European University Studies

May L-Y Wong

Adverbial Clauses in Mandarin Chinese

A Corpus-based Study

Peter Lang

What are adverbial clauses in Chinese? Do they all have subjects as their counterparts do in English? How do the semantic domains of adverbial clauses interact with the distribution of subjects? How do Chinese corpora help us explore these intriguing questions?

The aim of this study is to demonstrate the usefulness of corpus linguistics as a methodology in grammar studies. A problem-oriented tagging approach has been used to enable the exploration of adverbial clauses in the corpus and to identify eleven semantically based classes of adverbial clauses. While it is a well-known fact that Chinese adverbial clauses (CACs) are overtly marked by a subordinating conjunction, their subjects can be left unexpressed and recovered in the prior discourse. By analysing naturally occurring spoken and written samples from various corpora, the author examines this intriguing phenomenon of overt and non-overt subjects in adverbial clauses.

May L-Y Wong received her Ph.D. in Linguistics in 2005 from Lancaster University. She is currently Honorary Assistant Professor of Linguistics at the University of Hong Kong. Her research interests include quantitative corpus linguistics, varieties of English, cognitively-inspired theories of language and the integration of corpus linguistics and cognitive linguistics.

Adverbial Clauses in Mandarin Chinese

European University Studies
Europäische Hochschulschriften
Publications Universitaires Européennes

Series XXI
Linguistics

Reihe XXI Série XXI
Linguistik
Linguistique

Vol./Band 374

PETER LANG
Bern · Berlin · Bruxelles · Frankfurt am Main · New York · Oxford · Wien

May L-Y Wong

Adverbial Clauses in Mandarin Chinese

A Corpus-based Study

PETER LANG
Bern · Berlin · Bruxelles · Frankfurt am Main · New York · Oxford · Wien

Bibliographic information published by die Deutsche Nationalbibliothek
Die Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available on the Internet at ‹http://dnb.d-nb.de›.

British Library Cataloguing-in-Publication Data: A catalogue record for this book is available from The British Library, Great Britain

Library of Congress Cataloging-in-Publication Data

Wong, May Lai-Yin,
Adverbial clauses in Mandarin Chinese : a corpus-based study / May Lai-Yin Wong.
pages cm. – (Europäische Hochschulschriften. Reihe XXI, Linguistik,
ISSN 0531-7320 ; Bd. 734 = Publications universitaires européennes.
Série XXI, Linguistique v. 374 = European university studies. Series XXI, Linguistics v. 374)
Includes bibliographical references.
ISBN 978-3-0343-1120-5
1. Chinese language–Adverbial. 2. Chinese language–Clauses. I. Title.
PL1233.W66 2013
495.1'576–dc23
2012049416

ISSN 0721-3352
ISBN 978-3-0343-1120-5 pb. ISBN 978-3-0351-0497-4 eBook

Hochfeldstrasse 32, CH-3012 Bern, Switzerland
info@peterlang.com, www.peterlang.com

Printed in Switzerland

Contents

Acknowledgements

The main input for this book arouse out of research done at both the University of Hong Kong and Lancaster University. In the course of carrying out my research, I have been fortunate to have had the opportunity of interacting with a number of leading researchers (either in terms of having studied with them or correspondence via electronic mail) on the subject of adverbial clauses and Chinese grammar, of whom the following merit special mention: Professor Tony McEnery, Professor Luke Kang-kwong, Dr Owen Nancarrow, Professor Geoffrey Sampson, Dr Willem Hollmann, Dr Richard Xiao, Dr Scott Piao, Mrs Catherine Nancarrow, and Dr Ya-ling Chang.

The data and corpora cited in this book have been taken from various sources. The PFR People's Daily POS Tagged Chinese Corpus was produced by the joint effort of the Peking University, the People's Daily newspaper and the Fujitsu Research and Development Centre. I wish to thank the Institute of Computational Linguistics at Peking University for permission to draw from the PFR corpus in the book. I am grateful to Professor Tony McEnery and Dr Richard Xiao for permission to use the data drawn from the Lancasater Corpus of Mandarin Chinese (LCMC) created by them and their research team at Lancaster University, and I hasten to add that whatever factual errors made in reporting on the LCMC corpus are mine alone. Parts of Chapter Seven and Eight appeared in my journal article published in Corpora; thanks are due to the editors for allowing me to incorporate the essay in this book.

I wish to record my many thanks to Martina Räber and her colleagues at Peter Lang Publishing for their guidance, patience and excellent copy editing.

Finally, I am indebted to my family, friends, colleagues and students for providing me with their moral support and various types of assistance.

Abbreviations

ADVL	Adverbial marker (地 *de*)
BA	*Ba* construction (把 *ba*, 将 *jiang*)
BI	Comparative construction (比 *bi*)
C	Coordinating conjunction
CAC	Chinese adverbial clause
CL	Classifier (e.g. 個 *ge*, 件 *jian*)
COMP	Complementiser (得 *de*, 到 *dao*)
DE	*De* construction (的 *de*)
EAGLES	Expert Advisory Group for Language Engineering Standards
EXP	Experiential aspect marker (过 *guo*)
GB	Government and Binding Theory
GEN	Genitive marker (的 *de*)
LCMC	Lancaster Corpus of Mandarin Chinese
LDC	Linguistic Data Consortium
MLCT	Multilingual Corpus Toolkit
NLP	Natural language processing
NP	Noun phrase
PART	Particle (e.g. 了 *le*, 的 *de*, 嘛 *ma*, 呀 *a*, 呢 *le*, 啊 *a*, 哈 *ha*, 啦 *le*, 吧 *ba*)
PASSIVE	Passive marker (被 *bei*)
PERF	Perfective aspect marker (了 *le*)
PL	Plural marker (们 *men*)
POS	Part of speech
PRO	Non-overt (or null) subject in a Chinese adverbial clause
PROG	Progressive aspect marker (在 *zai*, 着 *zhe*, 正 *zheng*, 正在 *zhengzai*)
S	Subordinating conjunction
SGML	Standard Generalised Markup Language
TEI	Text Encoding Initiative
UCREL	University Centre for Computer Corpus Research on Language
VP	Verb phrase
XML	Extensible Markup Language

Chinese Adverbial Subordinators

甭管 *bengguan* “no matter what”
不单 *budan* “in addition to”
不管 *buguan* “no matter what”
不论 *bulun* “whether or, whatever”
不论是 *bulunshi* “whether or, whatever”
不说 *bushuo* “let alone …”
除非 *chufei* “unless”
从而 *conger* “in order that; as a result”
而是 *ershi* “rather”
故 *gu* “so that”
果真 *guozhen* “supposing that”
何况 *hekuang* “not to mention …”
假如 *jiaru* “supposing that”
假若 *jiaruo* “supposing that”
即便 *jibian* “even if”
尽管 *jinguan* “even though”
既然 *jiran* “since, as”
即使 *jishi* “even if”
就是 *jiushi* “even if”
哪怕 *napa* “even if”
且不说 *qiebushuo* “let alone …”
任 *ren* “no matter what”
如 *ru* “if”
如果 *ruguo* “if”
若 *ruo* “if”
若果 *ruoguo* “if”
若是 *ruoshi* “if”
如若 *ruruo* “if”
尚且 *shangqie* “even”
虽 *sui* “though”
虽然 *suiran* “although”
虽说 *suishuo* “while admitting that …”
随着 *suizhe* “as soon as”
倘 *tang* “supposing that”
倘若 *tangruo* “supposing that”
万一 *wanyi* “in case”
无论 *wulun* “whether or, whatever”
无论是 *wulunshi* “whether or, whatever”
要 *yao* “assuming that”
要不是 *yaobushi* “if not, otherwise”
要是 *yaoshi* “assuming that”
以 *yi* “in order that”
以便 *yibian* “in order that”
以免 *yimian* “in order that… not…”
因 *yin* “because”
因为 *yinwei* “because”
以致 *yizhi* “as a result”
以至于 *yizhiyu* “consequently”
由于 *youyu* “owing to the fact that”
与其 *yuqi* “rather than”
与其说 *yuqishuo* “rather than say that”
只是 *zhishi* “except that”
之所以 *zhisuoyi* “why there is a consequence of”
只要 *zhiyao* “provided that”
只有 *zhiyou* “only if”
纵 *zong* “even if”
纵使 *zongshi* “even if”

Chapter One
Introduction

1.1 Corpus-based approach to studying adverbial clauses

This book is motivated by the fact that none of the previous analyses of adverbial clauses in Chinese have based their illustrative examples and exposition on extensive corpus evidence. Rather, researchers have typically relied on their own intuitions about language (e.g. Liu et al., 1996; Chu and Chi, 1999), sometimes supplemented by adapting example sentences from influential novels (e.g. Ding et al., 1979).

Recent work by Wang (1995, 1998, 1999 and 2002) breaks fresh ground in studying adverbial clauses by adopting a corpus-based approach. She quantitatively analyses the distribution and information structure of four main types of adverbial clause (viz temporal, conditional, concessive and causal clauses) in spoken Chinese on the basis of a corpus of six hours worth of naturally occurring face-to-face, two party, and multi-party conversations and call-in broadcasts on local radio and television in Taiwan. However, her studies focus solely on a limited range of adverbial clauses and are largely based on the spoken register of Chinese. Also her spoken corpus is rather small and yields just some 700 adverbial clauses in total. Hence, the novelty of a corpus-based study on adverbial clauses in written Chinese as well as an in-depth analysis on the typology of adverbial clauses in Chinese argue for a more thorough quantitative and qualitative account of them in order to discover new insights into their use in written data. Furthermore, as far as adverbial clauses are concerned, theoretically informed corpus-based research is rare (cf. Quintero, 2002). More work can therefore be done on marrying corpus linguistics with linguistic theory in this area. This book aims to achieve such a marriage.

To investigate the use and structure of a grammatical construction, most researchers have found it profitable to investigate constructions that occur relatively frequently, since if a construction occurs too in-

frequently, it is often hard to make strong generalisations about its form and usage (Meyer, 1991). For this reason, to study infrequent linguistic constructions, it is often necessary to study reasonably large corpora, like the two corpora of written Chinese used in this book, both of which contain one million words, namely the PFR Chinese Corpus and the Lancaster Corpus of Mandarin Chinese (henceforth LCMC).

1.2 Research objectives and the organisation of the book

Given the need for a corpus-based approach to linguistic theory and the need for a more extensive corpus-based account of a wider spectrum of adverbial clauses in written Chinese, this book uses a skeleton treebank (i.e. a corpus annotated with basic level syntactic constituents) to explore the syntactic structure of the Chinese language in order to shed light on the following research questions.

(1) How does the sample skeleton treebank help in the identification of adverbial clauses and revealing the peculiarities of Chinese syntactic properties?
(2) What are the adverbial subordinators in Chinese that are responsible for overtly marking adverbial clauses?
(3) Which semantic roles do these adverbial clauses play in relation to the main clause they modify?
(4) Does the PRO theorem in the Government and Binding (GB) Theory apply in written Chinese?
(5) How do the semantic types of adverbial clauses vary across genres/text types in written Chinese?
(6) How does the distribution of PROs vary across both semantic domains and text types in written Chinese?
(7) Do research findings based on written Chinese hold for spoken Chinese?

In the course of exploring the adverbial subordinators in the PFR corpus, a critique will be provided of the catch-all term 连词 *lianci* "conjunction" as it has been used in Chinese grammars to refer to both a

coordinating conjunction and a subordinating conjunction (Lu and Ma, 1990; Hou, 1998). As will be demonstrated in Chapter Six (section 6.1), Chinese is a pro-drop language (Huang, 1989) i.e. a language which allows the omission of a subject in a clause. While there is an immense literature on null subjects in Chinese (see, for example, Huang, 1987 and Chen, 1990), the focus of the previous literature was on pro-drop in complement clauses and not on pro-drop in adverbial clauses. Hence, my book contributes by investigating null subjects in Chinese adverbial clauses in order to fill the gap in the literature of the pro-drop phenomenon. In particular, this book focusses upon the distribution of non-overt subjects across various semantic types of adverbial clauses because certain of these adverbial clause types (e.g. purpose and contrast clauses) may show a stronger tendency for dropping the subject than other types. By *a priori* reasoning, if a person performs an action (as described in the main clause), s/he must intend to do it for a particular purpose (as described in the adverbial clause of purpose). Thus the subject of the purpose clause is likely to be omitted, which is always the same as the subject of the associated main clause. Clauses of contrast make a contrast between two situations described in the main clause and the adverbial clause. The two situations are closely related to each other as they are in fact two contrasting descriptions regarding the same subject; the situation of the main clause is taken to be wrong and the situation of the adverbial contrastive clause is what is right about the subject of the main clause. It is therefore hardly surprising that the subject of the contrastive clause, which is co-referential with the subject of the main clause, can be dropped. Yet all of these predictions stem purely from intuitions about the behaviour of adverbial clauses. To test these introspective assumptions, a corpus-based analysis is conducted in this book into how null subjects distribute across adverbial semantic classes.

In pursuing these research objectives, my book is organised into three major parts. The first part, Chapter Two, deals primarily with issues relating to the PFR Chinese Corpus, including a brief history of the construction of this corpus and the annotation of sentence boundary markers in the part-of-speech (POS) tagged corpus; the LCMC corpus will also be briefly described. Though it is a corpus-based study, my book is not atheoretical. In my book, a corpus-based approach to theory is advocated. The approach taken to the investigating of my research questions is as follows: rather than set out to use corpus data to testify the validity

of theoretical assumptions, I start my research by examining my corpus data closely, looking for any systematic patterns in the behaviour of the adverbial clause; those patterns or properties of adverbial clauses are then explained in a theoretical framework that lends itself well to the analysis of similar phenomena. In other words, my work does not presuppose the use of nor the rejection of a particular theoretical framework; rather, when it becomes relevant to my discussion of the corpus data, a theory is selected on its merits and adopted to explain my findings. Hence in the second part of my book (Chapters Three to Five), as a prelude to the theory based approach of the third part, initial results are presented in relatively theory-neutral terms, to make the emerging patterns of adverbial clauses in Chinese as accessible to linguists as possible. In the third part (Chapters Six to Eight), the same data are analysed within the Government and Binding Theory Framework in order to understand the description of the adverbial clause developed in this book within a theoretical framework in which a theorem (PRO theorem) is important to the explanation of the occurrence of non-overt subjects in the adverbial clause. The findings concerning the distribution of non-overt subjects are then put to the test in the LCMC corpus which, unlike the PFR corpus, is a balanced corpus with fifteen text types of written Chinese and can therefore provide a sound basis for making reliable generalisations of the properties of adverbial clauses in written Chinese across a range of genres. A contrastive study of the distribution of non-overt subjects in the adverbial clauses in spoken and written Chinese is also conducted on the basis of the CALLHOME Mandarin Chinese Transcripts Corpus, which is a spoken Chinese corpus developed in 1996.

1.2.1 Brief chapter summaries

Chapter One, the present chapter, expressly states the rationale and research objectives of this book.

Chapter Two will present a brief review of the development of written Chinese corpora and their use in linguistics and beyond. The two written corpora used in this book, the PFR Chinese Corpus and the LCMC corpus, will also be described. A literature review of previous studies of Chinese adverbial clauses will be presented, most of which is concerned with the discourse functions of adverbial clauses in different

positions in a sentence – either they are placed clause-initially (i.e. before all obligatory elements) or clause-finally (i.e. after all obligatory elements) – and this positioning influences the discourse functions the clause takes. Following from this, I will justify why a comprehensive analysis of the interclausal semantic roles of adverbial clauses is worth undertaking.

Chapter Three will describe at length how to produce a skeleton treebank by using a sample text of approximately 100,000 word tokens (amounting to about 2,500 Chinese sentences) to build up a body of examples of adverbial clauses. A clearly-defined parsing scheme will be given, which comprises 17 constituent labels and 11 textual markers, followed by a detailed treatment of the bracketing system and parsing guidelines. A discussion of the difficulties that I encountered during the manual parsing process will be included and the sample skeleton treebank will be evaluated against some quality-control criteria.

Chapter Four is devoted to the identification of adverbial subordinators that occur in the PFR Chinese Corpus. I will offer a critique of the vague grammatical category *lianci* that was traditionally employed in Chinese grammars to refer to both a coordinating conjunction and a subordinating conjunction. I will develop a more fine-grained approach to classifying different types of conjunctions in Chinese. A working definition of an adverbial subordinator and a list of subordinators used in my corpus will also be given. Some observations about the use of these adverbial subordinators will also be made.

Chapter Five will concentrate on a detailed and systematic typology of the adverbial clauses found in my corpus. I adopt a problem-oriented tagging approach in identifying the adverbial clauses together with their associated main clauses, and propose eleven interclausal semantic domains of adverbial clauses. Adverbial subordinators which are used to mark these semantic relations will be given and illustrated with examples drawn from the corpus. It will be shown that some semantic classes of adverbial clauses are linked so that they even share the same introductory adverbial subordinator: both clauses of purpose and result can be introduced by the adverbial subordinator 从而 *conger* "in order to; consequently". The syntactic and semantic characteristics of two interesting subordinators, 因为 *yinwei* "because" and 而是 *ershi* "rather", will be discussed. Finally, a table of the frequency of occurrence of each adverbial clause type will be provided.

Chapter Six will focus on non-overt NPs, represented as PRO in Government and Binding Theory, which occur as the subject of the Chinese adverbial clause. I will first justify the postulation of such an empty category and then examine its features, distribution and interpretation. More specifically, the feature composition of PRO will be argued to be anaphoric and pronominal, and the interpretation of PRO will be shown to be regulated by control theory in different forms of control such as obligatory control and subject control. PRO will be proved to occur only in ungoverned positions, and its distribution among different semantic types of adverbial clauses will also be discussed.

Chapter Seven will discuss the occurrence of the eleven semantic classes of adverbial clauses in the fifteen genres of the LCMC corpus as well as the distribution of non-overt subjects across both semantic domains of adverbial clauses and genres of written Chinese. The semantic types of adverbial clauses will be shown to occur more frequently in certain text types. The results obtained in Chapter Six will also be compared with those obtained in this chapter. The hypothesis that the effect of adverbial semantic domain on the use of PRO depends on text type will be tested, and the influence of text type on the type of control of PRO will be considered.

Chapter Eight will present the differences between spoken and written Chinese in the use of adverbial clauses. Previous accounts of differences between spoken and written registers will be addressed when they become relevant to my discussion of the behaviour of the adverbial clause in these two modes. The results obtained from the two written Chinese corpora (i.e. the PFR and LCMC corpora) will also be compared with those obtained from the CALLHOME corpus. In this chapter, I will investigate whether or not the distribution pattern of adverbial clauses in narrative texts of written Chinese can be found in spoken Chinese, and discuss the differences between spoken and written Chinese in the contrast of the type of control of PRO and the distribution of subjects across adverbial semantic domains between the LCMC and CALLHOME.

Chapter Nine will conclude the book by summarising my research findings. The potential limitations of this book will also be discussed. Finally, future research on the exploitation of Chinese corpora in language studies will also be suggested.

Chapter Two

General Review

In this chapter, a definition of the term corpus (section 2.1) and a brief survey of Chinese corpora (section 2.2) will be given. A detailed description of the PFR Chinese Corpus used in this book will also be provided. Section 2.3 of this chapter is concerned with my object of study, adverbial clauses. This section provides a review of the literature on corpus-based accounts of adverbial clauses with a discussion of features such as the characterisation of adverbial clauses and their syntactic/semantic and discourse properties. This review of previous work on adverbial clauses will provide a justification of the focus of this book on the adverbial clause.

2.1 Defining corpus

Modern corpora are typically large, finite, collections of language data in machine readable form. Yet a corpus is more than a simple collection of data; rather it is "a large and principled collection of natural texts" (Biber, Conrad and Reppen, 1998: 12). A principled collection is distinct from a haphazard collection of materials; it attempts to represent language in such a way as to control for a range of variables. These variables may be wide ranging, e.g. genre, speaker's age or context/situation of utterances. In attempting to control such variables, most corpus builders appeal to the concepts of *sampling* and *representativeness*. Sampling involves the process of determining how many text samples are required for valid generalisations to be made about the variety of language under examination and what range of language users need to be selected for a valid representation of the population supplying the text samples (Meyer, 2002: 40). A corpus is representative of the language variety being examined if it accurately reflects

the tendencies of that variety (Biber, 1990; 1993a; 1993b). Hence, a refined definition of a corpus is that a corpus is a collection of machine-readable texts which are of finite size and carefully sampled in order to be maximally representative of a language variety being studied (Leech, 1992; Kirk, 1994; McEnery and Wilson, 2001: 32; McEnery, 2003: 449–450).

It is important to note, as other corpus linguists do (see, for example, Chafe, 1992: 88; Biber, Conrad and Reppen, 1998: 9; Kennedy, 1998: 7; McEnery and Wilson, 2001: 25), in advance of reviewing corpus-based studies, that corpus-based analysis should not be treated as being in competition with linguists' intuitions and other theories of language; rather it is complementary to these approaches. Corpus linguistics is a methodology that can easily be combined with other forms of linguistics (Leech, 1992). My book is contributing to corpus linguistics in this spirit, by combining corpus and theoretical linguistics. In the following subsections, a brief history of Chinese corpora and their uses in various fields will be considered.

2.2 Chinese corpora and their application

With the creation of English corpora, Chinese corpora began to come into being. While the earliest machine readable corpora in English are the Brown corpus (Francis and Kučera, 1964) and the LOB corpus (Johansson et al., 1978), both of which sample the American/British English language used in 1961[1], the earliest machine readable Chinese corpus is the Modern Chinese Word Frequency Corpus (Beijing Yuyan Xueyuan Yuyan Jiaoxue Yanjiusuo, 1986), which commenced constructing in 1979 and was completed in 1983, as reviewed in Feng (2002) and Sun et al. (2002). As my book is mainly based on data taken from

1 The Frown (Freiburg-Brown) corpus (Hundt et al., 1999) and the FLOB (Freiburg-Lancaster-Oslo-Bergen) corpus (Hundt et al., 1998), both of which sample the American/British English language used in 1991, are modern equivalents of the Brown and LOB corpora respectively.

written Chinese corpora, it is important in this chapter to give a survey of the Chinese written corpora.[2]

2.2.1 A survey of corpora of written Chinese

One of the most well-known written Chinese corpora is the Mandarin Chinese PH (*People's Republic of China Xin Hua*) Corpus (Guo, 1993).[3] It comprises news texts taken from the Xinhua News Agency which were written between January 1990 and March 1991. After the publication of the PH corpus, many Chinese corpora came into being such as the Mandarin Chinese News Text[4], the PFR People's Daily Chinese Corpus (Yu, 1999), the Chinese Penn Treebank (Xia et al., 2000; Chiou et al., 2001; Xue et al., 2002) and the Lancaster Corpus of Mandarin Chinese or LCMC (McEnery et al., 2003)[5]. These corpora, with the exception of the LCMC corpus, are not balanced: instead of containing a range of genres of written Chinese, they consist of a homogeneous collection of journalistic texts[6] taken roughly from the period 1994 to 1998. It is inevitable that the texts that they contain are rather restricted and not particularly representative of written Chinese. In contrast, the Lancaster Corpus of Mandarin Chinese is a balanced and representa-

2 The information about the Chinese corpora that are described below is provided by the following websites:
(1) Linguistic Data Consortium (LDC): <http://www.ldc.upenn.edu/Catalog/> (accessed 15 May 2012);
(2) Oxford Text Archive (OTA): <http://ota.ahds.ac.uk/> (accessed 15 May 2012);
(3) David Lee's web page: <http://www.uow.edu.au/~dlee/CBLLinks.htm> (accessed 15 May 2012);
(4) Barbara Manuel's web page: <http://www.bmanuel.org/> (accessed 15 May 2012).

3 The corpus is in GB code and its cleaned-up (with punctuations and clearly recognised proper names) segmented version can be freely downloaded by FTP as a single file. URL: <ftp://ftp.cogsci.ed.ac.uk/pub/chinese/> (accessed 15 May 2012).

4 See LDC's web page at <http://www.ldc.upenn.edu/> (accessed 15 May 2012) for description.

5 See LCMC's web page at <http://www.lancs.ac.uk/fass/projects/corpus/LCMC/> (accessed 15 May 2012).

6 While most Chinese corpora are part-of-speech tagged, the Chinese Penn Treebank is annotated with syntactic bracketing.

tive corpus, divided into 2000-word samples representing varying text types/genres of written Chinese, including press reportage, editorials, government documents, technical writing and fiction. The corpus was built as a match, in terms of sampling frame, for the FLOB corpus (Hundt et al., 1998).

In this book, I use the PFR corpus as the training corpus to explore as many features of the adverbial clause as possible. I use the LCMC corpus as the test corpus to obtain a more comprehensive analysis of the use of adverbial clauses in written Chinese. While the LCMC corpus will be described in more detail in Chapter Seven, the PFR corpus is discussed in the following subsections as it will be used in Chapters Three, Four, Five and Six.[7]

2.2.2 The PFR Chinese Corpus

The PFR (*Peita-Fujitsu-Renmin Ribao*) People's Daily POS Tagged Chinese Corpus (abbreviated to PFR Chinese Corpus hereafter) Release 1.0 was produced by the joint effort of the Institute of Computational Linguistics of the Peking University, the People's Daily Newspaper and the Fujitsu Research and Development (R&D) Centre Limited. With the permission of the People's Daily News and Information Centre, this corpus was based on extracts taken from one of the most popular Chinese newspapers, the People's Daily, from January to June in 1998. The corpus was composed of about 27 million Chinese characters, which were properly segmented into some 10 million words. These words were then annotated with part-of-speech tags.[8] In April 2001, the three corpus builders agreed to provide the public with free access to part of their research output, which contains the set of newspaper extracts assembled in January 1998, totalling some 3,000,000 characters, corresponding to about 1 million Chinese words. This subcorpus, on which part of my book is based, is freely available to the research community to use. Yet the PFR Chinese Corpus consists of texts more re-

7 The reason why I used the PFR corpus rather than the LCMC for the bulk of this research is that the latter was not ready early enough (until my second year of study).

8 As a by-product of the corpus construction, the institute has recently launched in its web page an online tool for word segmentation and POS tagging for short or medium-sized Chinese texts.

stricted than most researchers would ideally like as it covers only one publisher, one genre and a very narrow time span.[9] However, its value should not be ignored as it is one of the largest part-of-speech tagged Chinese corpora freely available at the time of writing.

2.2.2.1 The PFR tagset

Texts in the PFR corpus are arranged in ascending order according to date, page number and paragraph number. Moreover, every word of the texts is POS tagged. The tagset[10] comprises 26 basic word classes, including noun (*n*), time word (*t*), space word (*s*), directional locality (*f*), numeral (*m*), classifier (*q*), non-predicate adjective (*b*), pronoun (*r*), verb (*v*), adjective (*a*), descriptive[11] (*z*), adverb (*d*), preposition (*p*), conjunction (*c*), auxiliary (*u*), modal particle (*y*), interjection (*e*), onomatopoeia (*o*), idiom[12] (*i*), fixed expression (*l*), abbreviation (*j*), prefix[13] (*h*), suffix[14] (*k*), morpheme[15] (*g*), unclassified item[16] (*x*), and punctuation (*w*). Apart from this basic set of 26 POS markers, proper nouns were divided into personal names (*nr*), place names (*ns*), organisation names (*nt*) and other proper nouns (*nz*). Furthermore, another 20 markers which

9 As the subset of PFR was not chosen by me but was made freely available online by corpus compilers, I would assume that the question of how to "balance" the selections is an issue beyond the scope of this book.

10 The tagset that was used in the PFR Chinese Corpus was actually extended from the one proposed in Yu et al. (1998).

11 Descriptives are typically formed by reduplication or compounding, for example, 实实在在 *shishizaizai* "indeed, really, honestly", 绿茵茵 *lüyinyin* "green", 久远 *jiuyuan* "far back, ages ago, remote", 烂漫 *lanman* "bright-coloured; unaffected".

12 In Chinese, idoms, or 成语 *chengyu*, are expressions with a frozen internal structure. Their constituents and structure cannot be described in terms of morphological categories. They have to be treated as single morphological units. They should be distinguished from the fixed expressions, or 习用语 *xiyongyu*, the internal structure of which can be broken down into meaningful morphological units.

13 Examples include 非 *fei* "not", 超 *chao* "super", 无 *wu* "not", 过 *guo* "too", etc.

14 Examples include 儿 *er* "little", 们 *men* "expressing plurality", 型 *xing* "model, type" , 式 *shi* "type, style", etc.

15 Examples are 桌 *zhuo* "table", 身 *shen* "body", 鸭 *ya* "duck", etc.

16 Unlike morphemes, unclassified items do not carry any meaning at all. They must be combined with another unclassified item to give a meaningful word. Examples are 鹌 *an* (-鹑 *-chun*) "quail", 蟑 *zhang* (-螂 *-lang*) "cockroach", 蛤 *ge* (-蚧 *-jie*) "clam", etc.

address the peculiarities of Chinese linguistics were used, allowing for linguistic investigations specific to Chinese. The following table gives a description of the abbreviations used in this PFR tagset.

No.	*Tagset*	*POS (in Chinese)*	*POS (in English)*
1	Ag	形语素	Adjective Morpheme[17]
2	a	形容词	Adjective
3	ad	副形词	Adjective as Adverbial[18]
4	an	名形词	Adjective with Nominal Function[19]
5	Bg	区别语素	Non-predicate Adjective Morpheme
6	b	区别词	Non-predicate Adjective
7	c	连词	Conjunction
8	Dg	副语素	Adverb Morpheme
9	d	副词	Adverb
10	e	叹词	Interjection
11	f	方位词	Directional Locality
12	g	语素	Morpheme
13	h	前接成分	Prefix
14	i	成语	Idiom
15	j	简略语	Abbreviation
16	k	后接成分	Suffix
17	l	习用语	Fixed Expression
18	Mg	数语素	Numeric Morpheme
19	m	数词	Numeral
20	Ng	名语素	Noun Morpheme
21	n	名词	Common Noun
22	nr	人名	Personal Name
23	ns	地名	Place Name
24	nt	机构团体	Organisation Name
25	nx	外文字符	Nominal Character String
26	nz	其它专名	Other Proper Noun

17 The definition of morpheme was clearly stated in the institute's corpus annotation manual which states that a morpheme refers to the smallest meaningful unit which cannot be used independently. In Chinese, many characters may have their own meaning but they cannot stand alone (Norman, 1988: 154–156). They have to be combined with another character or word in the word formation process. Therefore, an adjective morpheme, resembling a common adjective semantically, is a morpheme signifying an attributive meaning to the word to which it is attached.

18 It refers to those adjectives functioning as adverbial without any modification to their morphological or phonological form.

19 It refers to those adjectives which can fulfil nominal functions in a clause.

No.	*Tagset*	*POS (in Chinese)*	*POS (in English)*
27	o	拟声词	Onomatopoeia
28	p	介词	Preposition
29	Qg	量语素	Classifier Morpheme
30	q	量词	Classifier
31	Rg	代语素	Pronoun Morpheme
32	r	代词	Pronoun
33	s	处所词	Space Word
34	Tg	时间语素	Time Word Morpheme
35	t	时间词	Time Word
36	Ug	助语素	Auxiliary Morpheme
37	u	助词	Auxiliary
38	Vg	动语素	Verb Morpheme
39	v	动词	Verb
40	vd	副动词	Verb as Adverbial[20]
41	vn	名动词	Verb with Nominal Function[21]
42	w	标点符号	Punctuation
43	x	非语素字	Unclassified Item
44	Yg	语气语素	Modal Particle Morpheme
45	y	语气词	Modal Particle
46	z	状态词	Descriptive

Table 1: The 46 PFR tagset.

2.2.2.2 The format of the PFR corpus

The PFR Chinese Corpus was saved in a plain text with the file extension "txt". In this corpus, every single line represents a paragraph or headline in the original newspaper scripts. At the beginning of each line, a string of numbers provide detailed information on the date (YYYYMMDD), page number, article number and paragraph number of each line. These are separated from one another by a hyphen –. For instance, as illustrated below, the numerical string "19980101-01-001-001" indicates that the piece of text under examination is the first paragraph of the first article appearing on the first page of the newspaper dated 1st January 1998. This crude metadata was simply tagged as a number in the corpus.

20 It refers to those verbs functioning as adverbial without any change to their form, both morphologically and phonologically.

21 It refers to those verbs that have acquired some nominal functions in a clause.

19980101-	01-	001-	001
Date (YYYYMMDD)	**Page**	**Article**	**Paragraph**

19980101-01-001-001/m 迈向/v 充满/v 希望/n 的/u 新/a 世纪/n ——/w 一九九八年/t 新年/t 讲话/n （/w 附/v 图片/n 1/m 张/q ）/w
……

Figure 1: An excerpt of PFR text (1).

While there was no blank line between paragraphs of the same article, one blank line was used to mark the boundary between two different articles. For each single line, words were carefully and consistently segmented and tagged in the form of "word/POS tag". These words were then separated by two whitespaces and the last word of each paragraph (with no following word) was also followed by two spaces to ensure consistency.

Furthermore, square brackets are used to group together single words belonging to the same proper noun. For instance, in the text string, "通过/p [中央/n 人民/n 广播/vn 电台/n]nt 、/w *tongguo*/p [*Zhongyang*/n *Renmin*/n *Guangbo*/vn *Diantai*/n]nt 、/w 'Through the Zhongyang Renmin Broadcasting Station,'", the pair of square brackets [ ] marks the beginning and the end of a proper noun, which refers to a broadcasting organisation, the whole unit being tagged "/nt". Each single word inside the square brackets still carries its own POS tag.

19980101-01-001-006/m 在/p １９９８年/t 来临/v 之际/f ，/w 我/r 十分/m 高兴/a 地/u 通过/p [中央/n 人民/n 广播/vn 电台/n]nt 、/w [中国/ns 国际/n 广播/vn 电台/n]nt 和/c [中央/n 电视台/n]nt ，/w 向/p 全国/n 各族/r 人民/n ，/w 向/p [香港/ns 特别/a 行政区/n]ns 同胞/n 、/w 澳门/ns 和/c 台湾/ns 同胞/n 、/w 海外/s 侨胞/n ，/w 向/p 世界/n 各国/r 的/u 朋友/n 们/k ，/w 致以/v 诚挚/a 的/u 问候/vn 和/c 良好/a 的/u 祝愿/vn ！/w

Figure 2: An excerpt of PFR text (2).