

## Chapter Seven

# Semantic Classes and Non-overt Subjects of CACs in the Lancaster Corpus of Mandarin Chinese

### 7.1 Introduction

The Lancaster Corpus of Mandarin Chinese (LCMC) is used in this book in large part to offset four major drawbacks of the PFR Chinese Corpus. Firstly, the PFR corpus is made up of newswire texts only. Journalistic writing is only one form of written language (Biber, 1988: 69) and thus a body of journalistic texts can hardly be sufficiently representative of written language as a whole. The LCMC corpus is, however, a balanced corpus of Mandarin Chinese, modelling its sampling frame on that of the Freiburg-LOB Corpus of British English or FLOB (Hundt et al., 1998): it has 500 samples of 2,000 words each, taken from fifteen different text types, as shown in Table 26 (McEnery and Xiao, 2004b). Secondly, while the tagset of the LCMC corpus is basically identical to that of the PFR corpus (see Chapter Two, section 2.2.2.1), it uses five additional tags i.e. conjunction morpheme (*cg*), sentential punctuation (*ew*), locality morpheme (*fg*), preposition morpheme (*pg*) and descriptive morpheme (*zg*). In particular, the distinction between non-sentential punctuation (*w*) and sentential punctuation (*ew*) is vital as it eases the identification of individual sentences in which an adverbial clause occurs so that the reference of the non-overt subject of the adverbial clause can easily be tracked down. Furthermore, unlike the written texts in most of the extant Chinese corpora, including the PFR corpus, which are encoded in GB2312 using simplified Chinese characters, the texts of the LCMC corpus were encoded in Unicode (UTF-8) which can be used in non-Chinese operating systems and can be searched in Unicode-compliant concordancers such as the LCMC web concordancer (see section 7.1.1.1), *Xaira* version 1.0 (Burnard and

Todd, 2003)<sup>1</sup> and the *Wordsmith* Tools version 4.0 (Scott, 2003). As McEnergy and Xiao (2004a) observe, a real problem in Asian corpus building is the existence of multiple and often competing encodings of Asian writing systems: the Chinese language can be encoded in GB2312, GB18030, HZ and Unicode. Xiao et al. (2004), however, recommend Unicode as the best encoding format, and support this by using two Unicode-compliant corpus tools that are available, *Xaira* and *Wordsmith* version 4, to explore the LCMC corpus. Given that the Unicode-encoded LCMC corpus can be explored efficiently with generally available corpus tools such as *Xaira* and *Wordsmith* 4 and its own web concordancer, it is of greater use than the PFR corpus in linguistic research, in particular in the investigation of complicated grammatical phenomena. I therefore decided to carry out an investigation of adverbial clauses in the LCMC to supplement my findings based on the PFR corpus. A final advantage of the LCMC corpus is that it is annotated with five levels of detail<sup>2</sup> in addition to POS tagging, notably including the annotation of sentence and paragraph boundary markers, without which I would have had to add these two levels of markup as I had to with the PFR corpus. In this chapter, I will apply the findings presented in the preceding chapters to the LCMC corpus in the hope of providing a more concrete description of adverbial clauses in written Chinese. This chapter is organised as follows: section 7.2 discusses the occurrence of the eleven semantic classes of Chinese adverbial clauses (CACs) in different text types; section 7.3 examines the distribution of non-overt subjects (PROs) across and within text types, and across the semantic domains of CACs. These results are compared to the results obtained on the basis of the PFR corpus to investigate whether the effect of adverbial semantic domain on the distribution of PROs as indicated in the previous chapter is indeed dependent on text type; section 7.4 discusses how the type of control of PRO is influenced by text type; section 7.5 summaries the findings presented in this chapter.

1 *Xaira* stands for “XML Aware Indexing and Retrieval Architecture”.

2 The five levels of annotation in the LCMC corpus are, namely, (1) text category, (2) file identifier, (3) paragraph, (4) sentence and (5) word, punctuation/symbol and elements omitted in transcriptions.

<i>Text Type</i>	<i>Description</i>	<i>Tokens</i>
Category A	Press: reportage	88,000
Category B	Press: editorials	54,000
Category C	Press: reviews	34,000
Category D	Religion	34,000
Category E	Skills, trades and hobbies	76,000
Category F	Popular lore	88,000
Category G	Biographies and essays	154,000
Category H	Miscellaneous: reports and official documents	60,000
Category J	Science: academic prose	160,000
Category K	General fiction	58,000
Category L	Mystery and detective fiction	48,000
Category M	Science fiction	12,000
Category N	Adventure and martial arts fiction	58,000
Category P	Romantic fiction	58,000
Category R	Humour	18,000

Table 26: The list of text types in the LCMC corpus.

### *7.1.1 Methodological issues*

In the previous chapters of this book, I used the PFR Chinese Corpus as a training corpus to gain insights into the use of adverbial clauses in written Chinese. I have explored various features of them, namely the use of a subordinating conjunction to mark the adverbial clause overtly (Chapter Three), the identification of 57 adverbial subordinators (Chapter Four), the classification of 2,417 adverbial clauses identified in the PFR corpus into eleven distinct interclausal semantic domains (Chapter Five) and the distribution of non-overt subjects within these semantic domains (Chapter Six). As noted above, all of the findings so far are based on the PFR corpus, a homogeneous collection of journalistic texts; these results cannot be generalised to written Chinese easily if at all (cf. Biber and Finegan, 1991: 211–213). As the distribution of both the adverbial clauses and their non-overt subjects may vary across text types, the LCMC corpus was used in my research as the test corpus in an attempt to offer a comprehensive account of adverbial clauses in written Chinese. The main procedures of exploiting the LCMC corpus for the purposes of this research are presented as follows.

### 7.1.1.1 The LCMC web concordancer

I used the LCMC web concordancer<sup>3</sup> for my search of adverbial clauses in the LCMC corpus. The discussion in Chapter Three (section 3.7) showed that Chinese adverbial clauses (CACs) are typically marked by adverbial subordinators and the discussion in Chapter Five (section 5.3) indicated that the semantic classes of adverbial clauses are commonly introduced by different adverbial subordinators, as highlighted in Table 27.

Semantic classes of CACs	Adverbial subordinators
Clauses of Time	随着 <i>suizhe</i>
Clauses of Cause/Reason	由于 <i>youyu</i> , 因为 <i>yinwei</i> , 因 <i>yin</i> , 既然 <i>jiran</i>
Clauses of Purpose	以 <i>yi</i> , 从而 <i>conger</i> , 以免 <i>yimian</i> , 以便 <i>yibian</i>
Clauses of Result	从而 <i>conger</i> , 之所以 <i>zhisuoyi</i> , 以致 <i>yizhi</i> , 故 <i>gu</i> , 以至于 <i>yizhiyu</i>
Clauses of Preference	与其 <i>yuqi</i> , 与其说 <i>yuqishuo</i>
Clauses of Contrast	而是 <i>ershi</i>
Clauses of Addition	何况 <i>hekuang</i> , 不单 <i>budan</i> , 且不说 <i>qiebushuo</i> , 不说 <i>bushuo</i>
Clauses of Exception	只是 <i>zhishi</i>
Clauses of Condition	如果 <i>ruguo</i> , 只要 <i>zhiyao</i> , 只有 <i>zhiyou</i> , 要是 <i>yaoshi</i> , 若是 <i>ruoshi</i> , 倘 <i>tang</i> , 倘若 <i>tangruo</i> , 哪怕 <i>napa</i> , 就是 <i>jiushi</i> , 要不是 <i>yaobushi</i> , 若果 <i>ruoguo</i> , 假若 <i>jiaruo</i> , 万一 <i>wanyi</i> , 如若 <i>ruruo</i> , 果真 <i>guozhen</i> , 要 <i>yao</i> , 除非 <i>chufei</i> , 即使 <i>jishi</i> , 如 <i>ru</i> , 若 <i>ruo</i> , 即便 <i>jibian</i> , 假如 <i>jiaru</i> , 纵 <i>zong</i> , 纵使 <i>zongshi</i>
Clauses of Concession	虽然 <i>suiran</i> , 尽管 <i>jinguan</i> , 虽 <i>sui</i> , 虽说 <i>suishuo</i> , 无论是 <i>wulunshi</i> , 无论 <i>wulun</i> , 不论是 <i>bulunshi</i> , 不论 <i>bulun</i> , 不管 <i>buguan</i> , 任 <i>ren</i> , 甬管 <i>benguan</i>
Clauses of Inference	尚且 <i>shangqie</i>

Table 27: Semantic classes of CACs and their respective adverbial subordinators.

3 See its website <<http://corpus.nie.edu.sg/cgi-bin/lcmc/conc.pl>> (accessed 15 May 2012).

I used the LCMC web concordancer for searching those adverbial clauses overtly marked by an adverbial subordinator. As illustrated in Figure 8, the adverbial subordinator *Yg ruguo* “if” was used as the search term across the fifteen text types of the LCMC corpus. This procedure was repeated until all of the adverbial subordinators identified in the PFR training corpus were searched for in the test corpus. As a result, I obtained a list of subordinator-marked (adverbial) clauses.

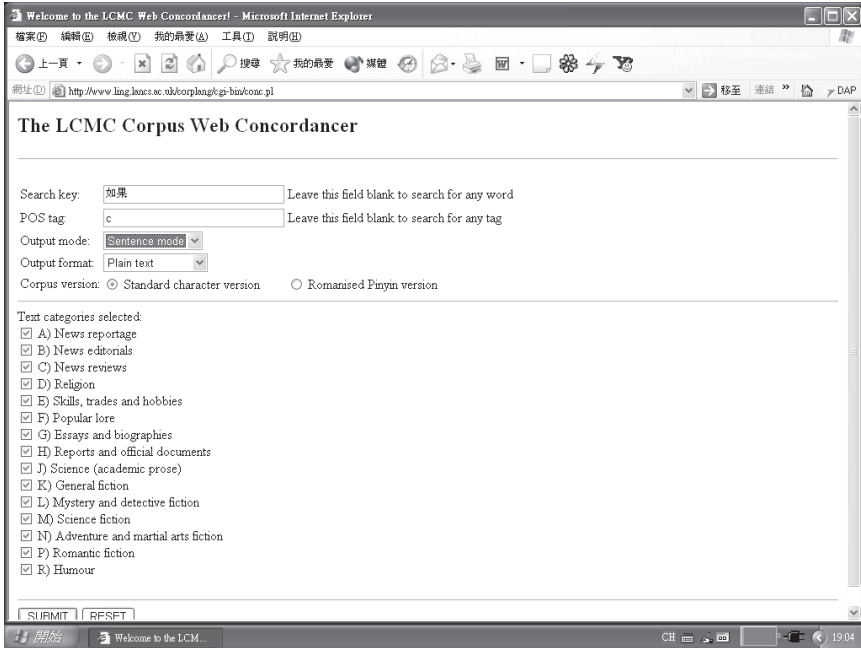


Figure 8: An illustration of the LCMC web concordancer.

### 7.1.1.2 Invalid adverbial clauses

I eliminated incorrect adverbial clauses from the list of subordinator-marked clauses obtained. There are three types of error causing these incorrect adverbial clauses to be drawn from the LCMC corpus by the concordancer. The first kind of error was already noted in Chapter Five (section 5.2) that some words of multiple word class membership may be misinterpreted as adverbial subordinators (e.g. 由于 *youyu* “because” can be a preposition and a conjunction) and the unwanted sentences

retrieved by those words should be disregarded. On the other hand, the sentences in which genuine adverbial subordinators were annotated with the wrong tag were added to the data set manually. The second kind of error arises from the LCMC web concordancer itself: it cannot distinguish adverbial subordinators having identical final characters e.g. 要 *yao* “if” and 只要 *zhiyao* “only if”, and 以 *yi* “in order to” and 之所以 *zhisuoyi* “the reason why...”. The adverbial clauses introduced by these subordinators were therefore sorted by hand. The third kind of error is a result of four borderline adverbial subordinators. In both the PFR and LCMC corpora, four of the adverbial subordinators sporadically violate the first criterion of the operational definition of a subordinating conjunction, which states that the clause overtly marked by an adverbial subordinator must be linked to another clause(s) in the same sentence, as Chinese adverbial clauses function only as an adjunct. These are: 因为 *yinwei* “because”, 从而 *conger* “in order to or as a result”, 何况 *hekuang* “let alone” and 故 *gu* “therefore”. These four subordinating conjunctions are at times used to introduce an independent sentence rather than a subordinate clause. However, I still considered them as subordinating conjunctions because in over sixty percent of their occurrences, these subordinators conform to the first operational criterion, as highlighted in Tables 28 and 29.

Borderline subordinators	因为 <i>yinwei</i>		从而 <i>conger</i>		何况 <i>hekuang</i>		故 <i>gu</i>	
	Freq.	%	Freq.	%	Freq.	%	Freq.	%
In independent sentence	63	35.2	5	2.8	4	36.4	2	22.2
In adverbial clause	116	<b>64.8</b>	173	<b>97.2</b>	7	<b>63.6</b>	7	<b>77.8</b>
Total	179	100	178	100	11	100	9	100

Table 28: Frequencies of borderline adverbial subordinators in independent sentences and adverbial clauses in the PFR corpus.

Borderline subordinators	因为 <i>yinwei</i>		从而 <i>conger</i>		何况 <i>hekuang</i>		故 <i>gu</i>	
	Freq.	%	Freq.	%	Freq.	%	Freq.	%
In independent sentence	68	32.5	2	2.8	7	38.9	6	31.6
In adverbial clause	141	<b>67.5</b>	69	<b>97.2</b>	11	<b>61.1</b>	13	<b>68.4</b>
Total	209	100	71	100	18	100	19	100

Table 29: Frequencies of borderline adverbial subordinators in independent sentences and adverbial clauses in the LCMC corpus.

### 7.1.1.3 Raw and normalised frequencies

As can be seen from Table 26, the fifteen text categories in the LCMC corpus are of different sizes. It is therefore impossible to compare the distribution of adverbial clauses within different text types on the basis of their raw frequencies. Hence, the frequency scores obtained need to be *normalised/standardised* to a common base in order to permit comparison across text types (McEnery et al., 2005: 47–48). A *normalised frequency* is a weighted frequency measure that allows for the easy and reliable comparison of data sets of different sizes (cf. Ball, 2002: 11). I initially chose to normalise frequencies to a common base of 10,000 word tokens because “the common base for normalisation must be comparable to the sizes of the corpora (or corpus segments) under consideration [...] as the results obtained on an irrationally enlarged or reduced common base are distorted” (McEnery et al., *ibid*). As the sizes of the fifteen text types included in the LCMC corpus range from 12K tokens to 160K tokens i.e. the average size of a text type is of approximately 60K tokens, it would be appropriate to normalise frequencies to a common base of 10K tokens. However, as shown in my PFR skeleton treebank (see Chapter Three, section 3.7), adverbial clauses are rare in written Chinese. With such a low frequency of occurrence, statistical tests of significance may not yield reliable results. Consequently, I decided to normalise frequencies to a common 100K-token base so as to avoid the potentially misleading outcomes resulting from significance tests relying on expected values less than 5. I assume that varying the common base for normalisation in this way will not distort the outcomes of the statistical tests significantly.

Take clauses of reason as an example. As shown in Table 30, the frequency of the reason clauses occurring in a text category was obtained by summing the frequencies of the four adverbial subordinators (i.e. *youyu*, *yinwei*, *yin* and *jiran*) responsible for introducing the clauses of reason in the text category. The frequency scores of individual text categories were then normalised to a common 100,000-word base so as to allow for the comparison of the distribution of reason clauses across text types.

Text types of LCMC	Raw frequency					Freq. per 100K tokens
	由于 <i>youyu</i>	因为 <i>yinwei</i>	因 <i>yin</i>	既然 <i>jiran</i>	Total	Total
Category A	11	3	1	2	17	19.3
Category B	12	8	0	1	21	38.9
Category C	9	1	0	0	10	29.4
Category D	9	6	0	11	26	76.5
Category E	23	13	4	4	44	57.9
Category F	31	22	0	3	56	63.6
Category G	20	20	0	8	48	31.2
Category H	7	0	1	0	8	13.3
Category J	96	39	3	8	146	91.3
Category K	0	6	0	5	11	19.0
Category L	10	5	0	6	21	43.8
Category M	2	0	0	5	7	58.3
Category N	2	5	2	10	19	32.8
Category P	5	8	0	5	18	31.0
Category R	1	5	0	4	10	55.6
Total:	238	141	11	72	462	46.2

Table 30: Raw and normalised frequencies of clauses of reason across text types.

#### 7.1.1.4 Overt and non-overt subjects across semantic domains and text types

While the frequencies of adverbial clauses across text types can be found by performing a simple addition as illustrated above, the occurrences of overt and non-overt subjects were found via manual annotation. In this annotation process, each of the adverbial clauses identified in the LCMC corpus was carefully screened to track down any occurrence of a non-overt subject. The raw frequencies of occurrence of overt and non-overt subjects were obtained for different semantic domains and text types. The raw frequencies of both subject types across semantic domains were not normalised as the distribution of subjects across semantic domains is not based on text types but the LCMC corpus as a whole, as shown in Table 31, whereas the raw frequencies of overt and non-overt subjects across text types were normalised to a common base of 100K word tokens, as shown in Table 32.



Semantic classes of CACs	Raw frequency		
	Overt subjects	PROs	Total
Clause of Time	2	0	2
Clause of Cause or Reason	381	81	462
Clause of Purpose	18	90	108
Clause of Result	65	81	146
Clause of Preference	2	11	13
Clause of Contrast	17	193	210
Clause of Addition	11	6	17
Clause of Exception	39	4	43
Clause of Condition	589	612	1,201
Clause of Concession	509	267	776
Clause of Inference	3	0	3
Total:	1,636	1,345	2,981

Table 31: Raw frequencies of overt and non-overt subjects across semantic domains.

Text types of LCMC	Raw frequency			Freq. per 100K tokens		
	Overt subjects	PROs	Total	Overt subjects	PROs	Total
Category A	71	45	116	80.7	51.1	131.8
Category B	103	107	210	190.7	198.1	388.9
Category C	40	47	87	117.6	138.2	255.9
Category D	60	78	138	176.5	229.4	405.9
Category E	183	170	353	240.8	223.7	464.5
Category F	201	170	371	228.4	193.2	421.6
Category G	194	142	336	126.0	92.2	218.2
Category H	14	22	36	23.3	36.7	60.0
Category J	369	313	682	230.6	195.6	426.3
Category K	72	58	130	124.1	100.0	224.1
Category L	64	55	119	133.3	114.6	247.9
Category M	18	17	35	150.0	141.7	291.7
Category N	132	58	190	227.6	100.0	327.6
Category P	88	52	140	151.7	89.7	241.4
Category R	27	11	38	150.0	61.1	211.1
Total:	1,636	1,345	2,981	163.6	134.5	298.1

Table 32: Normalised frequencies of overt and non-overt subjects across text types.

## 7.2 Semantic domains of CACs and text types of LCMC

In this section, I will explore how the semantic domains of adverbial clauses interact with the text types of the LCMC corpus. As can be seen from Table 33, clauses of time (i.e. 2 occurrences) and clauses of inference (i.e. 3 occurrences) are rare, making it impossible to study the distribution of these two types of adverbial clause across the fifteen text categories of the LCMC corpus. This corroborates my initial observation based on the PFR corpus that temporal adverbial clauses rarely occur in written Chinese (see Chapter 5, section 5.3.1). Other semantic classes of adverbial clauses are, however, closely related to certain text types of written Chinese. Hence, in this section, I will examine the interaction between (i) conditional and concessive clauses and text categories A, B and C (section 7.2.1); (ii) reason and result clauses and text categories D and J (section 7.2.2); (iii) purpose clauses and text categories E and F (section 7.2.3). Each contrast will allow me to gain a better understanding of how semantic classes of adverbial clauses distribute across a range of genres in written Chinese.

### *7.2.1 The interaction between conditional and concessive clauses and categories A, B and C*

As noted in Chapter Five (section 5.4), conditional and concessive clauses, among the eleven semantic categories of adverbial clauses in Chinese, make up almost half of the adverbial clauses studied in the PFR corpus. In the LCMC corpus, these two types of adverbial clause occur significantly more frequently than other semantic types (see Table 33).<sup>4</sup> By examining the distribution of conditional and concessive clauses in the press-related texts of the LCMC corpus, it was found that these two types of adverbial clause also dominate in press reportage, press editorials and press reviews (i.e. categories A, B and C) of the corpus. Table 34 shows the frequencies of adverbial clauses in both the PFR corpus and an *ad hoc* subcorpus composed of categories A, B and

4 The calculated log-likelihood (LL) value with 10 degrees of freedom (d.f.) is 5644.170, considerably greater than 29.59, the critical value for significance at  $p < 0.001$ .

C of the LCMC corpus. It can be seen from Table 34 that conditional and concessive clauses occur at least twice as frequently as the other kinds of adverbial clause in the press-related texts of categories A, B and C.

Semantic classes of CACs	Raw frequency
Clauses of Time	2
Clauses of Cause or Reason	462
Clauses of Purpose	108
Clauses of Result	146
Clauses of Preference	13
Clauses of Contrast	210
Clauses of Addition	17
Clauses of Exception	43
<b>Clauses of Condition</b>	1,201
<b>Clauses of Concession</b>	776
Clauses of Inference	3
Total:	2,981
Log-likelihood (10 d.f.):	5644.170 (>29.59)

Table 33: Frequencies of adverbial clauses in the LCMC corpus.

Semantic Types of CACs	PFR		LCMC (categories A, B and C)	
	Raw freq.	Freq. per 100K tokens	Raw freq.	Freq. per 100K tokens
Clauses of time	1	0.1	0	0.0
Clauses of reason	330	29.4	48	27.3
Clauses of purpose	348	31.1	48	27.3
Clauses of result	163	14.5	28	15.9
Clauses of preference	3	0.3	3	1.7
Clauses of contrast	177	15.8	36	20.5
Clauses of addition	13	1.2	2	1.1
Clauses of exception	25	2.2	2	1.1
<b>Clauses of condition</b>	779	<b>69.5</b>	166	<b>94.3</b>
<b>Clauses of concession</b>	575	<b>51.3</b>	115	<b>65.3</b>
Clauses of inference	3	0.3	0	0.0
Total:	2,417	215.7	448	254.5

Table 34: Frequencies of adverbial clauses in the PFR corpus and categories A, B, and C of the LCMC corpus.

I therefore hypothesise that journalistic writing is marked by clauses of condition and concession. To test this hypothesis, I tested the statistical significance of the difference in the occurrence of these two types of adverbial clause between the journalistic texts of categories A, B and C and non-journalistic texts of other categories of the LCMC corpus (i.e. categories D, E, F, G, H, J, K, L, M, N, P, and R).

Adverbial semantic classes	Freq. per 100K tokens	Freq. per 100K tokens	Fisher's Exact Test ( $p < 0.05$ )
	Journalistic texts (Categories A, B and C)	Non-journalistic texts (Categories D, E, F, G, H, J, K, L, M, N, P, R)	
Clauses of condition	94.3	125.6	0.747
Clauses of concession	65.3	80.2	

Table 35: Journalistic texts vs. non-journalistic texts in the LCMC corpus.

As Table 35 illustrates, conditional and concessive clauses are not a marked feature of journalistic writing; they do not occur significantly more frequently in journalistic texts than in non-journalistic texts.<sup>5</sup> This finding is in line with Biber et al. (1999: 823), who observe that in British English it is time and purpose clauses that are frequently used in news rather than conditional and concessive clauses: temporal clauses describe certain events in relation to others while clauses of purpose explain motivations behind events. In written Chinese, as shown in Table 36, conditional clauses are mostly used in category E (skills, trades and hobbies) for giving directions and instructions in hypothetical situations as in example (1), and concessive clauses are predominantly used in category N (adventure and martial arts fiction) for demonstrating a character's ability to handle some difficult events as in example (2). The LL values with 14 d.f. calculated on the basis of normalised frequencies for determining the distribution of clauses of condition and concession across text categories are 395.790 and 192.970 respectively, markedly greater than 36.12, the critical value for significance at  $p < 0.001$ . This suggests that clauses of condition and concession are more frequently used in categories E and N than in other text categories.

5 The Fisher's Exact Test (see Chapter Six, section 6.5.1) gives a high probability value (or  $p$ -value), greater than 0.05, the  $p$ -value at the level of significance of 95 percent.

Text types of LCMC	Clauses of condition		Clauses of concession	
	Raw freq.	Freq. per 100K tokens	Raw freq.	Freq. per 100K tokens
Category A	43	48.9 (2.8%)	38	43.2 (3.9%)
Category B	82	151.9 (8.6%)	54	100.0 (9.0%)
Category C	41	120.6 (6.8%)	23	67.6 (6.1%)
Category D	39	114.7 (6.5%)	40	117.6 (10.6%)
<b>Category E</b>	195	<b>256.6 (14.5%)</b>	62	81.6 (7.3%)
Category F	185	210.2 (11.9%)	72	81.8 (7.4%)
Category G	119	77.3 (4.4%)	115	74.7 (6.7%)
Category H	13	21.7 (1.2%)	3	5.0 (0.5%)
Category J	218	136.3 (7.7%)	159	99.4 (8.9%)
Category K	58	100.0 (5.6%)	38	65.5 (5.9%)
Category L	45	93.8 (5.3%)	32	66.7 (6.0%)
Category M	13	108.3 (6.1%)	6	50.0 (4.5%)
<b>Category N</b>	76	131.0 (7.4%)	75	<b>129.3 (11.6%)</b>
Category P	55	94.8 (5.4%)	52	89.7 (8.1%)
Category R	19	105.6 (6.0%)	7	38.9 (3.5%)
Total:	1,201	120.1 (100%)	776	77.6 (100%)
LL scores (14 d.f.):	395.790		192.970	

Table 36: Clauses of condition vs. clauses of concession in the LCMC corpus.

- (1) 白色衬衣经过多次穿用、洗涤，容易发黄，<Fm><Fa>[[如果]]经常用淘米水浸洗</Fa>，就不易发黄了。

</Fm>(LCMC\_E.xml/sn="0016")

*baise chenyi jingguo duoci chuanyong xidi rongyi*  
white shirt go.through several.times wear wash easy

*fahuang <Fm><Fa>ruguo jingchang yong taomishui*  
turn.yellow if frequently use water.for.washing.rice

*jinxi</Fa> jiu bu yi fahuang le</Fm>*  
soak.and.wash then not easy turn.yellow PART

“The white shirts easily become yellow after being washed several times, but if they are frequently soaked and washed in the water used for washing rice, they will not turn yellow so easily.”

- (2) <Fm><Fa>谭 公直 的 手 [[虽然]] 正在 开始 僵硬</Fa> ， 但 两 人 功 力 相 差 太 远 ， 媳 妇 还 是 扳 不 开 公 公 的 手 。

</Fm>(LCMC\_N.xml/sn="0021")

<Fm><Fa>Tan Gong zhi de shou suiran zhengzai  
 Tan Gong straight DE hand although PROG  
 kaishi jiangying</Fa> dan liangren gongli xiangcha  
 begin stiff but both.of.them marial.arts.skill differ  
 tai yuan xifu haishi banbukai Gonggong de  
 too far.away daughter.in.law still not.get.rid.of Tan's GEN  
 shou</Fm>  
 hand

“Although Tan Gong began to have difficulties in moving his hand, his daughter-in-law, being far less skilful and competent than Tan, still could not get rid of him.”

### 7.2.2 The interaction between reason and result clauses and categories D and J

As can be seen from Table 37, reason and result clauses are most frequently used in category D (religion) and category J (scientific academic prose) of the LCMC corpus. A log-likelihood test of significance was used to determine the distribution of these two kinds of adverbial clause across genres on the basis of their normalised frequencies. The calculated LL score (14 d.f.) for clauses of reason is 162.094, considerably greater than the critical value for significance at  $p < 0.001$  (36.12), whereas the calculated LL score (13 d.f.) for clauses of result is 87.465, much greater than the critical value for significance at  $p < 0.001$  (34.53). Therefore, the statistical tests clearly indicate that reason and result clauses occur significantly more frequently within the two text categories under consideration. Category D is a collection of 17 texts taken from reference books on a range of mainstream religions in mainland China such as Buddhist and Taoist faiths. Most of the texts give details about the origin and development of the religions (e.g. *D03* and *D05*), and a few of them are concerned with the theory and ritual of a particular religion (e.g. *D08*, *D11* and *D13*) as well as the comparisons between different religious beliefs and ideologies (e.g. *D04*, *D09* and *D17*). Thus, clauses of reason and result are commonly used for describing the development of a religion as in examples (3) and (5), and are also employed for

presenting the background/origin of an ideology (i.e. Confucianism) or a religion (i.e. Buddhism) as in examples (4) and (6) respectively.

Text types of LCMC	Clauses of cause or reason		Clauses of result	
	Raw freq.	Freq. per 100K tokens	Raw freq.	Freq. per 100K tokens
Category A	17	19.3 (2.9%)	4	4.5 (2.1%)
<b>Category B</b>	21	38.9 (5.9%)	18	<b>33.3 (15.6%)</b>
Category C	10	29.4 (4.4%)	6	17.6 (8.2%)
<b>Category D</b>	26	<b>76.5 (11.6%)</b>	11	<b>32.4 (15.2%)</b>
Category E	44	57.9 (8.7%)	15	19.7 (9.2%)
Category F	56	63.6 (9.6%)	13	14.8 (6.9%)
Category G	48	31.2 (4.7%)	8	5.2 (2.4%)
Category H	8	13.3 (2.0%)	4	6.7 (3.1%)
<b>Category J</b>	146	<b>91.3 (13.8%)</b>	47	<b>29.4 (13.8%)</b>
Category K	11	19.0 (2.9%)	5	8.6 (4.0%)
Category L	21	43.8 (6.6%)	6	12.5 (5.9%)
Category M	7	58.3 (8.8%)	2	16.7 (7.8%)
Category N	19	32.8 (5.0%)	6	10.3 (4.8%)
Category P	18	31.0 (4.7%)	1	1.7 (0.8%)
Category R	10	55.6 (8.4%)	0	0.0 (0.0%)
Total:	462	46.2 (100%)	4	4.5 (100%)
LL scores:	162.094 (14 d.f.)		87.465 (13 d.f.)	

Table 37: Frequencies of clauses of reason and result across text types.

- (3) 但在另一方面，<Fm><Fa> [[由于]] 教会在历次教案中获得了大量“赔款”和种种特权</Fa>，使它有可能在各地大肆建造教堂。</Fm>(LCMC\_D.xml/sn="0065")

*dan zai lingyifangmian <Fm><Fa>youyu jiaohui zai*  
but at on.the.other.hand because church in  
*lici jiaozhong huode le daliang peikuan*  
previous lawsuits obtain PERF plenty.of compensations  
*he zhongzhong tequan</Fa> shi ta you keneng zai*  
and various.kinds.of privileges thereby it have possibility at  
*gedi dasi jiangai jiaotang</Fm>*  
worldwide on.a.large.scale build churches

“On the other hand, as the Church received a huge sum in compensations and privileges of various kinds from previous lawsuits, it could build a lot of churches worldwide.”

- (4) <Fm>但 儒学 不能 代替 和 包容 传统 宗教 ， <Fa>[[因为]] 儒学 是 一 种 理论 形态 的 学术 文化 ， 宗教 祭祀 并非 儒学 题 中 应有 之 义 ， 儒学 的 重心 在 内 圣 外 王 之 学 ， 自 有 其 发展 脉络 ， 形 成 独 立 的 学 统</Fa> ； </Fm>(LCMC\_D.xml/sn="0045")

<Fm>dan ruxue buneng daiti he baorong chuantong  
but Confucian cannot replace and embrace traditional  
zongjiao <Fa>yinwei ruxue shi yi zhong lilun xingtai  
religions because Confucian be one CL theory mode  
de xueshi wenhua zongjiao jisi bingfei ruxue ti  
DE academic culture religion ritual not Confucian assumption  
zhong yingyou zhi yi ruxue de zhongxin  
within should.have DE meaning Confucian's GEN focus  
zai neishengwaiwang zhi xue zi you qi  
on spiritual DE ideology itself have its  
fazhang mailuo xingcheng duli de  
development route form independent DE  
xuetong</Fa></Fm>  
ideology

“However, Confucianism cannot replace or represent traditional religions because unlike religions it is based on theory, does not employ ritual and has developed into an ideology in its own right.”

- (5) <Fm>在 这 个 痛苦 的 探索 过程 中 ， 一 些 人 以 皈 依 宗 教 作 为 自 己 的 精神 归 宿 ， <Fa>[[从而]] 也 导 致 了 宗 教 的 复 兴</Fa> 。 </Fm>(LCMC\_D.xml/sn="0073")

<Fm>zai zhege tongku de tansuo guochengzhong yixie  
in this painful DE explore in.process.of some  
ren yi guiyi zongjiao zhuowei ziji de jingsheng  
people use follow religion serve.as their GEN spiritual  
guisu <Fa>conger ye dao zhi le zongjiao de  
shelter consequently also cause PERF religion DE  
fuxing</Fa></Fm>  
resurgence

“In the course of this painful adventure, a few of these people sought consolation in religious belief, bringing about the resurgence of religion.”



- (6) <Fm>6 年中经常风餐露宿，” 日食一麻一米，乃至七日食一麻一米”，在艰苦的修行中度过，<Fa>[[以致]]” 身形消瘦，有若枯木”，然而却一无所获</Fa>。

</Fm>(LCMC\_D.xml/sn="0054")

<Fm>liu nian zhong jingchang fengcanlusu ri shi  
 six years within frequently live.in.the.wild day eat  
 yi ma yi mi naizhi qi ri chi yi ma yi mi  
 one sesame one rice and.then seven days eat one sesame one rice  
 zai jianku de xiuxing zhong duguo<Fa> yizhi  
 in harsh DE training in.process.of go.through consequently  
 shengxing xiaoshou you ruo ku mu raner que  
 body.shape skinny have look.like dry wood yet but  
 yiwusuohuo</Fa></Fm>

in.vain

“Over the past six years, Buddha had lived in the wild. In the beginning, he ate a piece of sesame and rice each day, and later on he had sesame and rice once a week. Consequently he became very skinny but gained no insights into life.”

On the other hand, in category J, there are a total of 80 academic texts covering a wide spectrum of science and social science topics such as medicine (e.g. *J01* and *J03*), computing (e.g. *J07*, *J08* and *J68*), engineering (e.g. *J11*, *J67* and *J69*), physics (e.g. *J09* and *J10*), biology, (e.g. *J19* and *J20*) sociology (e.g. *J18* and *J49*), economics (e.g. *J24*, *J27*, *J34* and *J36*), journalism (e.g. *J32*, *J38* and *J53*), management (e.g. *J74* and *J77*), finance (e.g. *J54* and *J73*), psychology (e.g. *J21* and *J75*), education (e.g. *J12* and *J56*), linguistics (e.g. *J23*, *J57* and *J60*), philosophy (e.g. *J48* and *J71*), Chinese literature (e.g. *J04*, *J79* and *J80*), etc. In these texts, causal clauses are frequently used for justifying the methodology of a piece of research as in examples (7) and (8) while result clauses are mainly used for presenting the findings of a piece of research as in examples (9) and (10). Apart from discussing research findings, clauses of result are also frequently used by the official Xinhua news agency and several local newspapers represented in category B (press editorials) of the LCMC corpus. Result clauses are used to present an editor’s comments on the consequences of economic and political policies. As illustrated in examples (11) and (12), both of the examples are concerned with a commentator’s opinions on global political events.

- (7) <Fm><Fa>[[由于]] 要 针对 不同 的 需要</Fa> ， 语言 研究 的 目的 、 角度 和 侧重点 就 会 有所 不同 ， 研究 方法 也 会 有所 不同 。

</Fm>(LCMC\_J.xml/sn="0010")

<Fm><Fa>*youyu yao zhendui butong de xuyao*</Fa> *yuyan*  
 because have.to suit different DE needs language  
*yanjiu de mudi jiaodu he cezhongdian jiu hui*  
 research DE purpose perspective and focus then will.be  
*yousuobutong yanjiu fangfa ye hui yousuobutong*</Fm>  
 different research methodology also will.be different

“In an attempt to suit different purposes, linguistic analyses vary in their objectives, perspectives, focusses and methodologies.”

- (8) <Fm>在 词 处理 阶段 ， 我们 主要 解决 了 不 实行 分词 连写 的 汉语 的 词 处理 手段 问题 ， <Fa>[[因为]] 作为 语言 的 基础 毕竟 是 词 而 不是 ” 字 ” </Fa> 。

<Fm>*zai ci chuli jieduan women zhuyao jiejie le*  
 at word handling stage we mainly solve PERF  
*bu shixing fenci lianxie de hanyu de*  
 not put.into.use word.segmentation write DE Chinese.language DE  
*ci chuli shouduan wenti <Fa>yinwei zhuowei yuyan*  
 word handling strategy problem because serve.as language's  
*de jichu bijing shi ci er bu shi*  
 GEN foundation nonetheless be word but not be  
*zi*</Fa></Fm>

character

“In handling words, we have mainly focussed on the segmentation of words in Chinese as the basic unit of a language is the word not the character.”

- (9) <Fm>由于 它们 的 繁殖力 极强 ， 生长 速度 极快 ， 短期 内 就 会 产生 大量 的 后代 ， 所以 把 目的 基因 转入 这些 细菌 ， 就 能 在 短 时间 内 得到 大量 的 基因 拷贝 ， <Fa>[[从而]] 产生 大量 的 产物 </Fa> 。

<Fm>*youyu tamen de fanzhili ji qiang*  
 because their GEN reproductive.power extremely strong  
*shengzhang sudu ji kui duanqi nei jiu*  
 growth rate extremely quick short.spell during then

*hui chansheng daliang de houdai suoyi ba mudi*  
 can produce plenty.of DE next.generation so.that BA target  
*jiyin zhuangru zhexie xijun jiu neng zai duan*  
 genes inject.into these bacteria then can within short  
*shijian nei dedao daliang de jiyin kaobei*  
 period.of.time during obtain plenty.of DE genes replica  
 <Fa>**conger** *chansheng daliang de chanwu*</Fa></Fm>  
 consequently form plenty.of DE products

“Since they have strong reproductive power and a high growth rate and can reproduce within a short spell, if we inject our target gene into these bacteria, we will obtain, in a short period of time, a huge number of replicas of the target gene and this boosts productivity substantially.”

- (10) <Fm>有个连队收看电视连续剧《诽谤》，片面强调看节目就是受教育，不论工作多忙也要看到底，<Fa>[[以致]]有15次推迟早晨起床时间</Fa>。</Fm>(LCMC\_J.xml/sn="0024")

<Fm>*you ge liandui shoukan dianshi lianxuju*  
 there.exists CL military.officer watch TV drama.series  
*Feibang pianmian qiangdiao kan jiemu jiushi*  
 Rumours on.the.screen emphasise watch TV.programme that.is  
*shou jiaoyu bulun gongzhuo duo mang ye*  
 receive education no.matter job how busy yet  
*yao kandaodi <Fa>yizhi you 15 ci tuichi*  
 want.to watch.till.the.end as.a.result have 15 times delay  
*zaocheng qichuang shijian*</Fa></Fm>  
 morning wake.up time

“One of the military officers watched the TV drama series entitled ‘Rumours’ which purported to be an educational programme and he watched it no matter how busy he was and thus he had got up late fifteen times.”

- (11) <Fm>人们将密切注视，日本今后将以何种形式扩大自己的国际影响，<Fa>[[从而]]达到成为“政治大国”的目的</Fa>。</Fm>(LCMC\_B.xml/sn="0056")

<Fm>*renmen jiang miqie zhushi Riben jinhou*  
 people very.soon closely pay.attention.to Japan in.future

jiang yi hezhong xingshi kuoda ziji de guoji  
 very soon use which way expand its GEN international  
 yingxiang <Fa>conger dadao chengwei zhengzhi daguo  
 influence consequently reach become political giant.nation  
 de mudi</Fa></Fm>

DE purpose

“People will see how Japan expands its influence over other nations and achieves its goal of being one of the largest countries in the world.”

- (12) 路透社认为，<Fm><Fa>她 [[之所以]] 不象过去讨论成立货币联盟时那样强烈反对</Fa>，其原因是在国内政治上遇到了麻烦，威信下降，希望得到其他盟国的支持，再说，目前政治联盟的目标尚不具体，笼统地表示支持也无不可</Fm>。  
 (LCMC\_B.xml/sn="0025")

Lutoushe renwei <Fm><Fa>ta zhisuoyi bu xiang  
 Reuters believe she the.reason.why not like  
 guoqu taolun chengli huobi lianmeng shi  
 in.the.past discuss set.up currency union at.that.time  
 nayang qianglie fandui</Fa> qi yuanyin shi zai guonei  
 that.way fiercely object her reason be at within.the.country  
 zhengzhishang yudao le mafan weixin xiajiang  
 politically-speaking face PERF trouble credibility drop  
 xiwang dedao qita mengguo de zhichi zaishuo  
 hope obtain other united.nations' GEN support what.is.more  
 muqian zhengzhi lianmeng de mubiao shang bu  
 for.the.time.being political coalition's GEN goal yet not  
 juti longtong de biaoshi zhichi ye wu bu  
 flesh.out ostensibly ADVL indicate support yet have.not not  
 ke</Fm>

possible

“Reuters news agency believed the reason why she did not object severely to the creation of currency unions as she did previously was that she had political troubles at home and her credibility dropped, and thus she wished to gain support from other nations; additionally, the objective of the plan was not made clear yet and her ostensible support was by no means inappropriate.”

### 7.2.3 The interaction between purpose clauses and categories E and F

In the LCMC corpus, it was shown in Table 38 that purpose clauses are most frequently used in category E (i.e. skills, trades and hobbies) as in examples (13) and (14) and category F (i.e. popular lore) as in examples (15) and (16). The LL value with 9 d.f. for clauses of purpose calculated on normalised frequencies is 51.296, greater than the critical value for significance at  $p < 0.001$  (27.88), highlighting the fact that clauses of purpose are predominant in these two text categories.

Text types of LCMC	Clauses of purpose	
	Raw frequency	Freq. per 100K tokens
Category A	0	0.0 (0.0%)
Category B	10	18.5 (13.8%)
Category C	3	8.8 (6.6%)
Category D	5	14.7 (10.9%)
<b>Category E</b>	19	<b>25.0 (18.6%)</b>
<b>Category F</b>	22	<b>25.0 (18.6%)</b>
Category G	6	3.9 (2.9%)
Category H	8	13.3 (9.9%)
Category J	32	20.0 (14.9%)
Category K	1	1.7 (1.3%)
Category L	0	0.0 (0.0%)
Category M	0	0.0 (0.0%)
Category N	2	3.4 (2.5%)
Category P	0	0.0 (0.0%)
Category R	0	0.0 (0.0%)
Total:	108	10.8 (100%)
LL score (9 d.f.):	51.296	

Table 38: Frequencies of clauses of purpose across text types.

Category E contains 18 texts on a wide variety of topics which fall into three major categories, namely (a) hobbies e.g. fishing (*E02*), playing bridge (*E16*), playing badminton (*E18*), photography (*E32*), playing

cards (E38), etc., (b) practical skills in various fields such as craft (E37) and home decoration (E27 and E36), etc., and (c) tactics in trades and business such as stocks investment (E13 and E15). On the other hand, category F is composed of texts relating to local culture such as celebrities (F05 and F13), fashion (F37), marriage and family (F07 and F35), overseas study (F08) and love affairs (F16, F23 and F24), and texts relating to everyday life such as cosmetic preparations (F10), diet/health (F39, F40 and F41), school (F29) and driving (F01). Hence, as illustrated in examples (13) and (14) respectively, clauses of purpose are used for giving instructions on mastering a hobby i.e. photography and explaining the practical skills needed in making a bird cage, while in examples (15) and (16), they are used, respectively, for offering guidance on interpersonal skills and giving advice on careers.

(13) <Fm>色彩 选配 得当 ， 能 给 人 以 清 雅 大 方 之 感 ， <Fa>[[从而]]  
得到 艺 术 上 的 享 受</Fa> 。 </Fm>(LCMC\_E.xml/sn="0004")

<Fm>*secai xuanpei dedang neng gei ren yi*  
colour pick.and.match properly can give people as.regards  
*qingya dafang zhi gan <Fa>conger dedao yishushang*  
elegant generous DE impression in.order.that obtain aesthetical  
*de xiangshou</Fa></Fm>*

DE pleasure

“If the colour is used properly, the work will look elegant and aesthetic.”

(14) <Fm>洗 刷 后 的 竹 制 鸟 笼 不 能 任 烈 日 下 暴 晒 ， <Fa>[[以免]] 笼  
架 松 散</Fa> 。 </Fm>(LCMC\_E.xml/sn="0046")

<Fm>*xicha hou de zhuzhi niaolong buneng ren*  
wash after DE bamboo bird.cage cannot let  
*lieri xia baoshai <Fa>yimian longjia*  
strong.sunlight under expose in.order.not.to parts.of.cage  
*songsan</Fa></Fm>*

dismantle

“The washed bamboo bird cage should not be exposed to strong sunlight in order not to damage the cage.”

- (15) <Fm>他们 悄悄地 察看着 你的 脸， 察看 你的 相貌、衣着、举止， 倾听着 你的 言谈 和 回答， <Fa>[[从而]] 试着 判定 你 是 个 怎样 的 人</Fa>。 </Fm>(LCMC\_F.xml/sn="0048")

<Fm>tamen qiaoqiaode chakan zhe ni de lian  
 they quietly examine PROG your GEN face  
 chakan ni de xiangmao yizhuo juzhi  
 examine your GEN appearance clothing behaviour  
 qingting ni de yantan he huida <Fa>conger  
 listen.to your GEN talk and reply in.order.to  
 shi zhe panding ni shi ge zenyang de ren</Fa></Fm>  
 try PROG judge you be CL what.kind.of DE person  
 “They examined in secret your face, appearance, clothing and manner and  
 listened to what you said and how you replied in order to gauge what sort of  
 person you are.”

- (16) <Fm>在 明确 个人 的 职业 意向 及 心理 特点 后， 应 进一步 设法  
 了解 所 倾向 的 职业 要求， 培养 相 调节 自己 所 倾向 的 职业 所  
 需 的 各种 心理 品质， <Fa>[[以免]] 今后 的 工作 处于 被动</Fa>。  
 </Fm>(LCMC\_F.xml/sn="0046")

<Fm>zai mingque geren de zhiye yixiang ji  
 at realise individual's GEN career objective and  
 xinli tedian hou ying jinyibu shefa  
 psychological traits after should further make.every.effort.to  
 liaojie suo qingxiang de zhiye yaoqiu  
 understand belong.to.oneself intended DE career requirement  
 peiyang xiang tiaojie ziji suo qingxiang  
 foster accordingly adjust oneself belong.to.oneself intended  
 de zhiye suo xu de gezhong xinli  
 DE career belong.to.oneself need DE various.kinds.of psychological  
 pinzhi <Fa>yimian jinhou de gongzhuo chuyu  
 qualities in.order.not.to future DE work remain.in.a.state.of  
 beidong</Fa></Fm>  
 unmotivated  
 “After understanding one’s career objective and personality traits, one should  
 learn more about the requirements of the desired career and prepare one’s  
 mentality for it in order not to be unmotivated later at work.”

To conclude this section, I have demonstrated that semantic classes of adverbial clauses and text categories are in a reciprocal relationship in that adverbial semantic classes have a propensity to occur in certain text categories (for example, clauses of condition and purpose are frequently used in category E for giving instructions), and different genres tend to prefer a particular semantic class of adverbial clauses (for example, academic prose takes more clauses of reason to explain research methodological issues).

### 7.3 The distribution of PROs across text types and semantic domains

In this section, I will examine how overt and non-overt subjects are used in adverbial clauses across text categories and across the semantic domains of adverbial clauses, by comparison to the results obtained from the PFR training corpus.

#### 7.3.1 *Text type and choice of subject*

Table 39 shows the distribution of overt and non-overt subjects in different text types. I tested the statistical significance of the relatedness between text category and choice of subject, both *across* and *within* text types. As the text categories are of different sizes, the log-likelihood value for determining the distribution of subjects across text types was calculated on the basis of the normalised frequencies of subjects. However, the log-likelihood scores of individual text categories were calculated on the basis of the raw frequencies of overt and non-overt subjects, which both occur within the same text type. While the distribution of overt and non-overt subjects is significantly different across text types (the calculated LL value with 14 d.f. is 105.762, greater than 36.12, the critical value for significance at  $p < 0.001$ ), the distribution of subjects within certain text categories does not differ significantly at this level. Only category N indicates a significant outcome; its LL score with 1 d.f. (29.60) is greater than the critical value for significance



(10.83) at  $p < 0.001$ .<sup>6</sup> Only if  $p$  is lessened to 0.05 do more significant results occur (when 3.841 becomes the critical value). At that level overt subjects occur more frequently in categories A, G, J, N, P and R. Of the fifteen text categories, only six genres were shown statistically to prefer one type of subject (i.e. overt subject) over the other (i.e. PRO) in adverbial clauses. Consequently, I would not claim that the relatedness between text type and choice of subject is of significance. However, this does not entirely exclude the influence of text type on the distribution of subjects, though that significance is only measurable at a lower level (95%). In the following, I will demonstrate that text types do play a role in the choice of subject in the adverbial clause.

Text types	Raw frequency		Freq. per 100K tokens		LL (1 d.f.)
	Overt NP subjects	PRO	Overt NP subjects	PRO	
A	71	45	80.7	51.1	<b>5.88</b>
B	103	107	190.7	198.1	0.08
C	40	47	117.6	138.2	0.56
D	60	78	176.5	229.4	2.35
E	183	170	240.8	223.7	0.48
F	201	170	228.4	193.2	2.59
<b>G</b>	194	142	126.0	92.2	<b>8.08</b>
H	14	22	23.3	36.7	1.79
<b>J</b>	369	313	230.6	195.6	<b>4.60</b>
K	72	58	124.1	100.0	1.51
L	64	55	133.3	114.6	0.68
M	18	17	150.0	141.7	0.03
<b>N</b>	132	58	227.6	100.0	<b>29.60</b>
<b>P</b>	88	52	151.7	89.7	<b>9.36</b>
<b>R</b>	27	11	150.0	61.1	<b>6.95</b>
Log-likelihood (LL) value (14 d.f.)			105.762		

Table 39: Subject types vs. text types.

- 6 Category N, *adventure and martial arts fiction*, is a distinctive language type in that the language used in most stories of this type is influenced by vernacular Chinese, i.e. modern Chinese styled to appear like classical Chinese. In this way, the use of adverbial clauses in this text category may be different from that in other text categories due to the difference between modern Chinese and classical Chinese. However, as the focus of my book is on adverbial clauses in modern Chinese, I will not explore this issue further in this book. Interested readers can refer to Wiebusch (2005) for a quantitative investigation of Chinese numeral classifiers based on historical corpora.

### 7.3.2 Adverbial semantic class and choice of subject

The discussion in Chapter Six (section 6.5) reported that the distribution of non-overt subjects of adverbial clauses varies significantly across semantic types of adverbial clauses. To recapitulate, clauses of concession and reason favour the use of overt subjects whereas clauses of condition, purpose, contrast and result prefer using null subjects. Table 40 shows the raw frequencies of overt and non-overt subjects of adverbial clauses in both the PFR and LCMC corpora.<sup>7</sup> While both subject types are slightly more frequent in the LCMC corpus than in the PFR corpus, the difference in the distribution of these two subject types between the two corpora is not statistically significant: Fisher's Exact Test fails to detect a significant difference even at the 95% level ( $p=2.105$ ). Given that the distribution of subjects of adverbial clauses is not sensitive to genre variations (as noted in section 7.3.1), it is not surprising that the two corpora, which are structurally distinct from each other in terms of their composition of genres, show no statistically significant difference in the occurrence of overt and non-overt subjects of Chinese adverbial clauses. On the other hand, the semantic class of the adverbial clause plays a vital role in determining the distribution of subjects, as demonstrated in the previous chapter using the PFR corpus. As the PFR corpus contains a single genre i.e. journalistic writing, it is unclear whether or not the influence of adverbial semantic class on the distribution of overt and non-overt subjects depends on text type. I therefore decided to investigate this question by using the LCMC corpus. As the variation in genres has been proved to be insignificant in the distribution of subjects in the adverbial clause, I hypothesise that the effect of adverbial semantic domain on the distribution of subjects is not dependent on text type, i.e. a similar pattern of the distribution of overt subjects and non-overt ones (PROs) in certain semantic kinds of adverbial clauses can be found in both the PFR corpus (a homogenous corpus of journalistic texts) and the LCMC corpus (a balanced corpus of fifteen distinct text types). To test this hypothesis, I compared the distribution of both subject types (i.e. overt subject and PRO) within eleven semantic classes of adverbial clauses in the LCMC corpus. As

7 As both the PFR and LCMC corpora are of equal sizes, containing one million word tokens, normalisation of frequencies is not required.

illustrated in Table 41, the distribution of overt and non-overt subjects varies significantly across semantic domains (the LL value with 10 d.f. is 535.750, considerably greater than 29.59, the critical value for significance at  $p < 0.001$ ). Five semantic types of adverbial clauses give statistically significant results with the log-likelihood values (with 1 d.f.) greater than the critical value for significance at  $p < 0.001$  (10.83). Whilst overt subjects are most frequently used in clauses of concession and less frequently used in clauses of reason and exception, null subjects are overwhelmingly used in clauses of contrast and purpose.

Subject types	PFR Corpus	LCMC Corpus	Fisher's Exact Test ( $p < 0.05$ )
Overt NP subjects	1,083	1,636	2.105
PROs	1,334	1,345	

Table 40: Frequencies of overt and non-overt subjects of adverbial clauses in PFR and LCMC.

Semantic types of CACs	Raw frequency		LL (1 d.f.)
	Overt NP subjects	PROs	
Clauses of time	2	0	2.77
<b>Clauses of reason</b>	381	81	<b>211.52</b>
<b>Clauses of purpose</b>	18	90	<b>52.40</b>
Clauses of result	65	81	1.76
Clauses of preference	2	11	6.86
<b>Clauses of contrast</b>	17	193	<b>173.06</b>
Clauses of addition	11	6	1.49
<b>Clauses of exception</b>	39	4	<b>33.00</b>
Clauses of condition	589	612	0.44
<b>Clauses of concession</b>	509	267	<b>76.74</b>
Clauses of inference	3	0	4.16
Log-likelihood (LL) value (10 d.f.)	535.750		

Table 41: Subject types vs. semantic types of CACs.

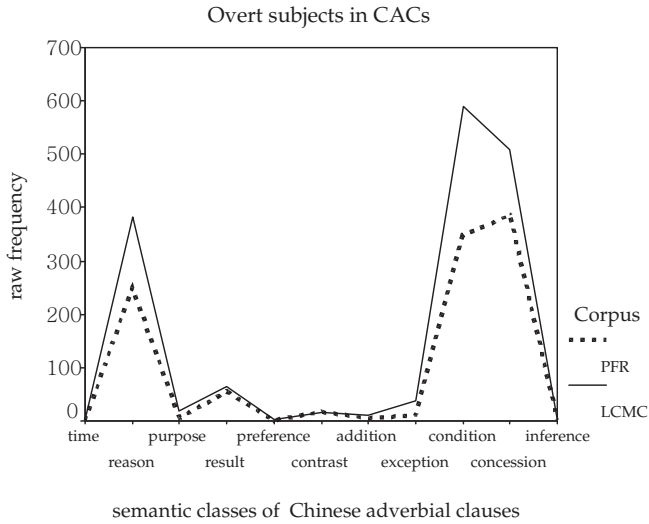


Figure 9: Contrasting overt subjects in PFR and LCMC.

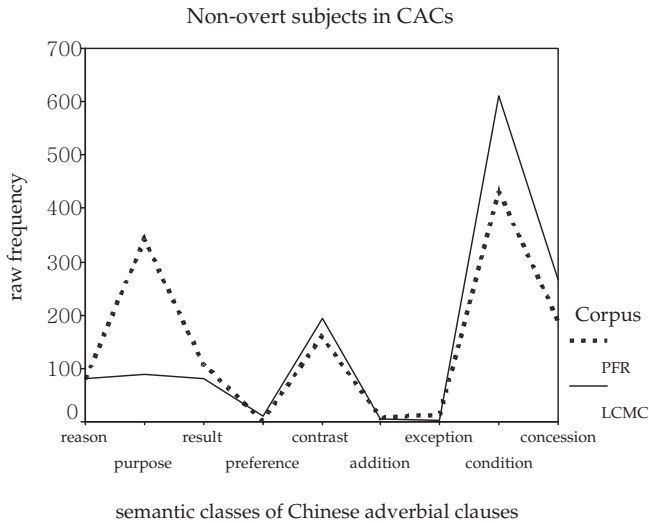


Figure 10: Contrasting non-overt subjects in PFR and LCMC.

LL scores for overt subjects vs. PROs					
Text types	Semantic classes of CACs				
	<i>Cause/reason</i>	<i>Purpose</i>	<i>Contrast</i>	<i>Exception</i>	<i>Concession</i>
A	<b>11.25</b>	0.00	<b>13.86</b>	2.77	<b>13.57</b>
B	6.06	<b>13.86</b>	<b>12.97</b>	0.00	9.23
C	1.65	0.34	5.55	0.00	0.04
D	3.95	1.93	9.75	0.34	0.40
E	<b>25.95</b>	2.64	<b>14.70</b>	1.39	1.62
F	<b>25.08</b>	2.98	5.78	4.16	0.89
G	<b>26.66</b>	2.91	8.71	<b>16.64</b>	4.63
H	<b>11.09</b>	<b>11.09</b>	0.00	0.00	0.34
J	3.17	<b>29.40</b>	<b>86.56</b>	8.32	<b>34.81</b>
K	2.36	1.39	<b>19.41</b>	1.39	6.95
L	2.38	0.00	2.94	4.16	6.34
M	7.72	0.00	5.55	2.77	0.00
N	<b>13.55</b>	2.77	0.34	5.55	<b>13.21</b>
P	5.88	0.00	9.70	0.00	<b>11.51</b>
R	0.40	0.00	0.00	2.77	1.33

Table 42: Contrasting the distribution of subjects across text types.

As far as the distribution of overt subjects is concerned, the results obtained from the PFR corpus can be perfectly mirrored in the data drawn from the LCMC corpus. As can be seen in Figure 9, the distribution patterns of overt subjects in the PFR and LCMC corpora are strikingly similar. However, as far as the distribution of non-overt subjects is concerned, two observations made on the basis of the LCMC corpus go against the results obtained from the PFR corpus. These two findings are telling arguments for the role of text type in the use of overt and non-overt subjects in the adverbial clause; this relationship should not be ignored simply because of the lack of significance when text type and choice of subject are contrasted (see section 7.3.1). As shown in Figure 10, purpose clauses take remarkably less PROs in the LCMC corpus than in the PFR corpus, possibly because the majority (87%) of the non-overt subjects of the purpose clauses occur only in six of the text types in the LCMC corpus i.e. scientific academic prose (17%), press editorials (17%), skills/trades/hobbies (15%), popular lore (15%), reports and official documents (12%) and religion (11%). Whilst a fall

in the use of non-overt subjects in clauses of purpose attaches some credence to the influence of text type on the distribution of subjects, the insignificant contrast between the distribution of overt and non-overt subjects in clauses of condition provides further evidence to support this influence. As shown in Table 41, conditional clauses do not show a preponderant use of PRO in the LCMC corpus (the calculated LL value with 1 d.f. is 0.44, much smaller than the critical value for significance, 10.83, at  $p < 0.001$ ). If the semantic domains of adverbial clauses had imposed an influence on the occurrence of PROs irrespective of genre variations, conditional clauses, which were proved to favour the use of non-overt subjects in the PFR corpus, would have taken significantly more PROs than overt subjects in the LCMC corpus. Thus it is necessary to examine more closely the five adverbial clause types which, as noted above, give significant results on their choice of subject, with the aim of identifying any difference in the distribution of subjects across genres in these adverbial semantic domains. Table 42 shows the log-likelihood scores of individual text categories in the contrast of overt and non-overt subjects, where statistically significant values are highlighted.<sup>8</sup> As illustrated in the table, the contrast between the distribution of overt and non-overt subjects in the five adverbial semantic classes under consideration is marked only in certain text types, and the influence of text type is most appreciable in clauses of exception in which the difference in the distribution of these two kinds of subject is significant only in category G. Hence, I rejected the null hypothesis and conclude that the effect of semantic domain on the distribution of subjects depends on text type. While text types do not have a direct influence on the distribution of PROs as indicated by the non-significant results between text type and subject type, they do indeed have an indirect influence on how the semantic domains of adverbial clauses determine the distribution of subjects in the adverbial clause. In the following section, I will investigate how text type influences the choice of referential controllers for PRO in CACs.

8 As both the overt and non-overt subjects of a particular kind of adverbial clause occur in the same text category, the log-likelihood test of significance was computed on the basis of the raw frequencies of the subjects. The calculated LL value (with 1 d.f.) is significant if it is greater than 10.83, the critical value for significance at  $p < 0.001$ .

## 7.4 Types of control of PRO and text types

The discussion of the properties of control of PRO in Chapter Six, section 6.4 brings out three distinct types of control for Chinese non-overt subjects: a non-overt subject of the Chinese adverbial clause is controlled by either an NP (subject or object) in the main clause or an NP outside the main clause i.e. in previous context, or it is not controlled by any NP and refers freely. While more or less the same number of non-overt subjects occur in both the PFR and LCMC corpora, the distribution of their controllers is different in the two corpora.

Chinese corpora	Main clause control		Control outside main clause	No control i.e. arbitrary interpretation	No. of PRO
	Subject control	Object control			
PFR	689 (52%)	12 (1%)	411 (31%)	222 (17%)	1,334 (100%)
LCMC <sup>9</sup>	436 (32%)	32 (2%)	424 (32%)	453 (34%)	1,345 (100%)

Table 43: Types of control of PRO in PFR and LCMC.

As can be seen in Table 43, subject control from the main clause slightly dominates the other two types of control in the PFR corpus. However, in the LCMC corpus, these three types of control average out in their distribution, with more or less the same frequency of occurrence. The effect of averaging out the three kinds of control of PRO in the LCMC corpus is caused by a surge in the use of non-overt subjects (PRO<sub>arb</sub>) which are not controlled by any element in the sentence and thus take an arbitrary reading, i.e. 453 instances in the LCMC corpus, the double of 222 instances in the PFR corpus. While more “arbitrary” non-overt subjects appear, there is a relatively reduced use of main clause subject control. In Figure 11, a drop in the use of main clause subject controllers can be clearly seen, as opposed to a rise in “no control” in the

9 For the record, object control in the LCMC corpus occurs in various text types, namely category B (*yinwei*, *jishi* and *ruo*), category C (*zhiyou*), category D (*yimian*), category E (*yimian*, *conger* (2) and *sui*), category F (*youyu*, *conger* (2), *ruguo*, *zhiyao*, *zhiyou* and *chufei*), category G (*youyu*, *yinwei*, *zhiyao*, *suiran*, *sui* and *buguan*), category J (*yimian*, *conger* and *suiran*), category K (*ruguo* and *napa*), category N (*ru*, *ruoshi* and *buguan*), and category P (*ruguo* and *yaoshi*).

LCMC corpus. As highlighted in Table 44, the evident growing use of  $PRO_{arb}$  is shown in clauses of condition, which have 314 cases of  $PRO_{arb}$  in the LCMC corpus, compared to 143 cases in the PFR corpus. More specifically, it was shown in Table 45 that clauses of condition take the greatest number of  $PRO_{arb}$  in category E among the fifteen text types of the LCMC corpus. This category comprises a wide range of articles on practical skills and hobbies. In articles of this sort, authors give instructions for how to manage a particular skill or hobby. As has been mentioned in section 7.2.1, conditional clauses are frequently employed in this genre to create a hypothetical situation in which the author provides an explanation. In these situations, there is no specific referent intended to be the reader. As a result, non-overt subjects with an arbitrary interpretation are overwhelmingly used in text type E, as illustrated in examples (17) and (18).

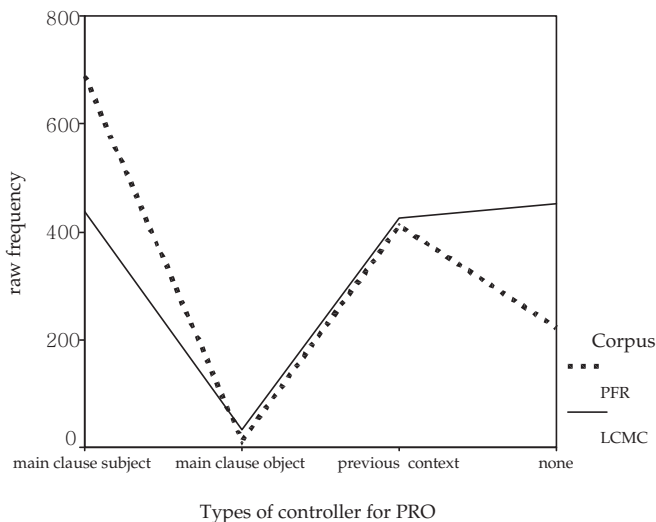


Figure 11: Contrasting types of controller for PRO in PFR and LCMC.



Non-overt subjects with an arbitrary interpretation (PRO <sub>arb</sub> )		
Semantic Classes of CACs	PFR Corpus	LCMC Corpus
Clause of Time	0	0
Clause of Cause or Reason	4	17
Clause of Purpose	35	21
Clause of Result	0	13
Clause of Preference	1	2
Clause of Contrast	5	1
Clause of Addition	1	4
Clause of Exception	0	3
<b>Clause of Condition</b>	<b>143</b>	<b>314</b>
Clause of Concession	33	78
Clause of Inference	0	0
Total:	222	453

Table 44: Distribution of PRO<sub>arb</sub> across semantic types of CACs in PFR and LCMC.

Text Types	Semantic Domains of Chinese Adverbial Clauses										
	<i>Time</i>	<i>Cause</i>	<i>Purpose</i>	<i>Result</i>	<i>Preference</i>	<i>Contrast</i>	<i>Addition</i>	<i>Exception</i>	<i>Condition</i>	<i>Concession</i>	<i>Inference</i>
<b>A</b>	0	0	0	0	0	0	1.1	0	11.4	4.5	0
<b>B</b>	0	3.7	0	1.9	1.9	0	0	0	38.9	1.9	0
<b>C</b>	0	0	0	0	0	0	0	0	35.3	17.6	0
<b>D</b>	0	14.7	0	0	0	0	0	5.9	50.0	29.4	0
<b>E</b>	0	1.3	6.6	2.6	0	0	0	0	<b>103.9</b>	13.2	0
<b>F</b>	0	2.3	4.5	0	0	0	1.1	0	48.9	10.2	0
<b>G</b>	0	0.6	0	0	0	0	0	0	14.9	4.5	0
<b>H</b>	0	0	1.7	0	0	0	0	0	3.3	0	0
<b>J</b>	0	1.9	6.9	6.3	0.6	0.6	0	0	46.9	11.9	0
<b>K</b>	0	0	0	0	0	0	1.7	0	41.4	5.2	0
<b>L</b>	0	2.1	0	0	0	0	2.1	0	20.8	6.3	0
<b>M</b>	0	0	0	0	0	0	0	0	25.0	0	0
<b>N</b>	0	1.7	0	0	0	0	0	0	6.9	10.3	0
<b>P</b>	0	0	0	0	0	0	0	1.7	12.1	3.4	0
<b>R</b>	0	5.6	0	0	0	0	0	0	5.6	0	0

Table 45: Distribution of PRO<sub>arb</sub> across semantic domains of CACs and text types of LCMC.

- (17) <Fm><Fa>[[如果]] **PRO<sub>arb</sub>** 递钩后，没见鱼上钩</Fa>，要上下提拉钓钩，以此诱鱼和显示钓饵可见度。

</Fm>(LCMC\_E.xml/sn="0023")

<Fm><Fa>*ruguo* **PRO<sub>arb</sub>** *di gou hou mei jian yu*  
 if PRO hold hook after have.not see fish  
*shanggou*</Fa> *yao shangxia ti la diaogou yici*  
 being.caught have.to up.and.down raise pull hookthereby  
*you yu he xianshi diaoer kejiandu*</Fm>  
 attract fish and show bait visibility

“If (you) hold the hook and do not see any fish being caught, (you) should move the hook up and down in order to attract prey and monitor the bait.”

- (18) <Fm><Fa>[[若]] **PRO<sub>arb</sub>** 用洗衣机洗衣</Fa>，只要在洗衣缸里放一杯醋，洗衣粉泡沫就会消失，不会溢出机外。

</Fm>(LCMC\_E.xml/sn="0004")

<Fm><Fa>*ruo* **PRO<sub>arb</sub>** *yong xiyiji xiyi*</Fa>  
 if PRO use washing.machine wash.clothes  
*zhiyao zai xiyigang li fang yi bei*  
 given.that in drum.of.washing.machine inside put one CL  
*cu xiyifen paomo jiu hui xiaoshi buhui*  
 vinegar washing.powder foam then will disappear will.not  
*yichu ji wai*</Fm>

splash.outmachine outside

“If (you) apply washing powder, (you) can pour a cup of vinegar into the drum of the washing machine so that the foam of the powder will disappear and will not spill out.”

## 7.5 Chapter summary

In this chapter, I used the Lancaster Corpus of Mandarin Chinese (LCMC), which is a balanced corpus containing fifteen distinct text types of written Chinese, as a test corpus and the PFR corpus as a training corpus. I examined the interrelationship between the semantic do-

mains of Chinese adverbial clauses (CACs) and the text types of the LCMC corpus. I also investigated how the non-overt subjects (PROs) of adverbial clauses are used in written Chinese and explored the influence of semantic domain and text type on the distribution of PROs in the adverbial clause.

In both the PFR and LCMC corpora, conditional and concessive clauses are the most common of the eleven semantic types of adverbial clauses. In a close examination of the press texts of the LCMC corpus exclusively (i.e. categories A, B and C), it is clear that conditional and concessive clauses also dominate all other semantic domains in these text categories. A hypothesis drawn from this is that journalistic texts are marked by conditional and concessive clauses. However, a statistical test of significance on the difference in the occurrence of these two kinds of adverbial clause between journalistic texts (of categories A, B and C) and non-journalistic texts (of the remaining categories in the LCMC) indicated that the difference is not significant. Hence, while conditional and concessive clauses are commonly used in press texts, they are not a marked feature of journalistic writing. Rather, they are most frequently used, respectively, in categories E (skills, trades and hobbies) and N (adventure and martial arts fiction). Other semantic classes of adverbial clauses also have a propensity to occur in certain text types of the corpus. Clauses of reason and result are closely related to categories D and J. Category D is chiefly concerned with the history of mainstream religions in mainland China and thus reason and result clauses are used to describe the origin and development of a religion, whereas category J is a collection of various science and social science academic texts. Thus reason clauses are commonly employed to give a rationale for adopting a particular research methodology and result clauses are used to present the research findings. As well as discussing research observations and results, result clauses are frequently used to give comments on the effects of political and economic events in category B which contains editorials from Xinhua news agency and local newspapers. Purpose clauses interact with category E (*ibid*) to give instructions for mastering a particular skill (e.g. craft and home decoration) and hobby (e.g. photography and fishing), as well as with category F (popular lore) to offer guidance on topics concerning local culture and daily life such as fashion, health, interpersonal communication skills, family, school and careers.

Whilst the distribution of subjects of adverbial clauses varies significantly across text types, overt and non-overt subjects do not differ significantly in their distribution within text types. This does not indicate, however, that text type does not influence the distribution of PROs in the adverbial clause. Results obtained from the PFR corpus have demonstrated that the distribution of non-overt subjects varies significantly across semantic domains of adverbial clauses: while non-overt subjects are used predominantly in clauses of condition, purpose, contrast and result, overt subjects are mostly used in concessive and causal clauses. In the LCMC corpus, the use of overt subjects confirms the results obtained from the PFR corpus, i.e. they occur significantly more frequently in clauses of concession, reason and exception. On the other hand, null subjects are used mostly in clauses of contrast and purpose. In contrast to the results obtained earlier from the PFR corpus, conditional clauses do not show a marked difference in the use of either subject type, though they were proved statistically to take significantly more PROs than overt subjects in the PFR corpus. One possible explanation is that text type has an influence on the distribution of PROs: in the five adverbial clause types which, as noted above, show a preference for a particular type of subject, the contrast between the distribution of overt and non-overt subjects is marked only in certain text categories of the LCMC corpus. This suggests that the effect of the semantic domains of adverbial clauses on the distribution of PROs, as shown in both the PFR and LCMC corpora, does indeed depend on text type. In other words, the distribution of PROs is a result of the semantic domains of adverbial clauses which is determined by text type.

Of the three distinct types of control of PRO identified in the PFR corpus (i.e. main clause control, control from previous context and no control), while main clause subject control slightly dominates in the corpus, these three kinds of control have more or less the same frequency of occurrence in the LCMC corpus. This averaging-out effect is brought about by a doubled number of PRO<sub>arb</sub> used in the LCMC (i.e. 453 occurrences) compared to that in PFR (i.e. 222 cases). The preponderant use of PRO<sub>arb</sub> is evident in text type E, which employs a huge number of conditional clauses to create a hypothetical situation for teaching a particular skill or hobby. As articles of this nature are not intended for a particular group of readers, it does not come as a surprise that non-

overt subjects with an arbitrary interpretation occur obtrusively in category E. Thus more occurrences of PRO<sub>arb</sub> are found in the LCMC corpus than in the PFR corpus, which is made up of press materials only. This demonstrates how text type influences the choice of the type of control of PRO in CACs.

