name	Size	Content	Period	Segmented	Accessibility		
					Text availability	Info on corpus size	Download of items
Lancaster Corpus of Mandarin Chinese version 2 (LCMCv2)	word tokens: 1,000,000 (incl. punctuation)	Balanced: Press (reportage, editorials, reviews); Religion; Skills, trades and hobbies; Popular lore; Biographies and essays; Miscellaneous (reports and official documents); Science: academic prose; General fiction; Mystery and detective fiction; Science fiction; Adventure and martial arts fiction; Romantic fiction; Humor	1991 (+/-2 years)	yes	full text (free: available from the Oxford Text Archive-Catalogue No. 2474) web search (free)	yes	d.n.a. no limit; randomization allowed

Name	Size	Content	Period	Segmented	Accessibility		
					Text availability	Info on corpus size	Download of items
National broadcast Media Language corpus	characters: 241,316,530 (incl. punctuations)	Broadcast. Media: video, broadcast; Mode: conversations, monologues, dialogues and general; register: conversations, announcements, tutorials.	2008-2013	yes	web search	yes	no limit
Peking University CCL Corpus	characters: 581,794,456	Balanced: fictions, newspaper, conferences, translated literature, blogs, etc.	n.a.	no	web search	yes	no limit
Sketch engine - 9 Chinese sub-corpora	9 sub-corpora with a total collection of 4,099,628,033 tokens	Newswire, web texts, and parallel corpora of web texts.	Depends on sub-corpora	yes	web search/full text	yes	Depends on sub-corpora

Name	Size	Content	Period	Segmented	Accessibility		Download of items
					Text availability	Info on corpus size	
The Corpus of Chinese Compound Sentences (汉语复句语料库)	compound sentences: 658,447; characters: 44,395,000	Newswire	n.a.	no	web search	yes	no limit
The Corpus of Contemporary Novels (当代小说语料库)	sentences: 657,136	Fictions.	n.a.	no	web search	no	no limit
The UCLA Written Chinese Corpus (2nd edition)	word tokens: 1,119,930	Same as LCMC	2000-2012	yes	web search	yes	no limit; randomization allowed
Tsinghua Chinese Treebank	characters: 1,000,000	Literature (fictions, proses, scripts), media (biographies, news), academic works, practical texts.	developed 1998 – 2003	yes	full text (charged, raw corpus available from CLDC, No. CLDC-LAC-2003-005	yes	d.n.a

Appendix 3: Tables of collocates

Table 1. Top 100 collocates ranked by G² (general context)

1-25		26-50		51-75		76-100	
words	dir	words	dir	words	dir	words	dir
wo 'I/me'	1	zhan 'occupy'	-1	juede 'feel'	1	hao 'good'	1
yinwei 'because'	1	laodong 'work'	-1	bingfei 'not'	-1	zai 'again'	1
youyu 'since'	1	cai 'only'	1	nimen 'you'(plural)	1	yizhi 'always'	1
ta 'she/her'	1	yi 'already'	-1	gei 'give'	1	pa 'be afraid of'	1
ni 'you'(singular)	1	jiazhi 'value'	-1	yinggai 'should'	1	zong'e 'total amount'	-1
ta 'he/him'	1	woguo 'our country'	-1	dan 'but'	1	di 'emperor'	-1
women 'we/us'	1	gen 'and'	1	jizai 'record literally'	-1	zengzhang 'increase'	-1
duome 'so much'(exclamatory mood)	-1	lai 'come'	1	ziben 'capital'	-1	yin 'because'	1
tamen 'they'	1	rang 'let'	1	you 'further'	1	juan 'roll'	-1
jiu 'just'	1	renwei 'believe'	1	danshi 'but'	1	men (plural affix)	1
yao 'want'	1	shihou 'time'	1	zai 'at'	1	zhege 'this one'	1
bu 'no'	1	na 'that'	1	tebie 'especially'	1	wei 'for'	-1
dou 'all'	1	bisai 'race'	1	de (particle)	1	shangnian 'last year'	-1
hedeng 'how much'(exclamatory mood)	-1	shengyu jiazhi 'surplus value'	-1	xianzai 'now'	1	yige 'one'	1
xiang 'think'	1	tai 'too'	1	gongren 'worker'	-1	bujin 'not only'	-1
hui 'would'	1	qi 'that'	-1	bili 'proportion'	-1	ju 'according to'	-1
zhi 'of'	-1	de 'of'	1	keyi 'can'	1	mei (negation)	1
ziji 'self'	1	di 'ly'	1	gaosu 'tell'	1	cidian 'dictionary'	-1
meiyou 'have not'	1	neng 'can'	1	zhongyao 'important'	-1	zhunbei 'prepare'	1
zhidao 'know'	1	zhe 'imperfective aspect marker'	1	bixu 'must'	1	kaishi 'begin'	1
qu 'go'	1	zuo 'make'	1	shangpin 'goods'	-1	shengchan 'produce'	-1
hen 'very'	1	yu 'left'	-1	zhi 'reach'	-1	fazhan 'develop'	-1
yuan 'Chinese dollar'	-1	daojiao 'Taoism'	-1	ba (disposal construction)	1	ren 'people'	1
shuo 'say'	1	keneng 'may'	1	shu 'book'	-1	jing 'suprisingly'(short)	-1
le (perfective aspect marker)	1	xiwang 'hope'	1	jingji 'economy'	-1	zhao 'find'	1

Table 2. Top 100 collocates ranked by G² (preceding clause)

1-25		26-50		51-75		76-100	
words	dir	words	dir	words	dir	words	dir
yinwei ‘because’	1	yu ‘left’	-1	dui ‘correct’	1	yinyong ‘quote’	-1
youyu ‘since’	1	juan ‘roll’	-1	zhi ‘reach’	-1	deng ‘wait’	-1
wo ‘I/me’	1	jizai ‘record literally’	-1	shou ‘first’	-1	bijiao ‘compare’	1
shi ‘is’	1	dan ‘but’	1	jin ‘now’	-1	chuan ‘pass’	-1
de ‘of’	1	yijing ‘already’	1	jurán ‘unexpectedly’	-1	gongzuo ‘work’	1
ta ‘she/her’	1	yin ‘because’	1	ju ‘sentence’	-1	yan ‘how’(literary)	-1
ni ‘you’(singular)	1	cheng ‘state’	-1	zengzhang ‘increase’	-1	jiu ‘just’	1
bu ‘no’	1	danshi ‘but’	1	yishang ‘above’	-1	da ‘reach’	-1
ta ‘he/him’	1	renwei ‘believe’	1	na ‘that’	1	jingran ‘suprisingly’(long)	-1
women ‘we/us’	1	zong’e ‘total amount’	-1	tongji ‘statistics’	-1	juede ‘feel’	1
hen ‘very’	1	zheng ‘straight’	1	san ‘three’	-1	shihou ‘time’	1
yuan ‘Chinese dollar’	-1	shu ‘book’	-1	pa ‘be afraid of’	1	jiazhi ‘value’	-1
meiyou ‘have not’	1	yiding ‘for sure’	1	zhe ‘imperfective aspect marker’	1	shi ‘family name’	-1
zhan ‘occupy’	-1	keneng ‘may’	1	chanzhi ‘output value’	-1	dongxi ‘things’	1
zhidao ‘know’	1	buguo ‘but’	1	ju ‘according to’	-1	wen ‘literal’	-1
tamen ‘they’	1	gai ‘change’	-1	chuban ‘first edition’	-1	you ‘further’	1
xiang ‘think’	1	gen ‘and’	1	yige ‘one’	1	zai ‘at’	1
tai ‘too’	1	zhege ‘this one’	1	nian ‘year’	-1	yinggai ‘should’	1
zhi ‘of’	-1	ren ‘people’	1	shi ‘poem’	-1	xing ‘gender’	1
dou ‘all’	1	xihuan ‘like’	1	liaojie ‘understand’	1	chutu ‘unearthed’	-1
ziji ‘self’	1	que ‘but’	-1	ji ‘collection’	-1	ce ‘volume’	-1
hui ‘would’	1	bili ‘proportion’	-1	jian ‘see’	-1	ru ‘if’	-1
yao ‘want’	1	yue ‘say’ (literary)	-1	zhe ‘this’	1	yin ‘print’	-1
jing ‘suprisingly’(short)	-1	woguo ‘our country’	-1	bisai ‘race’	1	yan ‘speak’(literal)	-1
wei ‘for’	-1	shengyu jiazhi ‘surplus value’	-1	yi ‘upon’	-1	mingbai ‘understand’	1

Table 3. Top 100 collocates ranked by G² (following clause)

1-25		26-50		51-75		76-100	
words	dir	words	dir	words	dir	words	dir
wo 'I/me'	1	qi 'that'	-1	jingji 'economy'	-1	guanxi 'relation'	-1
duome 'so much'(exclamatory mood)	-1	gei 'give'	1	de (particle)	1	renwei 'believe'	1
shi 'is'	-1	rang 'let'	1	dao 'reach'	1	kaishi 'begin'	1
jiu 'just'	1	jiao 'call'	1	gen 'and'	1	you 'further'	1
hedeng 'how much'(exclamatory mood)	-1	jiazhi 'value'	-1	shen 'very'	-1	juda 'huge'	-1
ta 'she/her'	1	zhi 'of'	-1	shehui 'society'	-1	wen 'ask'	1
shuo 'say'	1	ziji 'self'	1	ting 'listen'	1	yizhi 'always'	1
yi 'already'	-1	fazhan 'develop'	-1	shen 'deep'	-1	tebie 'especially'	1
ni 'you'(singular)	1	xiang 'think'	1	zuo 'make'	1	juede 'feel'	1
qu 'go'	1	neng 'can'	1	qiye 'company'	-1	zaiyu 'lie in'	-1
ta 'he/him'	1	daojiao 'Taoism'	-1	bing 'and'	-1	bixu 'must'	1
yao 'want'	1	yi 'one'	1	shangpin 'goods'	-1	zai 'again'	1
tamen 'they'	1	shihou 'time'	1	qianli 'potential'	-1	juzuqingzhong 'crucial'	-1
women 'we/us'	1	bian 'convenient'	1	zhao 'find'	1	miansha 'cotton'	-1
bingfei 'not'	-1	ci 'times'	1	zuoyong 'function'	-1	xiande 'seem'	1
zhongyao 'important'	-1	qing 'please'	1	shengyu 'surplus value'	-1	zhe 'imperfective aspect marker'	1
cai 'only'	1	shi 'time'	1	dang 'at'	1	liang 'good'	-1
lai 'come'	1	bisai 'race'	1	bujin 'not only'	-1	ziben 'capital'	-1
dangshi 'that time'	-1	woguo 'our country'	-1	de 'of'	-1	li 'inside'	1
di 'ly'	1	xiangdang 'considerably'	-1	jiang 'be about to'	1	yongxin 'attentively'	-1
le (perfective aspect marker)	1	bei (passive)	1	yinsu 'reason'	-1	zhidu 'system'	-1
laodong 'work'	-1	xiwang 'hope'	1	diwei 'status'	-1	meiyou 'have not'	1
dou 'all'	1	na 'that'	1	chengwei 'be stated as'	1	dai 'take'	1
ba (disposal construction)	1	xianzai 'now'	1	nimen 'you'(plural)	1	dui 'correct'	-1
hui 'would'	1	yijing 'already'	-1	gan 'dare'	1	zaoyi 'very early already'	-1

Table 4. Top 100 collocates ranked by G² (narrative genre)

1-25		26-50		51-75		76-100	
words	dir	words	dir	words	dir	words	dir
wo 'I/me'	1	meiyou 'have not'	1	zai 'again'	1	ren 'mercy'	-1
yinwei 'because'	1	jie 'session'	-1	gen 'and'	1	men (plural affix)	1
jiu 'just'	1	jurán 'unexpectedly'	-1	qin 'diligent'	-1	jie 'all'	-1
women 'we/us'	1	wu 'no'	-1	xi 'to be informed of'	-1	an 'we'(oral)	-1
tamen 'they'	1	cai 'only'	1	liang 'good'	-1	fuhe 'conform to'	-1
zhi 'of'	-1	jian 'see'	-1	mingming 'apparently'	-1	Lin Yuxiang (name)	-1
duome 'so much'(exclamatory mood)	-1	xiang 'mutual'	-1	bingfei 'not'	-1	zhi 'only'	1
ta 'she/her'	1	de 'of'	1	haocheng 'be known as'	-1	diren 'enemy'	1
yao 'want'	1	ba (disposal construction)	1	shang 'even'	-1	dou 'all'	1
qi 'that'	-1	lajiao 'peper'	-1	Stalin (Name)	-1	he 'and'	1
ta 'he/him'	1	que 'but'	-1	shouji 'collect'	-1	ding 'vessel'	-1
youyu 'since'	1	qu 'go'	1	bujin 'not only'	-1	shenme 'what'(literary)	-1
rang 'let'	1	zou 'walk'	1	jing 'suprisingly'(short)	-1	beibi 'mean'	-1
kejian 'so'	-1	zhe 'imperfective aspect marker'	1	guoran 'as expected'	-1	shengguo 'outrace'	-1
ni 'you'(singular)	1	yuanyi 'would like'	1	ji 'avoid'	-1	yongxin 'attentively'	-1
ziji 'self'	1	gei 'give'	1	neng 'can'	1	xinli 'psychology'	-1
zu 'block'	-1	hourou 'next generation'	-1	shen 'deep'	-1	zhineng 'can only'	1
shi 'family name'	-1	danshi 'but'	1	kandao 'see'	1	keneng 'may'	1
zai 'at'	1	biaodian 'punctuation'	-1	xie 'some'	1	Zhou Enlai (Name)	-1
xiang 'think'	1	shen 'very'	-1	zhong 'heavy'	-1	anquanju 'security office'	-1
hui 'would'	1	du 'poison'	-1	fangyang 'dialect'	-1	juejin 'tunnelling'	-1
zhidao 'know'	1	di 'ly'	1	la 'spicy'	-1	Shen Congwen (Name)	-1
renyi 'righteousness'	-1	chunqiu 'spring and autumn'	-1	bixu 'must'	1	yongtu 'purpose'	-1
shihou 'time'	1	bu 'no'	1	dao 'reach'	1	wufa 'have no mean'	1
yuan 'source'	-1	lai 'come'	1	zi 'character'	-1	zhi 'reach'	-1

Table 5. Top 100 collocates ranked by G² (non-narrative genre)

1-25		26-50		51-75		76-100	
words	dir	words	dir	words	dir	words	dir
youyu ‘since’	1	laodong ‘work’	-1	di ‘emperor’	-1	yi ‘also’	-1
yinwei ‘because’	1	ta ‘she/her’	1	qu ‘go’	1	xunlian ‘train’	1
wo ‘I/me’	1	neng ‘can’	1	yin ‘because’	1	chuban ‘first edition’	-1
women ‘we/us’	1	shengyu jiazhi ‘surplus value’	-1	xing ‘gender’	1	juede ‘feel’	1
duome ‘so much’(exclamatory mood)	-1	jiazhi ‘value’	-1	jizai ‘record literally’	-1	ziben ‘capital’	-1
hedeng ‘how much’(exclamatory mood)	-1	dan ‘but’	1	chuan ‘pass’	-1	chanzhi ‘output value’	-1
yao ‘want’	1	keneng ‘may’	1	nimen ‘you’(plural)	1	chang ‘space’	1
dou ‘all’	1	ta ‘it’	1	bingfei ‘not’	-1	miansha ‘cotton’	-1
tamen ‘they’	1	xiang ‘think’	1	canjia ‘participate’	1	chongdie ‘overlap’	-1
ni ‘you’(singular)	1	bixu ‘must’	1	gongren ‘worker’	-1	xianzai ‘now’	1
de ‘of’	1	dui ‘group’	1	shangnian ‘last year’	-1	jing ‘suprisingly’(short)	-1
bisai ‘race’	1	aoyunhui ‘Olympic game’	1	zai ‘at’	1	he ‘and’	1
bu ‘no’	1	daojiao ‘Taoism’	-1	kaishi ‘begin’	1	yizhi ‘always’	1
hui ‘would’	1	xiwang ‘hope’	1	di ‘ly’	1	shu ‘book’	-1
yuan ‘Chinese dollar’	-1	zuo ‘make’	1	zhunbei ‘prepare’	1	renhe ‘any’	1
shuo ‘say’	1	woguo ‘our country’	-1	zong’e ‘total amount’	-1	ce ‘volume’	-1
jiu ‘just’	1	bijiao ‘compare’	1	ji ‘collection’	-1	xianling ‘shilling’	-1
zhan ‘occupy’	-1	yinggai ‘should’	1	juan ‘roll’	-1	shengchanziliao ‘production means’	-1
hen ‘very’	1	feichang ‘very much’	1	you ‘further’	1	danshi ‘but’	1
renwei ‘believe’	1	cai ‘only’	1	shijian ‘time’	1	jinxing ‘going on’	1
meiyou ‘have not’	1	tebie ‘especially’	1	zui ‘most’	1	zhidao ‘know’	1
ziji ‘self’	1	shi ‘is’	1	chengwei ‘be stated as’	1	fubai ‘corruption’	-1
zhi ‘of’	-1	tai ‘too’	1	zhexie ‘these’	1	yiban ‘ordinary’	1
yi ‘already’	-1	ta ‘he/him’	1	bili ‘proportion’	-1	bei (passive)	1
yu ‘left’	-1	keyi ‘can’	1	gaosu ‘tell’	1	lai ‘come’	1

Appendix 4: Experimental items Chapter 4

Experiment 1 (Dutch)

1.

Jon raakte geblesseerd bij het skien, daardoor / moet hij / tot het volgende seizoen rust houden.

Jon raakte geblesseerd bij het skien, dus / moet hij / wel te veel risico hebben genomen.

Jon got injured while skiing, so he has to take rest until next season.

Jon got injured while skiing, so he must have taken too much risk.

2.

Er was vanmorgen een bommelding, daardoor / zal het gebouw / vandaag gesloten zijn.

Er was vanmorgen een bommelding, dus / zal het gebouw / vandaag wel niet veilig zijn.

A security breach occurred last night, so the building will be closed today.

A security breach occurred last night, so the building may not be safe today.

3.

De kat zat in een boom, daardoor / heeft de eigenaar / hem niet gezien.

De kat zat in een boom, dus / heeft de eigenaar / heeft hem naar buiten gelaten.

The cat was stuck in a tree, so it's owner did not see it.

The cat was stuck in the tree, so it's owner has left him out.

4.

De hond heeft zijn huiswerk opgegeten, daardoor / heeft Max / vandaag geen huiswerk bij zich.

De hond heeft zijn huiswerk opgegeten, dus / heeft Max / het laten slingeren.

The dog ate his homework, so Max does not have his homework today.

The dog ate his homework, so Max has left it lying around.

5.

Ze hebben aan hun CV gewerkt, daardoor / hebben ze / nu een verbeterde versie.

Ze hebben aan hun CV gewerkt, dus / hebben ze / plannen om te solliciteren.

They have been working on their resume, so they have an updated version.

They have been working on their resume, so they have plans to apply for a new job.

6.

Robbert heeft regelmatig getraind, daardoor / heeft hij / veel gewicht verloren.
Robbert heeft regelmatig getraind, dus / heeft hij / een heel aantrekkelijk lichaam.

Robbert has been exercising regularly, so he has lost much weight.
Robbert has been exercising regularly, so he must have a very attractive body.

7.

Daphne is weer ziek geworden, daardoor / moet ze / in bed blijven.
Daphne is weer ziek geworden, dus / moet ze / een zwak immuunsysteem hebben.

Daphne has caught a cold again, so she will have to stay in bed.
Daphne has caught a cold again, so she must have a weakened immune system.

8.

Jocelyn reist met de trein naar Italië, daardoor / zal ze / lang onderweg zijn.
Jocelyn reist met de trein naar Italië, dus / zal ze / wel last van vliegangst hebben.

Jocelyn travels by train to Italy, so she will be in transit for a long time.
Jocelyn travels by train to Italy, so she may be afraid of air travel.

9.

Dit restaurant heeft hippe decoraties, daardoor / heeft het / een jong publiek.
Dit restaurant heeft hippe decoraties, dus / heeft het / veel geld gekost het in te richten.

This restaurant has eclectic decorations, so it popular amongst young people.
This restaurant has eclectic decorations, so it has cost a lot of money to decorate it.

10.

Alain heeft ernstige hoogtevrees, daardoor / zal hij / thuis blijven.
Alain heeft ernstige hoogtevrees, dus / zal hij / wel een nare ervaring hebben gehad.

Alain is scared of heights, so he will stay home.
Alain is scared of heights, so he must have had a bad experience.

11.

Joanne heeft vijf koppen koffie gedronken, daardoor / is ze / heel erg hyper.
Joanne heeft vijf koppen koffie gedronken. dus / is ze / heel erg verslaafd aan cafeïne.

Joana drinks five cups of coffee a day, so she is very hyper.
Joana drinks five cups of coffee a day, so she is be addicted to caffeine.

12.

Het is heel druk in de winkel, daardoor / zal het / lang duren om te betalen.
Het is heel druk in de winkel, dus / zal het / wel uitverkoop zijn.

The store is busy with customers, so it takes a while to pay.
The store is busy with customers, so it's probably having a sale.

13.

De trein heeft vertraging, daardoor / zullen veel mensen / te laat op hun werk zijn.
De trein heeft vertraging, dus / zullen veel mensen / te laat op hun werk zijn.

The train is delayed, so many people will be late for work.
The train is delayed, so many people will be late for work.

14.

Angela heeft astronomie gestudeerd, daardoor / kan ze / alles vertellen over sterrenstelsels.
Angela heeft astronomie gestudeerd, dus / kan ze / solliciteren bij de NASA.

Angela studied astronomy at university, so she can explain everything about stellar constellations.
Angela studied astronomy at university, so she can apply for a job at NASA.

15.

Zijn fiets had een lekke band, daardoor / was het / niet mogelijk verder te fietsen.
Zijn fiets had een lekke band, dus / was het / niet mogelijk om verder te fietsen.

His bike had a flat tire, so it was impossible to drive on.
His bike had a flat tire, so it was impossible to drive on.

16.

Elise heeft op een debatclub gezeten, daardoor / is ze / goed in speechen.
Elise heeft op een debatclub gezeten, dus / is ze / goed in speechen.

Eletta can communicate effectively, so she is good at giving speeches.
Eletta can communicate effectively, so she is good at giving speeches.

17.

Het pak melk is over de datum, daardoor / moet het / weggegooid worden.
Het pak melk is over de datum, dus / moet het / al lang in de koelkast hebben gestaan.

The milk is past expiration, so it must be thrown away.
The milk is past expiration, so it must have been in the refrigerator for a while.

18.

Ze reizen vaak naar de Cariben, daardoor / zijn ze / veel geld kwijt aan vakanties.
Ze reizen vaak naar de Cariben, dus / zijn ze / vast dol op een tropisch klimaat.

They travel to the Carribean often, so they spend a lot of money on their vacations.
They travel to the Carribean often, so they must enjoy the tropical climate.

19.

Het is koud voor de tijd van het jaar, daardoor / is het / niet prettig om buiten te zijn.
Het is koud voor de tijd van het jaar, dus / is het / klimaat aan het veranderen.

The weather is unseasonably cold, so it is not nice to be outside.
The weather is unseasonably cold, so the climate is changing.

20.

Het bedrijf heeft het water vervuild, daardoor / heeft het / een milieuramp veroorzaakt.
Het bedrijf heeft het water vervuild, dus / heeft het / onverantwoordelijke eigenaar.

The factory has been polluting the water, so it has caused an environmental disaster.
The factory has been polluting the water, so it has an irresponsible owner.

Experiment 2 (Chinese)

1.

约翰上个月去滑雪时摔伤了胳膊，**所以/因而**他这段时间经受了巨大的痛苦。
约翰上个月去滑雪时摔伤了胳膊，**所以/可见**他当时**可能**是在很滑的雪坡上。

Yuehan shangge yue qu huaxue shi shuaishang le gebo, suoyi/yin'er ta zhe duan shijian jingshou le juda de tongku.

Yuehan shangge yue qu huaxue shi shuaishang le gebo, suoyi/kejian ta dangshi keneng shi zai hen hua de xuepo shang.

NAME last month go ski time fall:injure ASP(PFV) arm, **CONJ** 3SGM this CL time experience ASP(PFV) huge MOD pain.

NAME last month go ski time fall:injure ASP(PFV) arm, **CONJ** 3SGM that:time **may** COP at very icy MOD slopes up.

Jon injured his arm while skiing last month, suoyi/yin'er he experienced a lot of pain during this period of time.

Jon injured his arm while skiing last month, suoyi/kejian he may have been on icy slopes.

2.

昨天这座城市最高建筑的火险警报响了，**所以/因而**楼里的电梯自动停运了。
昨天这座城市最高建筑的火险警报响了，**所以/可见**这栋楼**可能**并不安全。

Zuotian zhe zuo chengshi zui gao jianzhu de huoxian jingbao xiang le, **suoyi/yin'er** louli de dianti zidong tingyun le.

Zuotian zhe zuo chengshi zui gao jianzhu de huoxian jingbao xiang le, **suoyi/kejian** zhe dong lou **keneng** bing bu anquan.

Yesterday this CL city most high architecture POSS fire alarm ring ASP(PFV), **CONJ** building MOD lift automatic stop:operate ASP(PFV).

Yesterday this CL city most high architecture POSS fire alarm ring ASP(PFV), **CONJ** this CL building **may** EMP NEG safe.

A fire alarm occurred in the highest building of the city yesterday, suoyi/yin'er the lift of this building stopped automatically.

A fire alarm occurred in the highest building of the city yesterday, suoyi/kejian the building may not be safe to enter.

3.

这只小猫被卡在了公园的一棵树上，**所以/因而**这棵树上的鸟儿都被惊跑了。
这只小猫被卡在了公园的一棵树上，**所以/可见**它的主人**可能**没有关好家里的门。

Zhe zhi xiao mao bei ka zai le gongyuan de yi ke shu shang, **suoyi/yin'er** zhe ke shu

shang de niao'er dou bei jing pao le.

Zhe zhi xiao mao bei ka zai le gongyuan de yi ke shu shang, **suoyi/kejian** ta de zhuren **keneng** meiyou guan hao jiali de men.

This CL little cat PASS stick at ASP(PFV) park MOD one CL tree up, **CONJ** this CL tree MOD birds all PASS scare away ASP(PFV).

This CL little cat PASS stick at ASP(PFV) park MOD one CL tree up, **CONJ** 3SGN POSS owner **may** NEG close well home MOD door.

The little cat was stuck in a tree in the park, *suoyi/yin'er* birds on this tree were scared away.

The little cat was stuck in a tree in the park, *suoyi/kejian* it's owner may have left the door open.

4.

迈克的狗把迈克写的数学作业给咬烂了, **所以/因而**他给老师交作业交迟了。

迈克的狗把迈克写的数学作业给咬烂了, **所以/可见**他之前**可能**把作业放在了狗可以够到的地方。

Maike de gou ba ta xie de shuxue zuoye gei yao lan le, **suoyi/yin'er** ta gei laoshi jiao zuoye jiao chi le.

Maike de gou ba ta xie de shuxue zuoye gei yao lan le, **suoyi/kejian** ta zhiqian **keneng** ba zuoye fang zai le gou keyi goudao de difang.

NAME POSS dog BA 3SGM write MOD math homework PASS bite damaged ASP(PFV), **CONJ** 3SGM for teacher submit homework submit late ASP(PFV).

NAME POSS dog BA 3SGM write MOD math homework PASS bite damaged ASP(PFV), **CONJ** 3SGM earlier **may** BA homework put at ASP(PFV) dog can reach MOD place.

Max's dog ate the math homework Max wrote, *suoyi/yin'er* he was late to submit homework to the teacher.

Max's dog ate the math homework Max wrote, *suoyi/kejian* he may have left it where the dog can reach it earlier.

5.

这些学生一整天都在讨论课程论文, **所以/因而**他们都不知道外面发生了什么。

这些学生一整天都在讨论课程论文, **所以/可见**他们几个**可能**是快到提交论文的期限了。

Zhexie xuesheng yi zhengtian dou zai taolun kecheng lunwen, **suoyi/yin'er** tamen dou bu zhidao waimian fasheng le shenme.

Zhexie xuesheng yi zhengtian dou zai taolun kecheng lunwen, **suoyi/kejian** tamen jige **keneng** shi kuai dao tijiao lunwen de qixian le.

These student one whole:day EMP ASP(IPFV) revise master course essay, **CONJ**

3PL EMP NEG know outside happen ASP(PFV) what.
These student one whole:day EMP ASP(IPFV) revise master course essay, **CONJ**
3PL several **may** COP quick reach submit essay MOD deadline ASP(PFV).

They have been discussing on their course essays for the whole day, **suoyi/yin'er** they have no idea of what happened outside.

They have been discussing on their master course essays for the whole day, **suoyi/kejian** they may have an approaching deadline.

6.

罗伯特从去年开始就坚持每周健身三次，**所以/因而**他瘦下来了很多。

罗伯特从去年开始就坚持每周健身三次，**所以/可见**他对身材**一定**是在乎的。

Luobote cong qunian kaishi jiu jianchi mei zhou jianshen san ci, **suoyi/yin'er** ta shou xialai le henduo.

Luobote cong qunian kaishi jiu jianchi mei zhou jianshen san ci, **suoyi/kejian** ta dui shencai **yiding** shi zaihu de.

NAME since last:year start EMP insist every week exercise three CL, **CONJ** 3SGM thin down ASP(PFV) very:much.

NAME since last:year start EMP insist every week exercise three CL, **CONJ** 3SGM for fitness **must** COP care PRT.

Robbert has been exercising three times a week since last year, **suoyi/yin'er** he has lost weight.

Robbert has been exercising three times a week since last year, **suoyi/kejian** he must care about being fit.

7.

安妮这周又感染了流行性病毒感冒，**所以/因而**她今天没有来开会。

安妮这周又感染了流行性病毒感冒，**所以/可见**她免疫力**可能**不行。

Anni zhe zhou you ganran le liuxingxing bingdu ganmao, **suoyi/yin'er** ta jintian meiyou lai kaihui.

Anni zhe zhou you ganran le liuxingxing bingdu ganmao, **suoyi/kejian** ta mianyili **keneng** bu xing.

NAME this week again infect ASP(PFV) epidemic virus cold, **CONJ** 3SGF today NEG come attend:meeting.

NAME this week again infect ASP(PFV) epidemic virus cold, **CONJ** 3SGF immune **may** NEG good.

Anne has caught an influenza this week again, **suoyi/yin'er** she was not present in the meeting today.

Anne has caught an influenza this week again, **suoyi/kejian** she may have a weakened immune system.

8.

史蒂夫每个月都坐火车去奥地利出差，**所以/因而**他熟知这趟列车的时刻表。
 史蒂夫每个月都坐火车去奥地利出差，**所以/可见**他们公司**可能**有跨国业务。

Shidifu mei ge yue dou zuo huoche qu Aodili chuchai, **suoyi/yin'er** ta shuzhi zhe tang lieche de shikebiao.

Shidifu mei ge yue dou zuo huoche qu Aodili chuchai, **suoyi/kejian** tamen gongsi **keneng** you kuaguo yewu.

NAME every CL month EMP sit train go Austria on:business, **CONJ** 3PM know:well this CL train MOD schedule.

NAME every CL month EMP sit train go Austria on:business, **CONJ** 3PM company **may** have international business.

Steven travels by train to Austria every month for work, suoyi/yin'er he knows the train schedule well.

Steven travels by train to Austria every month for work, suoyi/kejian their company may have international business.

9.

这家新开的餐厅有高档的艺术装饰品，**所以/因而**它吸引了不少艺术爱好者。
 这家新开的餐厅有高档的艺术装饰品，**所以/可见**它的老板**可能**花了大价钱装修。

Zhe jia xin kai de canting you gaodang de yishu zhuangshipin, **suoyi/yin'er** ta xiyin le bushao yishu aihaozhe.

Zhe jia xin kai de canting you gaodang de yishu zhuangshipin, **suoyi/kejian** ta de laoban **keneng** hua le da jiaqian zhuangxiu.

This CL new open MOD restaurant have high-grade MOD art decoration, **CONJ** 3SGN attract ASP(PFV) a:lot art lover.

This CL new open MOD restaurant have high-grade MOD art decoration, **CONJ** 3SGN POSS owner **may** spend ASP(PFV) big price fix:up.

This new restaurant has high-grade art decorations, **suoyi/yin'er** it attracts a lot of art lovers.

This new restaurant has high-grade art decorations, **suoyi/kejian** its owner may have paid a lot of money on them.

10.

小学二年级的琳达害怕站在高处，**所以/因而**她在一次爬山活动中几乎晕倒。
 小学二年级的琳达害怕站在高处，**所以/可见**她以前**可能**有过可怕的经历。

Xiaoxue ernianji de Linda haipa zhan zai gaochu, **suoyi/yin'er** ta zai yi ci pashan huodong zhong jihu yun dao.

Xiaoxue ernianji de Linda haipa zhan zai gaochu, **suoyi/kejian** ta yiqian **keneng** you guo rang ta haipa de jingli.

Primary:school second:grade MOD NAME be:afraid:of stand at high:place, **CONJ** 3SGF at one CL mountain:climbing activity almost fell down.

Primary:school second:grade MOD NAME be:afraid:of stand at high:place, **CONJ** 3SGF before **may** have ASP(PFV) terrible experience.

Linda from the second grade is scared of heights, suoyi/yin'er she almost fell in faint in a mountain climbing activity.

Linda from the second grade is scared of heights, suoyi/kejian she may have had some terrible experience before.

11.

乔安娜每天至少喝五杯加浓的咖啡, **所以/因而**她一直精力充沛。

乔安娜每天至少喝五杯加浓的咖啡, **所以/可见**她对咖啡**一定**是上瘾。

Qiaoanna meitian zhishao he wu bei jianong de kafei, **suoyi/yin'er** ta yizhi jingli chongpei.

Qiaoanna meitian zhishao he wu bei jianong de kafei, **suoyi/kejian** ta dui kafei **yiding** shi shangyin.

NAME everyday at:least drink five CL special:strong MOD coffee, **CONJ** 3SGF always energy sufficient.

NAME everyday at:least drink five CL special:strong MOD coffee, **CONJ** 3SGF for coffee **must** be addicted.

Joana drinks at least five cups of strong coffee a day, suoyi/yin'er she always stays hyper.

Joana drinks at least five cups of strong coffee a day, suoyi/kejian she must be addicted to caffeine.

12.

这家电子产品商店有很多顾客在选购，**所以/因而**排队付款花了很长时间。
这家电子产品商店有很多顾客在选购，**所以/可见**这家店**可能**在搞促销活动。

Zhe jia dianzi chanpin shangdian you henduo guke zai xuangou, **suoyi/yin'er** paidui fukuan hua le hen chang shijian.

Zhe jia dianzi chanpin shangdian you henduo guke zai xuangou, **suoyi/kejian** zhe jia dian **keneng** zai gao cuxiao huodong.

This CL electronic product shop have many customer ASP(IPFV) shop, **CONJ** queue pay spend ASP(PFV) very long time.

This CL electronic product shop have many customer ASP(IPFV) shop, **CONJ** this CL shop **may** ASP(IPFV) make sale activity.

The media store is very busy with customers, suoyi/yin'er it took a while to pay.

The media store is very busy with customers, suoyi/kejian so it may be having a sale.

13.

开往波茨坦方向的火车误点了一个小时，**所以/因而**很多人上班迟到了。
开往波茨坦方向的火车误点了一个小时，**所以/可见**铁路系统**可能**出了事故。

Kaiwang Bocitan fangxiang de huoche wudian le yi ge xiaoshi, **suoyi/yin'er** henduo ren shangban chidao le.

Kaiwang Bocitan fangxiang de huoche wudian le yi ge xiaoshi, **suoyi/kejian** tielu xitong **keneng** chu le shigu.

Towards Potsdam direction MOD train delay ASP(PFV) one CL hour, **CONJ** many people work late ASP(PFV).

Towards Potsdam direction MOD train delay ASP(PFV) one CL hour, **CONJ** railway system **may** happen ASP(PFV) accident.

The train to Potsdam was delayed for one hour, suoyi/yin'er many people were late for work.

The train to Potsdam was delayed for one hour, suoyi/kejian the railway system may have encountered an accident.

14.

丹尼尔在大学里面选修了行星天文学课，**所以/因而**他学到了许多天体的名称。
丹尼尔在大学里面选修了行星天文学课，**所以/可见**他对天体**一定**是感兴趣的。

Danni'er zai daxue limian xuanxiu le xingxing tianwenxue ke, **suoyi/yin'er** ta xuedao le xuduo tianti de mingcheng.

Danni'er zai daxue limian xuanxiu le xingxing tianwenxue ke, **suoyi/kejian** ta dui tianti **yiding** shi ganxingqu.

NAME at university inside choose:practice ASP(PFV) planetary astronomy course,
CONJ 3SGM learn ASP(PFV) many constellation POSS name.

NAME at university inside choose:practice ASP(PFV) planetary astronomy course,
CONJ 3SGM for constellation **must** COP interested.

Daniel followed planetary astronomy as a minor course at university, *suoyi/yin'er* he learnt the name of many constellations.

Daniel followed planetary astronomy as a minor course at university, *suoyi/kejian* he may be interested in constellations.

15.

这个年轻人的摩托车车胎没气了，**所以/因而**他花了很长时间才到家。

这个年轻人的摩托车车胎没气了，**所以/可见**他的车胎**可能**被刺破了。

Zhe ge nianqing ren de motuoche chetai meiqi le, **suoyi/yin'er** ta hua le hen chang shijian cai dao jia.

Zhe ge nianqing ren de motuoche chetai meiqi le, **suoyi/kejian** ta de chetai **keneng** bei cipo le.

This CL young man POSS motorbike tire no:air ASP(PFV), **CONJ** 3SGM spend ASP(PFV) very long time EMP get home.

This CL young man POSS motorbike tire no:air ASP(PFV), **CONJ** 3SGM POSS tire **may** PASS pierce:out ASP(PFV).

The young man's motorbike had a flat tire, *suoyi/yin'er* he spent a long time getting home.

The young man's motorbike had a flat tire, *suoyi/kejian* it may have a nail in it.

16.

艾利克上周的演讲受到了广泛好评，**所以/因而**他收到了新的演讲邀请。

艾利克上周的演讲受到了广泛好评，**所以/可见**他为此**一定**花了很多精力准备。

Ailike shangzhou de yanjiang shoudao le guangfan haoping, **suoyi/yin'er** ta shoudao le xin de yanjiang yaoqing.

Ailike shangzhou de yanjiang shoudao le guangfan haoping, **suoyi/kejian** ta wei ci **must** ta weici keneng hua le henduo jingli zhunbei.

NAME last:week MOD speech obtain ASP(PFV) widespread positive:feedback,
CONJ 3SGM receive ASP(PFV) new MOD speech invitation.

NAME last:week MOD speech obtain ASP(PFV) widespread positive:feedback,
CONJ 3SGM for this **must** spend ASP(PFV) a:lot effort prepare.

Eric's talk last week received lots of positive feedbacks, *suoyi/yin'er* he received another invitation for public speech.

Eric's talk last week received lots of positive feedbacks, *suoyi/kejian* he must have spent a great effort in preparation.

17.

这家超市的牛奶已经过了保质期，**所以/因而**牛奶闻起来有些异味。
这家超市的牛奶已经过了保质期，**所以/可见**这批牛奶**一定**在货架上放了一段时间了。

Zhe jia chaoshi de niunai yijing guo le baozhiqi, **suoyi/yin'er** niunai wenqilai you xie yiwei.
Zhe jia chaoshi de niunai yijing guo le baozhiqi, **suoyi/kejian** zhe pi niunai **yiding** zai huojia shang fang le duan shijian le.

This CL supermarket POSS milk already pass ASP(PFV) valid:date, **CONJ** milk smell have some odd:smell.
This CL supermarket POSS milk already pass ASP(PFV) valid:date, **CONJ** this CL milk **must** at shelf up set ASP(PFV) one period time ASP(PFV).

The milk in this supermarket is past expiration, **suoyi/yin'er** it smells bad.
The milk in this supermarket is past expiration, **suoyi/kejian** it must be on the shelf for quite a while.

18.

布朗夫妇夏天经常去加勒比海地区度假，**所以/因而**他们的皮肤被晒成了小麦色。
布朗夫妇夏天经常去加勒比海地区度假，**所以/可见**他们俩**一定**喜欢赤道地区的气候。

Bulang fufu xiatian jingchang qu Jialebihai diqu dujia, **suoyi/yin'er** tamen de pifu bei shaicheng le xiaomaise.
Bulang fufu xiatian jingchang qu Jialebihai diqu dujia, **suoyi/kejian** tamen lia **yiding** xihuan chidao diqu de qihou.

Brown couple summer often go Caribbean region on:vacation, **CONJ** 3PL POSS skin PASS sunburn ASP(PFV) tan.
Brown couple summer often go Caribbean region on:vacation, **CONJ** 3PL two **must** like tropical area MOD climate.

Mr.& Mrs. Brown travelled to the Caribbean area for holiday often, **suoyi/yin'er** they looked tanned on their skin.
Mr.& Mrs. Brown travelled to the Caribbean area for holiday often, **suoyi/kejian** they two must enjoy the tropical climate.

19.

今年冬天的天气异于往年地寒冷，**所以/因而**一些商店的防寒服都卖光了。
今年冬天的天气异于往年地寒冷，**所以/可见**全球变暖**可能**并非绝对趋势。

Jinnian dongtian de tianqi yiyu wangnian de hanleng, **suoyi/yin'er** yixie shangdian de fanghanfu dou mai guang le.

Jinnian dongtian de tianqi yiyu wangnian de hanleng, **suoyi/kejian** quanqiu biannuan **keneng** bing fei juegui qushi.

This:year winter MOD weather different:from previous:years MOD cold, **CONJ** some store POSS anorak EMP sell out ASP(PFV).

This:year winter MOD weather different:from previous:years MOD cold, **CONJ** global warming **may** EMP NEG absolute tendency.

The weather is unseasonably cold this winter, **suoyi/yin'er** anoraks have been sold out in some stores.

The weather is unseasonably cold this winter, **suoyi/kejian** global warming may not be an absolute tendency.

20.

这家私营的化工厂一直在排放污水，**所以/因而**附近河流里的鱼类大面积死亡。这家私营的化工厂一直在排放污水，**所以/可见**它的厂主**可能**并不关心环境保护。

Zhe jia siying de huagongchang yizhi zai paifang wushui, **suoyi/yin'er** fujin heliu li de yulei da mianji siwang.

Zhe jia siying de huagongchang yizhi zai paifang wushui, **suoyi/kejian** ta de changzhu **keneng** bing bu guanxin huanjing baohu.

This CL private MOD chemistry factory always ASP(IPFV) discharge polluted:water, **CONJ** nearby river inside MOD fish:kind big area die.

This CL private MOD chemistry factory always ASP(IPFV) discharge polluted:water, **CONJ** 3SGN POSS owner **may** EMP NEG care environment protection.

The private chemistry factory has been polluting the water, **suoyi/yin'er** fishes in the rivers nearby are dying in large scale.

The private chemistry factory has been polluting the water, **suoyi/kejian** its owner may not care about environment protection.

Appendix 5: Experimental items Chapter 5⁴

1.

这家公司|Ø/听说/竟然|有一半|的员工都被辞退了，|所以/可见|公司效益|可能|不容|乐观。|很多在职员工都开始考虑退路。

Zhe jia gongsi| Ø/**tingshuo/jingran**| you yiban| de yuangong dou bei citui le,| **suoyi/kejian**| gongsi xiaoyi| **keneng**| bu rong| leguan.| henduo zaizhi yuangong dou kaishi kaolü tuilu.

This CL company| Ø/**as reported/surprisingly**| have half| MOD employee all PASS fire ASP(PFV),| **CONJ**| company economic:benefit| **may**| NEG allow| optimism.| Many at:work employees even start consider backup:plan.

Ø/As reported/Surprisingly half of the employees of this company have been fired, **suoyi/kejian** the economic benefit of this company **may** not be optimistic. Many employees who are still at work start considering backup plans.

2.

这家小区车库|Ø/听说/竟然|在白天|时间被盗，|所以/可见|小区安保|可能|做得不到位。|一些居民要求增加小区摄像头。

Zhejia xiaoqu cheku| Ø/**tingshuo/jingran**| zai baitian| shijian bei dao,| **suoyi/kejian**| xiaoqu anbao| **keneng**| zuo de| bu daowei.| Yixie jumin yaoqiu zengjia xiaoqu shexiangtou.

This living:district garage| Ø/**as reported/surprisingly**| during day| time PASS rob,| **CONJ**| district security| **may**| do PRT| NEG enough. | Some residents request increase living:district camera.

Ø/As reported/Surprisingly the garage of this district was robbed during daytime, **suoyi/kejian** the security **may** not do enough. Some residents requested to increase the number of cameras in this district.

3.

长途游轮|Ø/听说/竟然|需要|提前一个小时准备登船，|所以/可见|游轮载客量|一定|非常|可观。|船舱里设有两个餐厅和若干吧台。

Changtu youlun| Ø/**tingshuo/jingran**| xuyao| tiqian yi ge xiaoshi zhunbei dengchuan,| **suoyi/kejian**| youlun zaikeliang| **yiding**| feichang| keguan.| Chuancang li sheyou liang ge canting he ruogan batai.

Long:distance ferry| Ø/**as reported/surprisingly**| need| advance one CL hour prepare boarding,| **CONJ**| ferry capacity:passengers| **must**| very| considerable.| Cabin inside set two CL canteen and several bar.

Ø/As reported/Surprisingly the long distance ferry needs one hour in advance for boarding, **suoyi/kejian** the capacity of passengers of the ferry **must** be considerable. Inside the cabin there are two canteens and several bars.

Verification statement:

⁴ In this experiment, participants saw one out of six possible combinations of a perspective marker (no stance marker (indicated with Ø), evidential stance marker or attitudinal stance marker) and a connective (*suoyi* of *kejian*).

这艘长途游轮提供餐饮服务。

This ferry provides food and drink service. (TRUE)

4.

小美毕业后|Ø/听说/居然|还一直|需要她父母照顾，|所以/可见|她自理能力|可能|有很大|问题。|她的同学们都去了全国各地。

Xiaomei biye hou| Ø/tingshuo/juran| hai yizhi| xuyao ta fumu zhaogu,| **suoyi/kejian**| ta zili nengli| **keneng**| you henda| wenti. | Ta de tongxue men dou qu le quanguo gedi.

NAME graduate after| Ø/ As reported/Unexpectedly| still always| need 3SGF parents care,| **CONJ**| 3SGF independent:living ability| **may**| have very:big| problem. | 3SGF MOD classmates all go ASP(PFV) country every:region.

Ø/ As reported/Unexpectedly, Xiaomei still needs her parents to take care of her after graduating from college, **suoyi/kejian** her ability to live alone **may** be problematic. Her classmates are distributed throughout the country.

5.

城际特快| Ø/听说/居然|误点了|一个多小时，|所以/可见|铁路系统|可能|出了|事故。|这种情况并不常见。

Chengji tekuai| Ø/tingshuo/juran| wudian le| yige duo xiaoshi,| **suoyi/kejian**| tielu xitong| **keneng**| chu le| shigu. | Zhezhong qingkuang bing bu changjian.

Intercity express| Ø/ as reported/unexpectedly| is:delayed ASP(PFV)| one more hour,| **CONJ**| railway system| **may**| happen ASP(PFV) accident. | This:kind situation EMP NEG common.

Ø/ As reported/Unexpectedly the intercity express is delayed for more than one hour, **suoyi/kejian** the railway system **may** have encountered an accident. This kind of situation is not common.

Verification statement:

城际特快经常会误点。

The intercity express is always delayed. (FALSE)

6.

莹莹的生日宴|Ø/听说/居然|没有|邀请她的室友们，|所以/可见|她和室友|可能|相处得|并不愉快。朋友们为莹莹准备了一个双层蛋糕。

Yingying de shengriyan| Ø/tingshuo/juran| meiyou| yaoqing ta de shiyoumen,| **suoyi/kejian**| ta he shiyou| **keneng**| xiangchu de bing bu yukuai. Pengyoumen wei Yingying zhunbei le yi ge shuangceng dangao.

NAME MOD birthday:party| Ø/ as reported/unexpectedly| NEG| invite 3SGF MOD roommates,| **CONJ**| 3SGF and roommate| **may**| get along PRT| EMP NEG pleasant. | Friends for NAME prepare ASP(PFV) one CL double-layer cake.

Ø/ As reported/Unexpectedly Yingying didn't invite her roommates to her birthday party, **suoyi/kejian** she and her roommates may not get along with each other. Friends prepared a double-layer cake for Yingying.

Verification statement:

莹莹今年没有庆祝生日。

Yingying didn't celebrate her birthday this year. (FALSE)

7.

Condition 1 & 2

安妮这周|**Ø/听说/不幸**|又感染|了流行性病毒感冒，|**所以/可见**|她的免疫力|**可能**|比较|脆弱。|医生给她开了抗病毒的药。

Anni zhe zhou| **Ø/tingshuo/buxing**| you ganran| le liuxingxing bingdu ganmao,| **suoyi/kejian**| ta de mianyili| **keneng**| bijiao| cuiruo.| Yisheng gei ta kai le kangbingdu de yao.

NAME this week| **Ø/As reported/unfortunately**| again infect| ASP(PFV) epidemic virus flu,| **CONJ**| 3SGF POSS immune| **may** | comparatively| weak.| Doctor give 3SGF prescribe ASP(PFV) anti-virus MOD medicine.

Ø/As reported/Unfortunately Anne was infected by an epidemic flu this week again, **suoyi/kejian** her immune system **may** be comparatively weak. The doctor prescribed an anti-virus medicine for her.

Verification statement:

安妮不止一次得过流感。

Anne got flu before. (TRUE)

8.

多家本土企业|**Ø/听说/不幸**|最近|都被跨国公司收购了，|**所以/可见**|本土企业|**可能**|在经济|全球化中发展乏力。|很多年轻人更喜欢在国际大公司工作。

Duo jia bentu qiye| **Ø/tingshuo/buxing**| zuijin| dou bei kuaguo gongsi shougou le,| **suoyi/kejian**| bentu qiye| **keneng**| zai jingji| quanqiuhua zhong fazhan falì.| henduo nianqingren geng xihuan zai guoji dagongsi gongzuo.

Many CL local company| **Ø/As reported/unfortunately**| recently| all PASS international company purchase PRT,| **suoyi/kejian**| local enterprise| **may**| at economy| globalization inside develop weak.| Many young:people more prefer at international company work.

Ø/As reported/Unfortunately a lot of local companies have been purchased by international companies, **suoyi/kejian** the development of local enterprise **may** be weak in globalization. Some young people prefer to work at big international companies.

9.

李玲的钱包|**Ø/听说/不幸**|在地铁|上被小偷偷走了，|**所以/可见**|她当时|**一定**|没有|留意周围的人。|她的同伴也没有觉察到异常。

Li Ling de qianbao| zai ditie| **Ø/tingshuo/buxing**| shang bei xiaotou touzou le,| **suoyi/kejian**| ta dangshi| **yiding**| meiyou| liuyi zhouwei de ren.| Ta de tongban ye meiyou juechadao yichang.

NAME POSS purse| **Ø/As reported/unfortunately**| at metro| on PASS thief steal ASP(PFV),| **CONJ**| 3SGF| that:time| **must**| NEG| pay:attention surrounding MOD people.| 3SGF MOD companion also NEG detect unusual.

Ø/As reported/Unfortunately Li Ling's purse was stolen by a thief in the metro,

suoyi/kejian she **must** be careless to the surroundings at that time. Her companion did not detect any unusual things either.

10.

这场火灾| Ø|听说|所幸|没有|造成什么人员伤亡，|所以/可见|消防部门|一定|及时地|组织了救援。|引起火灾的原因还在调查中。

Zhe chang huozai| Ø|tingshuo/suoxing| meiyou| zaocheng shenme renyuan shangwang,| **suoyi/kejian**|xiaofang bumen| **yiding**| jishi di| zuzhi le jiuyuan.| Yinqi huozai de yuanyin hai zai diaocha zhong.

This CL fire| Ø/As **reported/fortunately** [NEG| cause any life injury:and:death,| **CONJ**| fire department| **must**| in:time| MOD organize rescue.| Cause fire MOD reason still at investigation in.

Ø/As **reported/Fortunately** this fire didn't cause any injuries or deaths, **suoyi/kejian** the fire department **must** have organized the rescue in time. The cause of the fire is still under investigation.

11.

全部被绑人质|Ø|听说|所幸|已经|安全脱离险境，|所以/可见|公安人员|一定|组织|营救得当。|绑匪已被警方控制。

Quanbu Bei bang renzhi| Ø|tingshuo/suoxing| yijing| anquan tuoli xianjing,| **suoyi/kejian**| gongan renyuan| **yiding**| zuzhi| yingjiu dedang.| Bangfei yi bei jingfang kongzhi.

All PASS kidnap hostage| Ø/As **reported/fortunately** | already| safely leave dangerous:situation,| **CONJ**| police staff| **yiding**| organize| rescue appropriately.| Kidnapper already PASS police control.

Ø/As **reported/Fortunately** all hostage have already safely left the dangerous situation, **suoyi/kejian** the police **must** have organized the rescue appropriately. The kidnappers have been controlled by the police.

Verification statement:

这起绑架案并未伤及人质性命。

This kidnapping case didn't cause any deaths of the hostage. (TRUE)

12.

Condition 1 & 2

画展被盗画作| Ø|听说|所幸|在两周|内就被如数追回，|所以/可见|相关部门|一定|投入了|大量人力追缴。|这些画作多为宋元时期的珍品。

Huazhan beidao huazuo| Ø|tingshuo/suoxing| zai liangzhou| nei jiu bei rushu zhuihui,| **suoyi/kejian**| xiangguan bumen| **yiding**| touru le| daliang renli zhuijiao.| Zhexie zihua duo wei songyuan shiqi de zhenpin.

Exhibition PASS steal paintings| Ø/As **reported/fortunately** | in two weeks| within EMP PASS exactly:the:number recover,| **CONJ**| relevant department| **must**| invest ASP(PFV)| immense:amount effort recover.| These painting most COP Song:and:Yuan dynasty MOD treasure.

Ø/As **reported/Fortunately** the exact number of stolen paintings from the exhibition were recovered within two weeks, **suoyi/kejian** the relevant departments **must** have

spent immense amount of efforts to recover. Most of these paintings were treasures from the Song and Yuan dynasty.

13.

布朗夫妇|**Ø/确实/竟然**|今年|又去了加勒比海地区度假，|**所以/可见**|他们俩|**一定**|很喜欢|热带气候。|潜水运动是此地区一大旅游热点。

Bulang fufu| **Ø/queshi/jingran**| jinnian| you qu le Jialebihai diqu dujia,| **suoyi/kejian**| tamen lia| **yiding**| hen xihuan| redai qihou.| Qianshui yundong shi ci diqu yi da lüyou redian.

Brown couple| **Ø/truly/surprisingly**| this:year| again go ASP(PFV) Caribbean region on:vacation,| **CONJ**| 3PL two| **must**| very like| tropical climate.| Diving sport COP this area one big tourist attraction.

Ø/Truly/Surprisingly Mr.& Mrs. Brown travelled to the Caribbean for holiday this year again, **suoyi/kejian** they two **must** enjoy the tropical climate. Diving is one of the biggest attractions in this area.

Verification statement:

布朗夫妇今年第一次去加勒比海地区度假。

It was the first time for Mr. & Mrs. Brown to visit the Caribbean. (FALSE)

14.

这些学生|**Ø/确实/竟然**|一整天|都在讨论课程论文，|**所以/可见**|学生们|**可能**|快到提交论文的期限了。|期末考试也迫在眉睫。

Zhexie xuesheng| **Ø/queshi/jingran**| yi zhengtian| dou zai taolun kecheng lunwen,| **suoyi/kejian**| xueshengmen| **kengneng**| kuai dao| tijiao lunwen de qixian le.| Qimo kaoshi ye pozaimiejie.

These student| **Ø/truly/surprisingly**| one whole:day| EMP ASP(IPFV) revise master course essay,| **CONJ**| students| **may**| soon reach| submit essay MOD deadline ASP(PFV).| Final exam also imminent.

Ø/Truly/Surprisingly these students have been discussing on their master course essays for the whole day, **suoyi/kejian** students **may** have an approaching deadline. The final exam is also imminent.

15.

刘子明|**Ø/确实/竟然**|坐在|第一排都看不清楚黑板，|**所以/可见**|他的视力|**一定**|不太|理想。|他被建议配一副近视眼镜。

Liu Ziming| **Ø/queshi/jingran**| zuo zai| diyi pai dou kan bu qingchu heiban,| **suoyi/kejian**| ta de shili| **yiding**| bu tai| lixiang.| Ta bei jianyi pei yi fu jinshi yanjing.

NAME| **Ø/truly/surprisingly**| sit at| first row even see NEG clear blackboard,| **CONJ**| 3SGM MOD eyesight| **must**| NEG very| ideal.| He PASS suggest get one CL shortsighted glasses.

Ø/Truly/Surprisingly Liu Ziming can't see the blackboard clearly even at the first row, **suoyi/kejian** his eyesight **must** be very limited. He is suggested to get a pair of glasses for shortsighted.

16.

乔安娜|Ø/确实/居然|每天|至少喝五杯加浓的咖啡，|所以/可见|她对咖啡因|可能|非常|上瘾。|她的办公室同事更偏爱喝茶。

Qiaoanna| Ø/**queshi/juran**| meitian| zhishao he wu bei jianong de kafei,| **suoyi/kejian**| ta dui kafeiyin| **keneng**| feichang| shangyin. | Ta de bangongshi tongshi ze pianai hecha.

NAME| Ø/**truly/unexpectedly**| everyday| at:least drink five CL special:strong MOD coffee,| **CONJ**| 3SGF| for caffeine | **may**| very| addicted. | 3SGF MOD office mate more prefer drink:tea

Ø/**Truly/Unexpectedly** Joana drinks at least five cups of strong coffee a day, **suoyi/kejian** she **may** be very addicted to caffeine. Her office mate prefers tea.

17.

罗伯特|Ø/确实/居然|坚持|健身已经有一年多了，|所以/可见|他对身材|一定|是很|在乎的。|他的形象也有了显著改变。

Luobote| Ø/**queshi/juran**| jianchi| jianshen| yijing you yinian duo le,| **suoyi/kejian**| ta dui shencai| **yiding**| shi hen| zaihu de. | Ta de xingxiang ye you le xianzhu gaibian.

NAME| Ø/**truly/unexpectedly**| insist:on| exercise| already have one:year ASP(PFV),| **CONJ**| 3SGM for fitness| **must**| COP very| care PRT. | 3SGM MOD appearance also have ASP(PFV) notably change.

Ø/**Truly/Unexpectedly** Robbert has been exercising for more than one year, **suoyi/kejian** he **must** care about being fit. His appearance has changed notably.

18.

这家餐馆|Ø/确实/居然|需要|提前一个月预约才有位置，|所以/可见|这里的菜|一定|非常|可口。|餐馆提供免费的停车位。

Zhe jia canguan| Ø/**queshi/juran**| xuyao| tiqian yi ge yue yuyue caiyou weizhi,| **suoyi/kejian**| zheli de cai| **yiding**| feichang| kekou. | Canguan tigong mianfei de tingchewei.

This CL restaurant| Ø/**truly/unexpectedly**| need| in:advance one CL month reserve only have seat,| **CONJ**| here MOD food| **must**| very| delicious. | Restaurant provide free MOD parking:place.

Ø/**Truly/Unexpectedly** one needs to make reservation one month in advance for this restaurant, **suoyi/kejian** the food here must be delicious. The restaurant provides free parking place.

19.

赵雅莉|Ø/确实/不幸| 没有|被她报考的大学录取，|所以/可见|她高考|可能|发挥得|不太好。|她所在的中学成绩也不如往年。

Zhao Liya| Ø/**queshi/buxing**| meiyou| bei ta baokao de daxue luqu,| **suoyi/kejian**| ta gaokao| **keneng**| fahui de| bu tai hao| Ta suozai de zhongxue chengji ye buru wangnian.

NAME| Ø/**truly/unfortunately**| NEG| PASS 3SGF apply MOD university accept,| **CONJ**| 3SGF college:entrance:exam| **may**| perform| NEG very good .| 3SGF belong:to MOD middle:school record also not:as:good:as previous:years.

Ø/**Truly/Unfortunately** Zhao Liya was not accepted by the university she applied,

suoyi/kejian she **may** have performed not very well in the college entrance exam. The record of her middle school was also worse than previous years.

20.

历史上庞贝|Ø/确实/不幸|接连|遭遇地震和火山爆发，|所以/可见|它的位置|一定|正处在|地壳活跃地带。|罗马也曾多次受到地震威胁。

Lishi shang Pangbei | Ø/queshi/buxing| jielian| zaoyu dizhen he huoshan penfa,| **suoyi/kejian**| ta de weizhi| **yiding**| zheng chuzai| diqiao huoyue didai.| Luoma ye ceng duoci shoudao dizhen weixie.

History on Pompeii | Ø/truly/unfortunately| in:succession| suffer earthquake and volcano eruption,| **CONJ**| 3SG MOD position| **must**| right situate:at| crust active area.| Rome also before several:times PASS earthquake threaten.

Ø/Truly/Unfortunately Pompeii suffered earthquake and volcano eruption successively in history, **suoyi/kejian** its location **must** be right at an active area of the earth crust. Rome was also threatened by earthquakes several times.

Verification statement:

庞贝并非意大利唯一受地震威胁的城市。

Pompeii is not the only city in Italy that threatened by earthquakes. (TRUE)

21.

于丽娟|Ø/确实/不幸|没有|通过最后一轮面试，|所以/可见|公司上层|可能|已有了|更中意的人选。|这家公司的录取率大概为一比十。

Yu Lijuan| Ø/queshi/buxing| meiyou| tongguo zuihou yi lun mianshi,| **suoyi/kejian**| gongsi shangceng| **keneng**| yi you le| geng zhongyi de renxuan.| Zhe jia gongsi de luqulu dagai wei yibishi.

NAME| Ø/truly/unfortunately| NEG| pass last one CL interview,| **CONJ**| company managers| **may**| already have ASP(PFV)| more desirable MOD candidate.| This CL company MOD admission:rate approximately COP ten:percent.

Ø/Truly/Unfortunately Yu Lijuan didn't get through the last round interview, **suoyi/kejian** the managers of the company **may** already have better candidates. The admission rate of this company is approximately ten percent.

22.

这个小男孩|Ø/确实/所幸|在地震|中只受了点轻微的皮肉伤，|所以/可见|他的父母|可能|将他|保护得很好。|他的母亲伤势严重。

Zhe ge xiao nanhai| Ø/queshi/suoxing| zai dizhen | zhong zhi shou le dian qingwei de piroushang,| **suoyi/kejian**| ta de fumu| **keneng**| jiang ta| baohu de henhao.| Ta de muqin shangshi yanzhong.

This CL small kid| Ø/truly/fortunately| in earthquake| only suffer ASP(PFV) a:little light MOD wound,| **CONJ**| 3SGM MOD parents| **may**| JIANG 3SGM| protect PRT very:well.| 3SG MOD mother injury serious.

Ø/Truly/Fortunately the kid was only slightly wounded in the earthquake, **suoyi/kejian** his parents **may** protect him very well. His mother was seriously injured.

23.

杨丽娜|Ø/确实/所幸|并未|被舍友传染上水痘，|所以/可见|她的预防|一定|很及时|有效。|成年人得水痘远比幼童危险。

Yang Lina| Ø/**qushi/suoxing**| bing wei| bei sheyou chuanran shang shuidou,| **suoyi/kejian**| ta de yufang| **yiding**| hen jishi| youxiao.| Chengnianren de shuidou yuan bi youtong weixian.

NAME| Ø/**truly/fortunately**| EMP NEG| PASS roommate infect on chickenpox,| **CONJ**| 3SGF MOD precaution| **must**| very in:time| effective.| Adult infect chickenpox far compare:to children endangered.

Ø/**Truly/Fortunately** Yang Lina has not been infected with chickenpox by her roommate, **suoyi/kejian** her precaution **must** be in time and effective. Adults are far more endangered than children when infected with chickenpox.

24

此次原油泄漏|Ø/确实/所幸|只污染|了很小一片海域，|所以/可见|这块区域|一定|洋流|活动不强。|一次原油泄漏的影响往往会持续多年。

Zheci yuanyou xielou| Ø/**qushi/suoxing**| zhi wuran| le henxiao yi pian haiyu,| **suoyi/kejian**| zhe kuai quyul| **yiding**| yangliu| huodong bu qiang.| Yici yuanyou xielou de yingxiang wangwang hui chixu henduo nian.

This oil spill| Ø/**truly/fortunately**| only pollute| ASP(PFV) very:small one CL sea:area,| **CONJ**| this CL area| **must**| ocean:current| activity NEG strong.| One:time oil spill MOD effect often would last many years.

Ø/**Truly/Fortunately** the oil spill this time only polluted a very limited sea area, **suoyi/kejian** the ocean current of this area **must** be inactive. The effect of oil spill of one time would often last for many years.

Verification statement:

此次原油泄漏造成了严重的污染。

The oil spill this time caused serious pollution. (FALSE)

25.

这家店门口|Ø/显然/竟然|有很多|顾客排着长队抢购，|所以/可见|这家店|可能|在搞|促销活动。|电子产品往往最受顾客青睐。

Zhe jia menkou| Ø/**xianran/jingran**| you henduo| guke pai zhe changdui qianguo,| **suoyi/kejian**| zhe jia dian| **keneng**| zai gao| cuxiao huodong.| Dianzi chanpin wangwang zui shou guke qinglai.

This CL shop outdoor| Ø/**apparently/surprisingly**| have many| customers line ASP(IPFV) long:queue shopping:rush,| **CONJ**| this CL shop| **may**| ASP(IPFV) make| sale activity.| Electronic product often most PASS customer favor.

Ø/**Apparently/Surprisingly** many customers are lining up in a long queue for a shopping rush outside the shop, **suoyi/kejian** so it **may** be having a sale. Electronic products are often favored most.

Verification statement:

促销活动中电子产品无人问津。

The electronic products do not attract too much attention in this promotion. (FALSE)

26.

这家工厂|**Ø/显然/竟然**|一直在|往附近的河流排放污水，|**所以/可见**|它的厂主|**可能**|并不|关心环境保护。|周围居民已经多次举报污染问题。

Zhe jia gongchang| **Ø/xianran/jingran**| yizhi zai| wang fujin de heliu paifang wushui,| **suoyi/kejian**| ta de changzhu| **keneng**| bing bu| guanxin huanjing baohu.| Zhouwei jumin yijing duoci jubao wuran wenti.

This CL factory| **Ø/apparently/surprisingly**| always ASP(IPFV)| towards nearby MOD rivers discharge polluted:water,| **CONJ**| 3SGN MOD owner| may| EMP NEG| care environment protection.| Around resident already many:times report pollution problem.

Ø/Apparently/Surprisingly the factory has been polluting the rivers nearby, **suoyi/kejian** its owner **may** not care about environment protection. The residents around have reported the pollution problem several times.

Verification statement:

当地居民受到水污染问题困扰。

The local residents were annoyed by the water pollution. (TRUE)

27.

夏洛特|**Ø/显然/竟然**|最近|一直在买含脂量低的酸奶喝，|**所以/可见**|她这几天|**可能**|正在|努力减肥。|低脂酸奶的味道略差于普通酸奶。

Xialuote| **Ø/xianran/jingran**| zuijin| yizhi zai mai hanzhiliang di de suannai he,| **suoyi/kejian**| ta zhexie tian| **keneng**| zheng zai| nuli jianfei.| Dizhi suannai de weidao lue chayu putong suannai.

NAME| **Ø/apparently/surprisingly**| recently| always at buy fat low MOD yogurt drink,| **CONJ**| 3SGF these days| **may**| right ASP(IPFV) |work:hard loss:weight.| Low:fat yogurt MOD slightly taste worse:than regular yogurt.

Ø/Apparently/Surprisingly Charlotte always buys yogurt with low fat recently, **suoyi/kejian** she **may** be on diet these days. Low fat yogurt is less tasty than regular yogurt.

28.

这个镇子|**Ø/显然/居然**|没有|本地的小学中学，|**所以/可见**|这个地方|**可能**|人口规模很小。|学生要到十几公里外上学。

Zhege zhenzi| **Ø/xianran/juran**| meiyou| bendi de xiaoxue zhongxue,| **suoyi/kejian**| zhege difang| **keneng**| renkou| guimo henxiao.| Xuesheng yao dao shiji gongli wai qu shangxue.

This village| **Ø/apparently/unexpectedly**| NEG| local MOD primary:school middle:school,| **CONJ**| this place| **may**| population| scale very small.| Student must go more:than:ten kilometer outside school.

Ø/Apparently/Unexpectedly this village doesn't have local primary schools or middle schools, **suoyi/kejian** this place **may** be small in population scale. Students must go to school more than ten kilometers away.

29.

李晓涵|**Ø/显然/居然**|能说|一口流利的意大利语，|**所以/可见**|她学语言|**一定**|花了很多精力。|她的英语也不错。

Li Xiaohan| **Ø/xianran/juran**| neng shuo| yi kou liuli de yidaliyu,| **suoyi/kejian**| ta xue yuyan| **yiding**| hua le| henduo jingli.| Ta de yingyu ye bucuo.

NAME| **Ø/apparently/unexpectedly**| can speak| one CL fluent MOD Italian,| **CONJ**| she learn language| **must**| spend ASP(PFV)| very:much effort.| 3SGF POSS English also good.

Ø/Apparently/Unexpectedly Li Xiaohan can speak very fluent Italian, **suoyi/kejian** she **must** have spent a lot of efforts in learning language. She also speaks good English.

Verification statement:

李晓涵至少会两门外语。

Li Xiaohan speaks at least two foreign languages. (TRUE)

30.

罗小兰|**Ø/显然/居然**|没有|把卢浮宫列入巴黎游览计划，|**所以/可见**|她的爱好|**可能**|并不在|古典艺术上。|她计划船游塞纳河。

Luo Xiaolan| **Ø/xianran/juran**| meiyou| ba lufugong lieru bali youlan jihua,| **suoyi/kejian**| ta de aihao| **keneng**| bing bu zai| gudian yishu shang.| Ta jihua chuanyou Sainahe.

NAME| **Ø/apparently/unexpectedly**| NEG| BA Louvre list Paris travel schedule,| **CONJ**| 3SGF MOD interest| **may**| EMP NEG at| ancient art on.| 3SGF plan travel:by:boat the:Seine.

Ø/Apparently/Unexpectedly Luo Xiaolan didn't include Louvre in her travel schedule for Paris, **suoyi/kejian** she **may** have no interest in ancient art. She planned to travel along the Seine by boat.

31.

新奥尔良|**Ø/显然/不幸**|很多年|都未从飓风灾难中恢复过来，|**所以/可见**|那场飓风|**一定**|破坏力|极强。|当年不少居民被迫远走他乡。

Xinao'erliang| **Ø/xianran/buxing**| henduo nian| dou wei cong jufeng zainan zhong huifu guolai,| **suoyi/kejian**| na chang jufeng| **yiding**| pohuai li| ji qiang.| Dangniang bu shao jumin beipo yanzoutaxiang.

New Orleans| **Ø/apparently/unfortunately**| many year| EMP NEG from hurricane disaster in recover back,| **CONJ**| that CL hurricane| **must**| destructive power|extremely strong.| That:year NEG few resident be:forced leave:home.

Ø/Apparently/Unfortunately New Orleans hasn't recovered from the hurricane after so many years, **suoyi/kejian** the destructive power of the hurricane **must** be extremely strong. Back then, a lot of residents were forced to leave home.

Verification statement:

很多居民受飓风影响离开了新奥尔良。

Many residents left New Orleans because of the hurricane. (TRUE)

32.

史蒂夫|Ø/显然/不幸|错过了|上午起飞的飞机，|所以/可见|他早上|可能|起得太晚了。|这已经不是他第一次错过飞机了。

Shidifu| Ø/xianran/buxing| cuoguo le| shangwu qifei de feiji,| suoyi/kejian| ta zaoshang| keneng| qi de| tai wan le.| Zhe yijing bu shi ta diyici cuoguo feiji le.

NAME| Ø/apparently/unfortunately| miss ASP(PFV)| morning departure MOD flight,| CONJ| he morning| may| get:up PRT| too late ASP(PFV).| This already NEG COP 3SGM first:time miss flight PRT.

Ø/Apparently/Unfortunately Steve missed the morning flight, CONJ he may get up too late in the morning. This was not his first time missing his flight.

33.

孙国嘉|Ø/显然/不幸|没能|按时完成项目任务，|所以/可见|他最近|可能|工作|状态不佳。|他的妻子并不理解他的工作压力。

Sun Guojia| Ø/xianran/buxing| mei neng| anshi wancheng xiangmu renwu, suoyi/kejian| ta zuijin| keneng| gongzuo| zhuangtai bu jia.| Ta de qizi bing bu lijie ta de gongzuo yali.

NAME| Ø/apparently/unfortunately| NEG able:to| in:time complete project task,| CONJ| 3SGM recently| may| work| state| NEG good.| 3SGM MOD wife EMP NEG understand 3SGM MOD work pressure.

Ø/Apparently/Unfortunately Sun Guojia was unable to complete his task of the project in time, suoyi/kejian he may be in a bad state recently. His wife doesn't understand his work pressure.

Verification statement:

孙国嘉的妻子很支持他的工作。

Sun Guojia's wife supports him in work. (FALSE)

34.

金融危机|Ø/显然/所幸|没有|给香港带来太大损失，|所以/可见|金融机构|一定|施行了|有效的应对措施。|东南亚很多国家损失惨重。

Jinrong weiji| Ø/xianran/suoxing| meiyou| gei xianggang dailai tai da sunshi,| suoyi/kejian| jinrong jigou| yiding| shixing le| youxiao de yingduicuoshi.| Dongnanya henduo guojia sunshi canzhong.

Financial crisis| Ø/apparently/fortunately| NEG| get Hongkong bring very big loss,| CONJ| financial organization| must| apply ASP(PFV)| effective MOD countermeasures.| Southeast:Asia many country loss badly.

Ø/Apparently/Fortunately the financial crisis didn't bring Hongkong big loss, suoyi/kejian the financial organizations must have applied effective countermeasures. Many countries in Southeast Asia suffered great loss.

35.

这场事故|Ø/显然/所幸|并未|伤及车内的乘客，|所以/可见|当时车速|可能|不是|太快。|事故造成了公路被暂时封闭。

Zhe chang shigu| Ø/xianran/suoxing| bing wei| shangji chenei de chengke,|

suoyi/kejian dangshi chesu| **keneng** bu shi| tai kuai.| Shigu zaocheng le gonglu bei zanshi fengbi.

This CL accident| **Ø/apparently/fortunately** EMP NEG| hurt inside:car MOD passenger,| **CONJ** that:time car:speed| **may** NEG COP| too fast.| Accident cause ASP(PFV) road PASS temporarily closed.

Ø/Apparently/Fortunately this accident didn't hurt the passengers, **suoyi/kejian** the car speed **may** be not very fast at that time. Due to the accident the road is temporarily closed.

36.

卢谦祥|**Ø/显然/所幸**|在未知|丛林中也并没有迷路, |**所以/可见**|他认路本领|一定|非常|过硬。|与他同行的人主要依靠他找方向。

Lu Qianxiang| **Ø/xianran/suoxing** zai weizhi| conglin zhong ye bing meiyou milu,| **suoyi/kejian** ta renlu benling| **yiding**|feichang| guoying.| Yu ta tongxing de ren zhuyao yikao ta zhao fangxiang.

NAME| **Ø/apparently/fortunately** at unknown| jungle inside still EMP NEG get:lost,| **CONJ** 3SGM navigation skill| **must**| very| sophisticated.| With 3SGM walk:together MOD people mainly rely:on 3SGM find direction.

Ø/Apparently/Fortunately Lu Qianxiang didn't get lost in an unknown jungle, **suoyi/kejian** his navigation skills **must** be sophisticated. The people with him mainly relied on him to find the direction.

37.

郑亦凡|**Ø/据说/竟然**|在一个|月之内三次到国外出差, |**所以/可见**|他们公司|一定|有跨国|业务。|他本人并不喜欢出差。

Zheng Yifan| **Ø/jushuo/jingran** zai yi ge| yue zhinei san ci dao guowai chuchai,| **suoyi/kejian** tamen gongsi| **yiding** you kuaguo| yewu.| Ta benren bing bu xihuan chuchai.

NAME| **Ø/it:is:said/surprisingly** at one CL| month within three times go abroad on:business,| **CONJ** 3PM company| **must** have international| business.| 3SGM himself EMP NEG like business:trip.

Ø/It is said (that)/Surprisingly Zheng Yifan travels abroad for work three times within one month, **suoyi/kejian** their company must have international business. He doesn't like business trip himself.

38.

张安华|**Ø/据说/竟然**|在自己|经济困难时还资助朋友, |**所以/可见**|他这个人|一定|有副|热心肠。|朋友们都对他评价很高。

Zhang Anhua| **Ø/jushuo/jingran** zai ziji| jingji kunnan shi hai zizhu pengyou,| **suoyi/kejian** ta zhe ge ren| **yiding** you fu| re xinchang.| Pengyoumen dui ta dou pingjia hen gao.

NAME| **Ø/it:is:said/surprisingly** at self| economy tough time still financial:support friend,| **CONJ** 3SGM this person| **must** have CL| warm heart.| Friends towards 3SGM all comment very high.

Ø/It is said (that)/Surprisingly Zhang Anhua still support his friends when he is on financial budget himself, **suoyi/kejian** he **must** have a warm heart. His friends all

speak highly of him.

39.

刘伊珊|Ø/据说/竟然|每年|有一整个月的带薪年假，|所以/可见|她的公司|一定|给员工|的福利不错。|她通常利用年假时间陪伴家人。

Liu Yishan| Ø/jushuo/jingran| meinian| you yi zhengge yue de daixin nianjia,| suoyi/kejian| ta de gongsi| yiding| gei yuangong| de fuli bucuo.| Ta tongchang yong nianjia shijian peiban jia ren.

NAME| Ø/it:is:said/surprisingly| every:year| have one whole month MOD with:salary vacation,| CONJ| 3SGF MOD company| must| provide employee| MOD welfare NEG:bad.| She usually use vacation accompany family.

Ø/It is said (that)/Surprisingly Liu Yishan has one whole month vacation with salary ever year, suoyi/kejian her company must provide good welfare for its employees. She usually spends the vacation with her family.

40.

丹尼尔|Ø/据说/居然|在主修|课程之外又选修了天文学，|所以/可见|他对宇宙|一定|有着|浓厚的兴趣。|那门课并没有太多人选。

Danni'er| zai zhuxiu| Ø/jushuo/juran| kecheng zhiwai you xuanxiu le tianwenxue,| suoyi/kejian| ta dui yuzhou| yiding| you zhe| nonghou de xingqu.| Na men ke bing meiyou tai duo ren xuan.

NAME| Ø/it:is:said/unexpectedly| at major| course besides also choose:practice ASP(PFV) astronomy course,| CONJ| 3SGM for the:universe| must|have ASP(IPFV)| huge MOD interest.| That CL course EMP NEG too many people choose.

Ø/It is said (that)/Unexpectedly Daniel followed astronomy as a minor course besides his major, suoyi/kejian he must have a huge interest in the universe. Not many people chose that course.

Verification statement:

丹尼尔主修天文学。

Daniel studies astronomy as his major. (FALSE)

41.

张小倩|Ø/据说/居然|在一年|之内升职两次，|所以/可见|她的工作|一定|深受|老板认可。|毕业两年就有这样的成绩实在难得。

Zhang Xiaoqian| Ø/jushuo/juran| zai yinian| zhinei shengzhi liang ci,| suoyi/kejian| ta de gongzuo| yiding| shen shou| laoban renke.| Biye liangniang jiu you zheyang de chengji shizai nande.

NAME| Ø/it:is:said/unexpectedly| during one:year| within promote two time,| CONJ| 3SGF MOD work| must| deeply PASS| boss appreciate.| Graduate two:years EMP have this:kind MOD achievement indeed impressive.

Ø/It is said (that)/Unexpectedly Zhang Xiaoqian has been promoted two times in one year, suoyi/kejian her work must be deeply appreciated by her boss. It is impressive to have this kind of achievement two years after graduation.

Verification statement:

张小倩有多年工作经验。

Zhang Xiaoqian has been working for many years. (FALSE)

42.

今年冬季均温|**Ø/据说/居然**|比往年|低了整整三度，|**所以/可见**|全球变暖|**可能**|并非|绝对趋势。|人类活动只是影响气候的众多因素之一。

Jinniang dongji junwen| **Ø/jushuo/juran**| bi wangnian| di le zhengzheng san du,| **suoyi/kejian**| quanqiu biannuan| **keneng**| bing fei| juedui qushi.| Renlei huodong zhi shi yingxiang qihou de zhongduo yinsu zhiyi.

This: year winter average:temperature| **Ø/it:is:said/unexpectedly**| compare last:year| lower ASP(PFV) full three degree,| **CONJ**| global warming| **may**| EMP NEG| absolute tendency.| Human activity only COP influence climate MOD many factor one:of.

Ø/It is said (that)/Unexpectedly the average temperature this winter is three degrees lower than last year, **suoyi/kejian** global warming **may** not be an absolute tendency. Human activity is just one of the many factors influencing the climate.

43.

舟曲泥石流|**Ø/据说/不幸**|造成了|五万多人流离失所，|**所以/可见**|这场泥石流|**一定**|覆盖的|面积很广。|很多村落被夷为平地。

Zhouqu nishiliu| **Ø/jushuo/buxing**| zaocheng le| wu wan duo ren liulishisuo,| **suoyi/kejian**| zhe chang nishiliu| **yiding**| fugai de| mianji hen guang.| Hendo cunluo bei yiweipingdi.

PLACE mud flow| **Ø/it:is:said/unfortunately**| cause ASP(PFV)| five more ten:thousand people homeless,| **CONJ**| this CL mud flow| **must**| cover MOD| area very vast.| Many village PASS raze:to:the:ground.

Ø/It is said (that)/Unfortunately the Zhouqu mud flow made more than fifty thousand of people homeless, **suoyi/kejian** this mud flow **must** have covered a vast area. Many villages were razed to the ground.

44.

王小明|在上个|**Ø/据说/不幸**|月滑雪时摔伤了胳膊，|**所以/可见**|他当时|**可能**|是没太|留神。|之前他从未受过伤。

Wang Xiaoming| **Ø/jushuo/buxing**| zai shang ge| yue huaxue shi shuaishang le gebo,| **suoyi/kejian**| ta dangshi| **keneng**| shi mei tai |liushen.| Zhiqian ta congwei shou guo shang.

Name| **Ø/it:is:said/unfortunately**| at last CL| month ski time hurt ASP(PFV) arm,| **CONJ**| 3SGM that:time| **may**| COP NEG very | careful.| Before 3SGM never PASS ASP(PFV) injure.

Ø/It is said (that)/Unfortunately Wang Xiaoming hurt his arm in a skiing activity, **suoyi/kejian** he may have been not very careful at that time. He has never been injured before.

45.

瓦萨号战舰|**Ø/据说/不幸**|首航|不到一海里便沉没了，|**所以/可见**|这艘船|**可能**

有严重|的结构缺陷。|沉船直到三百年之后才被打捞出来。

Wasahao zhanjian| **Ø/jushuo/buxing**| shouhang| budao yi haili bian chenmo le,| **suoyi/kejian**|zhe sou chuan| **keneng**| you yanzhong| de jiegou quexian.| Chenchuan zhidao sanbai nian hou cai bei dalao chulai.

Vasa warship| **Ø/it:is:said/unfortunately**| first:voyage| less:than one sea:mile EMP sink ASP(PFV),| **CONJ**| this CL ship| **may**| have serious| MOD structure defect. | sink:ship till three:hundred year later only PASS salvage out.

Ø/It is said (that)/Unfortunately the Vasa warship sank after sailing for less than one sea mile in her first voyage, **suoyi/kejian** this warship **may** have serious structural defects. The ship was not salvaged from water until three hundred years later.

Verification statement:

瓦萨号沉没后没有马上被打捞出来。

The Vasa warship was not salvaged right after she sank. (TRUE)

46.

刘云英|**Ø/据说/所幸**|五分钟内就顺利逃离了火灾现场，|**所以/可见**|她以前|**可能**|接受过|逃生训练。|消防员赶到时楼已经烧了半边。

Liu Yunying| **Ø/jushuo/suoxing**| wu fenzhong| nei jiu shunli taoli le huozai xianchang,| suoyi/kejian| ta yiqian| **keneng**| jieshou guo| taosheng xunlian.| Xiaofangyuan gandao shi lou yijing shao le banbian.

NAME| **Ø/it:is:said/fortunately**| five minute| within EMP successfully escape ASP(PFV) fire scene,| **CONJ**| 3SGF before| **may**| receive ASP(PFV)| evacuation training.| Fireman arrive when building already on:fire ASP(PFV) half.

Ø/It is said (that)/Fortunately Liu Yunying successfully escaped from the fire accident within five minutes, **suoyi/kejian** she **may** have received evacuation training before. Half of the building was already on fire when the firemen arrived.

47.

何玉兰|**Ø/据说/所幸**|很快|就从离婚阴影中走了出来，|**所以/可见**|她这个人|**一定**|很积极|乐观。|她的朋友们给了她很多鼓励与支持。

He Yulan| **Ø/jushuo/suoxing**| hen kuai| jiu cong lihun yinying zhong zou le chulai,| **suoyi/kejian**| ta zhe ge ren| **yiding**| hen jijì| leguan.| Ta de pengyoumen gei le ta henduo guli yu zhichi.

NAME| **Ø/it:is:said/fortunately**| very fast| just from divorce shadow inside walk ASP(PFV) out,| **CONJ**| 3SGF this CL person| **must**| very positive| optimistic.| 3SGF MOD friends give ASP(PFV) 3SGF a:lot encouragement and support.

Ø/It is said (that)/Fortunately he Yulan walked out from the shadow of divorce very fast, **suoyi/kejian** she as a person **must** be very positive and optimistic. Her friends offered her a lot of encouragement and support.

48.

这栋大楼|**Ø/据说/所幸**|在八级|地震中都未严重受损，|**所以/可见**|其建筑质量|**一定**|是符合|标准的。|当年工程的造价也相当可观。

Zhe dong dalou| **Ø/jushuo/suoxing**| zai baji| dizhen zhong dou wei yanzhong

shousun,| **suoyi/kejian**| qi jianzhu zhiliang| **yiding**| shi fuhe| biao zhun de.| Dangnian gongcheng de zaojia ye xiangdang keguan.

This CL building| **Ø/it:is:said/fortunately**| at eight:degree| earthquake even NEG serious damage,| **CONJ**| that architecture quality| **must**| COP conform:to| standard PRT.| That:year construction MOD cost also quite considerable.

Ø/It is said (that)/Fortunately this building was not severely damaged even in an eight-degree earthquake, **suoyi/kejian** the architectural quality of it **must** have conformed to the standard. Back then the construction cost was also quite considerable.

Samenvatting in het Nederlands

Causale connectieven en perspectiefmarkeerders in het Chinees: corpus- en verwerkingsstudies naar subjectiviteit in discourse

1. Subjectiviteit in causale relaties

Mensen communiceren met elkaar in talige eenheden die samen een coherent geheel vormen: we spreken van tekst of *discourse*. Hierbij vervullen taalgebruikers in al hun uitingen de rol van verslaggever of verteller, als ze beschrijven wat er in de wereld om hen heen gebeurt of onder woorden brengen wat zij vinden van of voelen bij een bepaalde situatie. In sommige uitingen vervullen sprekers en schrijvers ook nog een andere rol. Waar (1) slechts de beschrijving van een gebeurtenis is, verwoordt de spreker in (2) en (3) een eigen mening of conclusie. Dit kan met subtiele indicatoren zoals *misschien* in (2) duidelijk gemaakt worden, maar de spreker kan zich ook expliciet presenteren door bijvoorbeeld *ik denk* te gebruiken, zoals in (3). Daarmee kan de spreker zelf verantwoordelijk gehouden worden voor de conclusie dat ‘Marie boos is op Jan.’ Met andere woorden: uiting (1) is objectief, waar (2) en (3) subjectief zijn.

- (1) Marie is boos op Jan.
- (2) Misschien is Marie boos op Jan.
- (3) Ik denk dat Marie boos is op Jan.

Subjectiviteit is een veelomvattend talig fenomeen, maar ook een belangrijke cognitieve notie, die vaak gedefinieerd wordt als de betrokkenheid van een spreker in een uitspraak (Finegan, 1995). Dit proefschrift richt zich op subjectiviteit in coherentierelaties. In causale coherentierelaties kun je de mate van subjectiviteit gebruiken om twee typen relaties te onderscheiden: objectieve relaties die gebeurtenissen in de werkelijkheid verbinden in een oorzaak-gevolg- of gevolg-oorzaakconstructie, en subjectieve relaties die argumenten met beweringen of andere taalhandelingen van een spreker verbinden (Sanders, 1997; Sweetser, 1990).

Om subjectiviteit te kunnen interpreteren, moeten hoorders/lezers rekening houden met de aanwezigheid van een *Subject of Consciousness* (SoC), iemand die verantwoordelijk is voor de evaluatie, het oordeel of de redenering in de zin. Dat kan de spreker zijn, maar ook een personage in het verhaal (Sanders, Sanders & Sweetser, 2009). De aanwezigheid van een SoC (al dan niet expliciet in de zin vermeld) leidt tot een ander verwerkingspatroon voor subjectieve relaties zoals (4a), in vergelijking met objectieve relaties zoals (4b) (Traxler, Sanford, Aked, & Moxey, 1997).

- (4)
- a. Susan lost her money and credit cards because she left her purse at the bus stop.
Susan verloor haar geld en creditcards, omdat ze haar portemonnee liet liggen bij de bushalte.
 - b. Susan was careless with money and credit cards because she left her purse at the bus stop.
Susan ging onzorgvuldig om met haar geld en creditcards, aangezien ze haar portemonnee liet liggen bij de bushalte.
- (Traxler et al., 1997: 91)

Eerdere studies hebben subjectiviteit en perspectief veelal beschouwd als afzonderlijke concepten. In dit proefschrift heb ik voorgesteld om het concept perspectief te gebruiken als instrument om subjectiviteit te operationaliseren: waar subjectiviteit aanduidt *dat* we te maken hebben met iemands mening of opvatting over de gegeven informatie, kan perspectief weergeven *wie* deze persoon is en *hoe* diegene bij de informatie betrokken is. SoCs kunnen in een uiting onder andere op de voorgrond treden via de dimensie van *epistemic stance* – als zij hun mening of oordeel uitdrukken – of via de dimensie van *attitudinal stance* – als ze hun houdingen of gevoelens weergeven (Conrad & Biber, 2000).

Informatie over de betrokkenheid van een spreker kan met een variëteit aan talige middelen worden uitgedrukt (Traugott, 2003). Sommige talen markeren de subjectiviteit van een relatie met specifieke connectieven. Voorbeelden zijn *dus* en *want* in het Nederlands (Pander Maat & Degand, 2001; Pander Maat & Sanders, 2000; Stukker & Sanders, 2012) of *kejian* ‘dus’ in het Chinees (Li, Evers-Vermeul & Sanders, 2013). Dergelijke talige middelen beïnvloeden de manier waarop subjectieve relaties worden verwerkt. Zo laten Canestrelli, Mak en Sanders (2013) zien dat het subjectieve connectief *want* in (5a) een onmiddellijke verwerkingsvertraging veroorzaakt in vergelijking met het objectieve connectief *omdat* in (5b).

- (5)
- a. Jeff werd boos op zijn burens, **omdat** ze altijd hun stereo hard hadden aanstaan tot diep in de nacht.
 - b. Jeff had asociale burens, **want** ze hadden altijd hun stereo hard aanstaan tot diep in de nacht.

Het verwerkingsverschil door de aanwezigheid van subjectieve dan wel objectieve connectieven kan worden beïnvloed door de aanwezigheid van perspectiefmarkeerders. Als bijvoorbeeld de perspectiefmarkeerder *volgens Lieke* het perspectief van Lieke aanduidt in het segment vóór het connectief, zoals in (5c),

wordt de veroorzaakte verwerkingsvertraging van het subjectieve *want* in vergelijking met het objectieve *omdat* geneutraliseerd (Canestrelli et al., 2013).

- c. Volgens Lieke had Jeff asociale burens, **want** ze hadden altijd hun stereo hard aanstaan tot diep in de nacht.
(Canestrelli et al., 2013: 170)

Het faciliterende effect van perspectiefmarkeerders op de verwerking van subjectieve *want*-relaties ondersteunt de hypothese dat de vertraagde verwerkingstijden na *want* gerelateerd zijn aan de aanwezigheid van iemands perspectief. Het kost lezers meer cognitieve inspanning als zij bij de interpretatie van uitingen rekening moeten houden met iemands perspectief.

Verscheidene vragen over de interpretatie van subjectiviteit en de rol van talige subjectiviteitsindicatoren zijn tot nu toe onbeantwoord gebleven: hoe combineren taalgebruikers verschillende talige indicatoren van subjectiviteit? Hoe dragen de verschillende talige indicatoren van subjectiviteit precies bij aan het begrip van subjectieve relaties als ze in combinatie gebruikt worden? Wat betekent subjectiviteit in termen van de mentale representatie die taalgebruikers van een zin maken? Deze vragen vormen het hart van de huidige studie. Bij het beantwoorden van die vragen is gebruik gemaakt van verschillende onderzoeksmethoden.

Dit proefschrift richt zich op subjectiviteit in causale relaties in het Mandarijn Chinees. Het Mandarijn Chinees heeft een connectiefsysteem met zowel een connectief *suoyi* ‘dus/daardoor’ dat ondergespecificeerd is voor subjectiviteit, als specifieke connectieven die subjectieve respectievelijk objectieve relaties uitdrukken: *kejian* ‘dus’ en *yin’er* ‘daardoor’. Dit connectiefsysteem maakt het mogelijk om de contextuele kenmerken en verwerkingseffecten van connectieven met een verschillende mate van subjectiviteit te onderzoeken.

2. Samenvatting van de belangrijkste bevindingen

Drie empirische studies zijn uitgevoerd om de talige markering en de verwerking van subjectiviteit in coherentierelaties te onderzoeken: een corpusonderzoek (meer specifiek: een collocatiestudie), een oogbewegingsstudie in het *Visual World Paradigm* en een leesexperiment.

2.1 Collocatiestudie

Eerdere corpusstudies richtten zich voornamelijk op voorkeuren in het gebruik van connectieven en de relaties die zij uitdrukken. Degand en Pander Maat (2003) stelden bijvoorbeeld vast dat het connectief *want* vooral gebruikt wordt in subjectieve relaties, terwijl het connectief *omdat* met name voorkomt in objectieve relaties (zie ook Pit, 2003; Sanders & Spooen, 2015). Een vergelijkbaar onderscheid wordt door Li et al. (2013) gerapporteerd voor het Chinees: *kejian* ‘dus’

wordt altijd gebruikt in subjectieve relaties, terwijl *yin'er* ‘daardoor’ samengaat met objectieve relaties; een derde connectief, *suoyi* ‘dus/daardoor’, kan worden gebruikt om beide typen relaties uit te drukken en is daarmee ondergespecificeerd voor subjectiviteit.

Met een exploratieve collocatiestudie heb ik onderzocht hoe verschillende indicatoren van subjectiviteit met elkaar gecombineerd worden. Hierin zijn de contexten van twee Chinese connectieven met een verschillende mate van subjectiviteit vergeleken: het gespecificeerde subjectieve connectief *kejian* ‘dus’ en het generieke, ondergespecificeerde *suoyi* ‘dus/daardoor’. In deze collocatiestudie stonden de volgende onderzoeksvragen centraal:

1. Verschillen connectieven die zich onderscheiden in mate van subjectiviteit in hun collocaties?
2. Meer specifiek: verschillen connectieven in het soort perspectiefmarkeerders waarmee zij gecombineerd worden?

Door de contexten waarin *kejian* in verschillende genres voorkwam direct te vergelijken met die van *suoyi*, bleek dat voornaamwoorden, modale werkwoorden (bijvoorbeeld *keneng* ‘kunnen’) en cognitie-werkwoorden (*renwei* ‘overwegen’, *zhidao* ‘weten’) vaker samengaan met *suoyi*. Uitdrukkingen van verrassing (*jingran* ‘verrassend’) en belang (*zhongyao* ‘belangrijk’, *jiazhi* ‘waarde’) bleken daarentegen onderscheidende collocaties te zijn van *kejian*. Connectieven die een verschillende mate van subjectiviteit uitdrukken verschillen dus in hun soort collocaties – in het bijzonder in de verdeling van perspectiefmarkeerders. *Suoyi* ging vaker samen met *epistemic stance markers*, die aangeven dat er een SoC in het spel is die een bepaalde mate van (on)zekerheid tot uitdrukking brengt, terwijl de collocaties van *kejian* markeerders van *attitudinal stance* waren, die attitudes en gevoelens van de SoC uitdrukken. De invloed van verschillende typen perspectiefmarkeerders op de verwerking van subjectieve coherentierelaties is getest in een leesexperiment (zie 2.3).

2.2 Oogbewegingsstudie in het *Visual World Paradigm*

In eerdere leesonderzoeken is gevonden dat er bij de verwerking van subjectieve relaties een vertraging optreedt (Canestrelli et al., 2013; Traxler et al., 1997). Vanuit een cognitief perspectief is aangevoerd dat de mentale representatie van subjectieve causale relaties, het zogenoemde situatiemodel, een SoC moet bevatten (Sanders, Sanders & Sweetser, 2012). Het identificeren, opbouwen of openen van een mentale representatie van de SoC zou complexer zijn en daarmee meer tijd kosten. In de huidige studie zijn twee oogbewegingsexperimenten in het *Visual World Paradigm* (VWP) uitgevoerd (één in het Nederlands en één in het Chinees) om te achterhalen of de langere leestijden voor het verwerken van subjectiviteit verband houden met

het achterhalen van de SoC voor het situatiemodel. De volgende onderzoeksvragen stonden centraal:

3. Omvat de verwerking van subjectiviteit het achterhalen van de SoC?
4. Wat is hierbij de rol van connectieven en modale werkwoorden?

In de VWP-experimenten zagen de proefpersonen twee afbeeldingen: een afbeelding van de situatie die werd beschreven, en een afbeelding van de persoon die over de situatie sprak (de SoC). Uit de resultaten van de experimenten blijkt dat taalgebruikers de neiging hebben om meer naar de visuele weergave van de spreker te gaan kijken (de relevante SoC in deze VWP-experimenten) als ze een subjectief connectief verwerken dan wanneer ze een objectief connectief verwerken. In beide talen was er meer aandacht voor de foto met een spreker na een subjectief connectief (*dus* en *kejian* ‘dus’) dan na een objectief connectief (*daardoor* en *yin'er* ‘daardoor’). Wanneer in een subjectieve relatie het connectief ondergespecificeerd was qua subjectiviteit (*suoyi* ‘dus/daardoor’), was een modaal werkwoord (*yiding* ‘moet’/ *keneng* ‘kan’) in het vervolg van de zin de eerste aanwijzing dat er sprake was van een subjectieve relatie. In de ondergespecificeerde gevallen leidde het modale werkwoord tot een toename in aandacht voor de spreker in vergelijking met de conditie met een subjectief connectief, *kejian*.

De VWP-bevindingen bevestigen daarmee eerdere verklaringen over de complexiteit van subjectieve relaties: het verwerken van subjectiviteit houdt in dat lezers/luisteraars een SoC in hun mentale representatie opnemen. Dit proces kost cognitieve energie, wat in eerdere studies tot uiting kwam in langere verwerkingstijden voor subjectieve relaties. Hierbij functioneerden talige middelen van subjectiviteit bovendien als verwerkingsinstructies die de aandacht naar de SoC leidden; als het connectief onvoldoende informatie gaf over de mate van subjectiviteit, had het modale werkwoord de rol om de aandacht alsnog naar de spreker te leiden.

2.3 Leesexperiment

Een volgende vraag was of lezers gevoelig zijn voor de verschillende combinaties van connectieven en andere talige markeerders van subjectiviteit die in de collocatiestudie naar voren kwamen. In een leesexperiment onderzocht ik de rol van perspectiefmarkeerders en connectieven tijdens de verwerking van subjectieve relaties. Hierbij stonden de volgende onderzoeksvragen centraal:

5. Hoe beïnvloedt de aanwezigheid van perspectiefmarkeerders de verwerkingspatronen in de connectiefregio en in latere regio's in subjectieve relaties met een argument-claimstructuur?

6. Gelden de effecten van *epistemic stance markers* op de verwerking van subjectieve relaties ook voor *attitudinal stance markers*?

Voor het leesexperiment werden zinnen met een subjectieve causale relatie met een argument-claim-structuur geconstrueerd, waarin het type *stance marker* (geen *stance marker* vs. *epistemic stance marker* vs. *attitudinal stance marker*) en het type connectief (*kejian* vs. *suoyi*) werd gemanipuleerd (zie (6)). Alle uitingen bevatten een modaal werkwoord in het tweede segment om ervoor te zorgen dat deze zin werd geïnterpreteerd als een bewering.

(6)

Chengji tekui *Ø/tingshuo/juran* wudian le yige duo xiaoshi, *suoyi/kejian* tielu xitong *keneng* chu le shigu. Zhezhong qingkuang bing bu changjian. De intercity *Ø/ schijnt/ onverwachts* te zijn:vertraagd ASP(PFV) nog een uur, *suoyi/kejian (CONJ)* railway system *may* happen ASP(PFV) accident. This:kind situation EMP NEG common. *Ø/ Het schijnt dat/ onverwachts* [is] de intercity meer dan een uur vertraagd [is], *suoyi/kejian* de spoorwegen hebben mogelijk een ongeluk [...]. Deze situatie is niet alledaags.

In overeenstemming met de voorspellingen veroorzaakte het subjectieve connectief een onmiddellijke verwerkingsvertraging: langere leestijden werden gevonden voor *kejian* ‘dus’ in vergelijking met het ondergespecificeerde connectief *suoyi* ‘dus’. Dit is in lijn met de bevindingen van Canestrelli et al. (2013) over het subjectieve connectief *want* en met het idee dat subjectieve connectieven een proces in gang zetten waarbij de lezer de SoC probeert te achterhalen. In de onderwerpsregio net na het connectief was *kejian* echter bevorderlijker voor de verwerking dan het ondergespecificeerde *suoyi*. Dit verwerkingsgemak na *kejian* kan zijn veroorzaakt doordat *kejian* lezers de moeite bespaart om meerdere interpretaties van de relatie in het geheugen vast te houden.

De aanwezigheid van perspectiefmarkeerders aan het begin van het eerste segment (het argument in de argument-claim-structuur) beïnvloedde niet de verwerkingspatronen in de connectiefregio, maar wel die van het modale werkwoord in het tweede segment, het talige middel waardoor lezers opnieuw (of voor het eerst) een expliciete aanwijzing kregen dat de relatie subjectief moest zijn. Deze bevinding is belangrijk voor onze kennis over de manier waarop mensen subjectiviteit verwerken: perspectiefmarkeerders die betrekking hebben op de mening van een SoC of die een SoC impliceren door zijn/haar houding erbij te betrekken, fungeren als verwerkingsinstructies die de lezer bewust maken van de aanwezigheid van een SoC. *Attitudinal stance markers* en *epistemic stance markers* verschillen niet in de mate waarin zij de verwerking van modale werkwoorden beïnvloedden.

Tegen de verwachting in was er geen interactie tussen het type perspectiefmarkeerder en het type connectief bij de verwerking van het connectief, het onderwerp van het segment en het modale werkwoord. Het interactie-effect in de laatste regio van het tweede segment was daarentegen wel in overeenstemming met de patronen uit de collocatiestudie. *Attitudinal stance markers*, zoals *jurán* ‘onverwacht’ in (6), kwamen in natuurlijke taaldata vaak voor in combinatie met *kejian*. Precies deze combinatie leidde, in vergelijking met andere combinaties, tot snellere leestijden in de laatste regio van het tweede segment.

Eerdere leesstudies hebben aangetoond dat de aanwezigheid van een SoC (dat wil zeggen: subjectiviteit) het begrip van coherentierelaties beïnvloedt. De huidige studie heeft aan dit onderzoeksveld bijgedragen door te laten zien dat de wijze waarop een SoC op de voorgrond treedt, oftewel de specifieke dimensie van subjectiviteit, eveneens van belang is bij de verwerking van coherentierelaties.

3. Theoretische kwesties en vervolgonderzoek

De bevindingen uit drie studies met een verschillende methodologie (i.e., corpusonderzoek, VWP-oogbewegingsonderzoek en registratie van leespatronen) verfijnen ons beeld van subjectiviteit. Ze bieden inzicht in een aantal theoretische kwesties, zoals de mentale representatie van subjectiviteit, de weergave van subjectiviteit in taalgebruik en de rol van deze talige middelen in taalverwerking.

3.1 Subjectiviteit en perspectief

De empirische bevindingen uit de VWP-studie zijn in lijn met de hypothese dat het interpreteren van subjectiviteit onder andere inhoudt dat de lezer een mentale representatie van de SoC maakt. Dit vereist dat lezers de informatie in een uiting verankeren in het perspectief van een spreker of van een andere SoC die verantwoordelijk is voor de subjectieve relatie.

De collocatiepatronen en de leestijden leidden ook tot nieuwe inzichten in de relatie tussen subjectiviteit en perspectief. Perspectief kan worden gebruikt als instrument om subjectiviteit in discourse te operationaliseren. De aan- dan wel afwezigheid van een SoC bepaalt de mate van subjectiviteit. Als een uiting een SoC veronderstelt, bijvoorbeeld doordat iemand daarin een evaluatie geeft op basis van een aantal argumenten, dan is de relatie subjectiever dan in uitingen waarin slechts een gebeurtenis wordt beschreven. Dat een uiting een SoC veronderstelt, kan talig worden gemarkeerd met bijvoorbeeld perspectiefmarkeerders, modale werkwoorden en subjectieve connectieven.

Perspectief specificceert vervolgens langs welke dimensie van subjectiviteit de SoC in de uiting betrokken is. In de traditionele definitie van de notie subjectiviteit omvat subjectiviteit zowel subjectieve overtuigingen en standpunten (Finegan, 1995; Lyons, 1987) als subjectieve houdingen (Lyons, 1987; Traugott, 1995). Deze aspecten van subjectiviteit kunnen worden uitgedrukt met talige middelen zoals *Jan*

denkt (standpunt), *belangrijk* (attitude), *gelukkig* (emotie) en *waarschijnlijk* (zekerheid). Een dergelijke definitie behandelt subjectiviteit als een homogeen concept. Ik stel echter voor om het concept subjectiviteit te analyseren langs verschillende dimensies die tot nu vooral in literatuur over perspectief werden onderscheiden: *epistemic stance*, *attitudinal stance* en *style stance* (Bednarek, 2006; Conrad & Biber, 2000). *Epistemic stance* markeert evidentialiteit en de (on)zekerheid van de spreker; *attitudinal stance* betreft de houding en emoties van de spreker; en *style stance* geeft weer hoe iets wordt uitgedrukt. Het leesexperiment toonde aan hoe ten minste twee van deze dimensies van subjectiviteit invloed hebben op verwerking, en de collocatiestudie liet verschillende patronen in taalgebruik zien die middels deze dimensies kunnen worden verklaard. Wanneer in (7a) bijvoorbeeld een *attitudinal stance* werd uitgedrukt, in de combinatie van een *attitudinal stance marker* en *kejian* – een voorkomend collocatiepatroon in taalgebruik –, werd deze zin sneller verwerkt dan (7b).

(7)

- a. Deze brand heeft **gelukkig** geen verwondingen of sterfgevallen veroorzaakt, dus de brandweer moet de redding op tijd in gang hebben gezet.
- b. Deze brand heeft **beslist** geen verwondingen of sterfgevallen veroorzaakt, dus de brandweer moet de redding op tijd in gang hebben gezet.

Style stance markers, die aangeven *hoe* iets wordt gesproken of geschreven en die eveneens subjectiviteit uitdrukken, manifesteerden zich niet in de collocatieresultaten. Ook zijn hun effecten op de verwerking nog niet getest, wat mogelijkheden biedt voor vervolgonderzoek.

3.2 Pragmatische principes in het gebruik van connectieven en perspectiefmarkeerders

Het gebruik van connectieven en perspectiefmarkeerders dat in de collocatiestudie naar voren kwam, blijkt in lijn met Horns principes van *Relatie* en van *Kwaliteit*. Als de epistemische dimensie al werd gemarkeerd door het subjectieve *kejian*, kwamen *epistemic stance markers* rondom het connectief minder frequent voor; ze waren niet efficiënt volgens het principe van *Relatie* (het verminderen van de productie-inspanning van de spreker). Met het ondergespecificeerde *suoyi* wordt de epistemische dimensie van subjectiviteit nog niet gespecificeerd. In dit geval zorgden de *epistemic stance markers* voor de benodigde informatie voor de interpretatie van subjectiviteit door lezers/luisteraars; dit is in overeenstemming met het principe van *Kwaliteit* (het verminderen van de begripsinspanning van de luisteraar).

Net als *epistemic stance markers* en het subjectieve *kejian* duiden *attitudinal stance markers* (e.g. *jingran* ‘verrassend’ en *zhongyao* ‘belangrijk’) op de aanwezigheid van een SoC. Deze *attitudinal stance markers* drukken subjectiviteit echter in een andere dimensie uit: ze geven de houding of gevoelens van de SoC over de informatie weer en overlappen dus niet met *kejian* qua subjectiviteitsdimensie. Dit verklaart waarom de combinatie van *attitudinal stance markers* en *kejian* niet ingaat tegen Horns principe van *Relatie*.

De collocatieresultaten kunnen ook goed worden verklaard vanuit de *Uniform Information Density Theory*. Deze theorie claimt dat de informatiedichtheid van een uitspraak uniform moet worden gehouden, zonder te veel overlap in betekenis te veroorzaken (Frank & Jaeger, 2008; Jaeger, 2006; Levy & Jaeger, 2007). Als gebruik wordt gemaakt van *epistemic stance markers* (zoals cognitie-werkwoorden, modale werkwoorden), is de inhoud op de epistemische dimensie van subjectiviteit zeer voorspelbaar (hoge waarschijnlijkheid en lage informatie). Er is dan een grotere kans op een ondergespecificeerd connectief als *suoyi*. Daarentegen maken uitingen met weinig *epistemic stance markers* de inhoud van de volgende zin minder voorspelbaar (lage waarschijnlijkheid en hoge informatie). In deze gevallen heeft het gebruik van het specifieke connectief *kejian* de voorkeur. De gangbaarheid van *epistemic stance markers* in de context van *suoyi* en hun minder frequente aanwezigheid bij *kejian* zorgt ervoor dat de informatiedichtheid in de zin op het vlak van subjectiviteit relatief uniform blijft.

3.3 Incrementele verwerking van subjectiviteit

In de literatuur zijn contrasterende hypothesen over de verwerking van coherentierelaties te vinden. Volgens de vertraagde-integratiehypothese wordt de interpretatie van relaties uitgesteld tot het einde van het tweede segment, waar de representaties van de twee segmenten worden geïntegreerd (Kintsch & Van Dijk, 1978; Millis & Just, 1994). Recente inzichten in taalverwerking ondersteunen echter een incrementele-verwerkingshypothese waarin wordt gesteld dat lezers een relatie al direct tijdens het lezen van de segmenten interpreteren (Canestrelli et al., 2013; Cozijn, Noordman, & Vonk, 2011; Koornneef & Sanders, 2013; Traxler, Bybee & Pickering, 1997; Van Silfhout, Evers-Vermeul, Mak, & Sanders, 2014). De vraag is hoe subjectiviteitsmarkeerders die eerder in de zin voorkomen de verwerkingseffecten van subjectiviteitsmarkeerders later in de zin beïnvloeden.

Het leesexperiment bevestigde net als eerdere studies de incrementele-verwerkingshypothese van zinsverwerking: de verwerking van subjectiviteit werd niet tot het einde van de zin uitgesteld, maar kwam juist incrementeel tot stand, onmiddellijk tijdens het lezen van de zin. De verwerking werd gestuurd door talige middelen zoals connectieven en modale werkwoorden. Elke aanwijzing van subjectiviteit in de zin droeg direct bij aan de mentale representatie, en daarmee aan de verwerking van nieuw inkomend materiaal. Werde de subjectieve relatie

bijvoorbeeld uitgedrukt door *kejian*, zoals in (6), dan volgde een onmiddellijke verwerkingsvertraging in vergelijking met de *suoyi*-conditie. Deze vertraging kan worden toegeschreven aan de complexiteit van het interpreteren van informatie over de subjectiviteit van de relatie: zodra lezers wisten dat de relatie subjectief was, begonnen ze met het construeren van een mentale representatie die een SoC bevatte.

Ook het VWP-oogbewegingsexperiment leverde evidentie voor een incrementele visie op de verwerking van subjectiviteit in causale relaties. In zowel het Nederlands als het Chinees gaven subjectieve connectieven de aanzet tot aandacht voor de spreker – dat was de relevante SOC in het experiment. In het Chinees vertoonden modale werkwoorden een soortgelijke functie door de aandacht van de lezer te leiden naar de visuele weergave van de spreker, direct wanneer zij een eerste aanwijzing waren voor subjectiviteit in de zin. Met andere woorden, zodra lezers/hoorders indicaties kregen over de subjectiviteit van de relatie, hetzij door het connectief of het modale werkwoord, begonnen zij in hun mentale representaties met het construeren van een SoC.

Daarnaast lieten het leesexperiment en de VWP-studie zien dat de verwerking van subjectiviteitsmarkeerders later in de zin wordt beïnvloed door subjectiviteitsmarkeerders eerder in de zin. Als er bijvoorbeeld een markeerder van subjectiviteit in de voorafgaande context stond (bijvoorbeeld een perspectiefmarkeerder als *jushuo* ‘het is gezegd’ of het subjectieve *kejian*), werd de verwerking van het modale werkwoord in de daarop volgende context beïnvloed, zowel in het leesexperiment als in de VWP-studie. Wanneer perspectiefmarkeerders in de context ontbraken, fungeerden modale werkwoorden (zoals *keneng* ‘mogen’) in het tweede segment als verwerkingsaanwijzingen die onmiddellijk leidden tot interpretatie van de aanwezige subjectiviteit. Het kostte in deze situaties meer tijd om het modale werkwoord te verwerken dan in situaties waarin ook perspectiefmarkeerders voorkwamen.

3.4 Inzichten vanuit onderzoek naar subjectiviteit in het Chinees

Het bestuderen van het lexicon van Chinese causale connectieven heeft ons een nauwkeuriger beeld gegeven van de codering en verwerking van subjectiviteit. Eerdere studies bekeken subjectieve versus objectieve relaties, of subjectieve versus objectieve connectieven. In de huidige studie is gekeken naar het Chinees, een taal die in het lexicon een connectief heeft dat ondergespecificeerd is voor subjectiviteit, en daarnaast ook specifieke subjectieve en specifieke objectieve connectieven kent. Het vergelijken van de verwerkingseffecten van subjectieve en objectieve connectieven met het ondergespecificeerde connectief gaf een duidelijk beeld van de precieze rol van connectieven als verwerkingsinstructies: het subjectieve connectief *kejian* leidde tot verhoogde aandacht voor de SoC in vergelijking met het ondergespecificeerde connectief en veroorzaakte daarmee een onmiddellijke

verwerkingsvertraging; het objectieve connectief *yin'er* ‘daardoor’ leidde de aandacht juist weg van de SoC.

Ook de combinatie van de gehanteerde onderzoeksmethoden bood meerwaarde. Door een directe vergelijking van de context van een specifiek subjectief connectief en dat van een ondergespecificeerd connectief, ontstond in de collocatiestudie een beeld van de favoriete manieren waarop taalgebruikers connectieven en verschillende soorten perspectiefmarkeerders combineren. Deze typen perspectiefmarkeerders werden in combinatie met connectieven onderzocht in een leesexperiment dat opnieuw de incrementele aard van de verwerking van subjectiviteit aantoonde. Tot slot testte de VWP-studie de theorie over de manier waarop mensen subjectiviteit in discourse begrijpen: de verwerking van subjectiviteit omvat het achterhalen van een SoC en het construeren van een mentale representatie van de coherentierelatie waarin deze SoC een plaats krijgt.

In vervolgonderzoek kan de interactie tussen het gebruik van connectieven en perspectiefmarkeerders nader bestudeerd worden. De collocatiestudie liet bijvoorbeeld zien dat zinnen met een connectief zonder informatie over subjectiviteit in de epistemische dimensie vaker perspectiefmarkeerders bevatten die deze informatie wél uitdrukten. Het is interessant om te achterhalen hoe dit werkt bij andere typen informatie die connectieven kunnen markeren, zoals volitionaliteit of temporaliteit.

Vanwege het exploratieve karakter van de collocatie-analyse is er in dit corpusonderzoek geen onderscheid gemaakt tussen *suoyi*-zinnen die een subjectieve relatie uitdrukten en *suoyi*-zinnen die een objectieve relatie uitdrukten. Het zou hierdoor kunnen dat sommige contextuele kenmerken van *suoyi* (in vergelijking met *kejian*) het gevolg zijn van de typerende kenmerken van de objectieve *suoyi*-relaties. Om te achterhalen of dit inderdaad het geval is, is een vervolgstudie nodig waarin de relaties gemarkeerd met *suoyi* zijn geannoteerd voor subjectiviteit.

In het Chinese leesexperiment zijn in de categorie objectieve relaties alleen non-volitionele causale zinnen getest. Het is goed denkbaar dat de verwerking van dergelijke relaties, die geen intenties uitdrukken, verschilt van die van volitionele causale relaties, waarin personen doelbewust bepaalde handelingen verrichten. Toekomstig experimenteel onderzoek kan een beter beeld geven van eventuele verwerkingsverschillen tussen deze twee typen objectieve relaties.

4. Conclusie

In dit proefschrift is vastgesteld dat Chinese connectieven en perspectiefmarkeerders een rol hebben als linguïstische indicatoren van subjectiviteit. De collocatiestudie illustreerde voorkeuren die taalgebruikers hebben om deze talige middelen te combineren. In de verwerkingsstudies (het leesexperiment en de VWP-studie), is de rol van deze middelen als verwerkingsinstructies aangetoond. Juist de combinatie

194 | Samenvatting

van de drie onderzoeksmethodes bleek vruchtbaar en heeft nieuwe inzichten opgeleverd in het verschijnsel subjectiviteit.

Acknowledgements

I have been fantasizing about writing this part of my thesis since my first year of my PhD. Finally, here we are. But when I started writing, I found it was never easy – summarizing the experience with the dear people around me in the past few years brought overwhelming feelings itself. Finishing this thesis requires rational analyses and logical arguments, although it is about subjectivity. Only in writing this part, I have a precious moment and a justified reason to let subjective emotions lead my mind.

I would first like to thank my supervisors: Ted Sanders, Jacqueline Evers-Vermeul and Pim Mak. Ted, I cannot use words to describe how lucky I feel to have you as my supervisor, so I will tell by a story. I once told a friend that I did not even dare to ask a question in conferences when I first came here. He asked, then what happened? I answered, then I met Ted! Yes, so many changes in my life, huge and small, started from the moment when Ted became my supervisor. Thank you, Ted! Jacqueline, you have taught me so many things – how to do a literature review, how to formulate a line of reasoning, how to write in an academic style, how to present and communicate ideas... You have made such an effort to help me grow with your enthusiasm for research and caring for students. And in my way of growing into an independent researcher, I feel assured to have you standing by me. Pim, I still remember that we were thrilled as children when we saw the results from the experiment. They were exactly what we drew hypothetically on the blackboard! You said *Great* three times in one row. Curiosity and courage – the attitude you have towards science encourages me through the adventure of making a breakthrough. It ever has and ever will.

I also owe great thanks to Dirk Speelman (University of Leuven), who has explained the complex statistical methods to me and helped me to apply them in practice. Everyone knows what a huge task that is. Dirk has selflessly offered his time, effort and knowledge to help me. From him, I learned that this should be the way to treat others who need help in academia, which I set as a requirement for myself. I thank prof. Hongyin Tao (UCLA), who has been an outside advisor of my PhD project since the very start. In the past four years, he always offered his advice and suggestions with no hesitation. In every email and every discussion, I felt the patience and ardor he has to promote young researchers in linguistics.

Many thanks to all the researchers and professors who have enlightened me in research as well as in personal development by sharing their brilliant ideas and experience: prof. José Sanders, prof. Hugo Quené, dr. Rein Cozijn, prof. Elsi Kaiser, prof. Pia Knoeferle, prof. Arie Verhagen, prof. Wilbert Spooren, prof. Chu-ren Huang, prof. Xianwen Cao, dr. Quanzhen Zhang, dr. Gang Chen, prof. Elizabeth Traugott, prof. Ronald Langacker, prof. Sandrine Zufferey, prof. Henriette de Swart,

prof. Huub van den Bergh, prof. Jos van Berkum, dr Aoju Chen, dr. Hans Van de Velde, dr. Joost Zwarts, Yang Yang, Yan Gu, Jing Lin, Cunliang Geng and Yao Tong. I want to give special thanks to dr. Quan Wan and dr. Yongwei Zhang, who had offered tremendous help in the corpus-based studies of my project.

A special *thank you* goes to the staff at the UiL-OTS lab – Iris Mulders, Maarten Duijndam and Chris van Run, and also Martijn van der Klis from the Digital Humanities lab. Without them, the empirical studies in this dissertation would never have been accomplished. Special thanks to Iris, who trained me in the eye-tracking techniques in the first place and showed me what a fun thing it is to do experiments! I also want to thank Caprice Sola, who helped me to prepare the beautiful stimuli for our experiment, Yan Xia and Manou van Montfort, who assisted me in running the experiments. Many thanks go to all the participants in my experiments!

I want to thank all the members of TASK committee, Maaïke Schoorlemmer, Biene Meijerman, Erna Kas, Rianne Giethoorn, Jose Aelst, Simone Veld, Nike Stam and Sophie Spaan, who have formed a great team that I'm so proud to be part of. Erna, you are my Icon! I also want thank Maaïke as the research coordinator of UiL-OTS, who has amazingly enabled a lot of things to happen through management and creativity. Many thanks to the secretaries of UiL OTS and TLC, Yvonne van Adrichem, Mariëlle Hilken and Zarah Farhan, and the secretary of LOT Martine Paulissen.

I received a lot of support and inspiration from the language and communication research group – Jet Hoek, Andrea Santana Covarrubias, Fang Li, Suzanne Kleijn, Nina Sangers, Merel Scholman, Björn 't Hart, Marijn Struiksma, Anita Eerland, Laura Brands, Theresa Redl, Martin Groen, Hongling Xiao, Inés Recio, Fernando Moncada. Either in Codex or DRP, sparkling ideas always made impressive discussions. Special thanks to Jet and Andrea who will support me as my Paranympths, and to Nina who helped me with the Dutch translation of the summary.

During the past four years, a lovely group of people made Trans a delightful place, a second home for me, where I felt happy and secure. Thank you, my dear colleagues and friends in Trans, for making and sharing the best memories – Zenghui Liu, Anqi Yang, Mengru Han, Ao Chen, Shuangshuang Hu, Heidi Klockmann, Maartje Schulpen, Anja Goldschmidt, Alexia Guerra Rivera, Liquan Liu, Xiaoli Dong, Bert Le Bruyn, Yingtong Liu, Xin Li, Jingwei Zhang, Luying Hou, Shanshan Fan, Mike Olson, Jolien Scholten, Ileana Grama, Quy Ngoc Thi Doãn, Chou Mo, Yuan Xie, Bei Peng, Myrthe Bergstra, Marjolein van Egmond, Dominique Blok, Anne van Leeuwen, Anne-France Pinget, Franca Wesseling and Stavroula Alexandropoulou.

I want to give special thanks to my office mates: Brigitta Keij, who helped me to get acquainted with the new environment at the very start of my PhD, and Suzanne Bogaerds-Hazenberg, who accompanied me in the harsh time of the final

stage with Louise. I also thank my other roomies who have brought so much joy to the office: Eva Poortman, Kirsten Schutter, Rob Zwitterlood, Marko Hladnik, and Captain Loes Janssen.

I owe a tremendous *thank you* to all my dear friends, who have made my experience of living abroad so much more colorful and fun: Li Yang and Yan Jiang – my best companions of traveling, shopping and everything; Tianruo Shao – from you I learned that a cool attitude makes everything simpler; Jeroen Breteler – every talk between us has a philosophical ending; Congrui Qiao – let's be female Junzi together; Lijie Zheng – the person with whom I can talk about anthropology; Tony – we have the best experience exploring food and music; Zhang Miao – the best photographer in Utrecht; Jieyu Chen and Yinghe Huo – thank you for being there for me as always. I also want to thank my friends, who have made Utrecht such a lovely place full of memories: Jing Ying (Qingqing), Zhongzheng Tan, Jinqiang Zhang, Rui Jin, Xiaoqian Li, Yihong Liu, Yueru Ni, Zhan Wu, Daixiao Zheng, Maria Koutiva, Maurício Martins, Lotte Hendriks, Emiel van Miltenburg, Loes Burgers, Merel van Witteloostuijn, Sander Karst, Devin Vartija, Jianye Chen, Yi Ke, Yuxing Huang, Haoran Yang, Jingyang Liu, Zidan Mao, Wenjing Lu, Yulong Zhao, Xiaolin Zang, Yumao Zhang, Mei Liu, Yujie He, Yue Song, Yingying Tan, Yili Mo, Jianan Liu, Yuchen Li and Jiangyi Huang.

Fengyuan, thank you for your understanding and support, which make life such a wonderful trip to enjoy and to expect. Thank you for being in my life, Ouyang and Fei, my sister-like friends, and Xinyao, my friend-like sister. Most importantly, I want to thank my parents for all the love and encouragement they have given me in the past 29 years.

雏鸟振飞，不忘哺育之恩；寸草微心，难报三春之晖。从蹒跚学步、牙牙学语，到游学四方、远赴重洋，我走出的每一步，都离不开父亲母亲给我的爱与支持。愿父母大人福寿长安！

Curriculum vitae

Yipu Wei was born on the 13th of October, 1988 in Henan, China. In 2011, she obtained a BA from Nanjing University (China), majoring in teaching Chinese as a second language. She then started a research master program in linguistics at Utrecht University, enabled through Utrecht Excellence Scholarship. She obtained a research MA (cum laude) specializing in language use and psycholinguistics in 2013. She started her PhD research at the Utrecht Institute of Linguistics-OTS, Utrecht University in 2013. This dissertation is the result of the research she carried out during that period.