A Corpus-Based Investigation of the Distributional Patterns of English and Chinese Pronouns

Hengbin Yan
Faculty of English Language and Culture
Guangdong University of Foreign Studies
China

Yinghui Li*
School of English and Education
Center for Linguistics and Applied Linguistics
Guangdong University of Foreign Studies
China

Abstract

In this study, we adopt a corpus-based approach to the analysis of the distributional patterns of major types of pronouns across different genres in two comparable balanced corpora in English and Chinese. Utilizing results from state-of-the-art grammatical parsers, we find considerable variation in the distribution of pronouns in different genres. While English tends to employ consistently more pronouns in every genre than Chinese, the distributional patterns of pronouns in the two languages across the genres are highly patterned and significantly correlated with one another, suggesting that pronouns can play similar functional roles in varying contextual situations in the two languages. Of the subtypes of pronouns in the two languages, five are found to be directly comparable. Personal pronouns are found to have the most similar (correlated) genre distribution in the two languages, while demonstrative pronouns share the least similarity. The distributional patterns for each pronoun type are investigated and their underlying functional and cultural implications discussed. Our study suggests that the identification of these classes of pronouns can in large part be automated with the help of state-of-the-art part-of-speech taggers and dependency parsers. The results of this study can inform future research and application involving pronouns, with implications ranging from cross-linguistic studies of grammatical features to second language acquisition.

Keywords: Pronouns, Cross-linguistic contrastive study, Corpus Linguistics

1. Introduction

Pronouns have been known to serve a wide range of linguistic and non-linguistic functions. As one of the most common classes of function words, pronouns play an important role in first language acquisition and have pedagogical implications in second language acquisition (Meissner, 2008; Tsao, 1982). In psychological studies, it has been found that pronouns can reveal personality traits and roles in interpersonal relationships, and help us interpret emotions and psychological health (Pennebaker, 2011).

Given its importance, research in pronouns in two of the most popular languages, English and Chinese, has received much attention in the past decades, culminating in a large number of studies investigating their syntactic, semantic, and functional properties(Campbell & Pennebaker, 2003; Hartung et al., 2016). Most of studies discuss only pronouns in one language, disregarding differences in pronoun usage in the two languages. Of the contrastive studies conducted, the majority have not been based on corpus data, relying instead on a handful of carefully hand-crafted examples from which to derive generalized rules. The remaining few corpus-based studies have mostly focused on one specific type of pronouns in general corpora without considering the relationships between different pronoun types across different text genres. In this study, we aim, through a comprehensive corpus-based study, to provide an overview of the distributional patterns of different pronoun types across different genres.

*Corresponding Author: Yinghui Li, e-mail:liyinghui@gdufs.edu.cn

Utilizing results from state-of-the-art grammatical parsers, we consider major comparable types of pronouns in English and Chinese, providing an in-depth and thorough survey of the pronoun landscape. Specifically, we aim to answer the following research questions:

(1) How does the distribution of different types of pronouns vary across genres?

(2) What are the similarities and differences between English and Chinese in the distribution of comparable types of pronouns across genres?

(3) Can current state-of-the-art grammatical analyzers be utilized for the classification and extraction of subtle features of grammatical structures such as pronouns?

2. Literature Review

Pronouns are a set of closed-class words with noun-like function. They behave syntactically like noun phrases but their meaning is general and undetermined, and only interpretable in context. Pronouns can be used to refer to entities in the shared perceptual environment (e.g. personal, demonstrative pronouns), co referential with a referring expression elsewhere in the sentence (e.g. possessive, reflexive pronouns), or with quantifier expressions as antecedents (e.g. indefinite pronouns) (Evans, 1980).

Type of Pronoun	English Examples in Corpus	Chinese Examples in Corpus
Personal	I, we, you, they	我/wo,我们/women, 你/ni, 您/nin, 他们/tamen
Possessive	mine, yours, hers, theirs	Personal Pronoun + 的/de
Reflexive	myself, yourself, herself, themselves	自/zi, 己/ji., 自己/ziji, 本身/benshen
Demonstrative	this, that, these, those	这个/zhege, 这些/zhexie, 那些/naxie, 那里/nali
Indefinite	anyone, everything, nobody	有些/youxie, 一切/yiqie, 每人/meiren, 大家/dajia, 人人/renren
Reciprocal	each other, one another	彼此/bici, 相互/xianghu, 互相/huxiang
Relative	that, who, whom, which, when, where	N/A
Interrogative	who, what, which, whom, whose	哪/na, 哪些/naxie, 哪里/nali, 什么/shenme, 谁/shui
Enumerative	N/A	等 等 /dengdeng, 谁 谁 /shuishui, 什 么 什 么 /shenme shenme

Table 1. Major types of pronouns in English and Chinese with examples

Table 1. lists the nine major types of pronouns in English and Chinese. Among these, relative pronouns are found only in English, and enumerative pronouns only in Chinese (their English semantic equivalents are considered adverbs). Reciprocal pronouns are a very infrequent type of pronouns with a very limited and specialized use (Biber et al., 1999:346). Possessive pronouns in Chinese belong under the type of personal pronouns, formed by adding the particle DE to a personal pronoun, and are conflated with possessive adjectives. For this reason, we exclude these four pronoun types from our contrastive analysis and focus on the remaining five: personal, reflexive, demonstrative, indefinite and interrogative pronouns.

The use of pronouns has rich implications in Second Language Acquisition. As (Tsao, 1982) points out, Chinese pronouns are often used for stressing and highlighting the reference. For this reason, Chinese learners of English often make the mistake of deleting pronouns in places where they are deemed necessary. They also mistakenly give too much stress to pronouns in English. Fortunately, these errors could be amended since students tend to imitate expert practices of using pronouns when they are prompted (Thonney, 2013).

The study of pronouns also has implications for advanced non-native users of English. Gao (2015)'s study of first person pronouns in English journals reveals that the use of pronouns varies cross-linguistically and cross-disciplinarily. Papers in physics tend to have more first-person pronouns, while Chinese scholars tend to use more of these than native speakers.

The number of first person pronouns in the conclusion section is also higher than in the introduction. Similar findings were presented in (Zhang, 2008), who argues that the differences in the use of first-person pronouns in academic papers is due to the fact that Chinese scholars tend to have an outdated view of the use of first-person pronouns.

In the field of politics, pronominal references are used by politicians to create multiple identities of themselves and others and to show affiliation or distance from others (Bramley, 2001).The reviewed researches suggest that the frequency and types of pronouns used in different contexts and genres can have cross-cultural, pedagogical as well as socio-political significance. The overuse and underuse of certain pronouns can mean purposeful manipulation of emotions or potential areas for improvement for language learners. However, for the study of the "norm" in the distribution of pronouns requires studies in corpus-based investigations across different genres, which is currently unavailable. This study aims to fill this gap by providing an overview of the common types of pronouns in English and Chinese.

3. Method

3.1 Corpora

Two balanced, comparable corpora were used for the contrastive analysis in this paper: the CLOB corpus and the TORCH corpus.

The CLOB corpus (Xu & Liang, 2012) is a corpus of written contemporary English modelled after the sampling frames of the original Brown corpus, with one million words collected around the year 2009. Following Brown, the corpus is divided into 15 genre categories and can be used for synchronic as well as diachronic studies of language change.

The one-million-word TORCH (Texts of Recent Chinese) corpus[†] is a balanced corpus of written Chinese designed by the same group of researchers to be directly comparable with other Brown-family corpora. It follows the same collection guidelines as the CLOB corpus and contains texts sampled around the same year (2009) as CLOB, divided into 15 genre categories. As such, it is directly comparable with CLOB.

Table 2. shows the genre categories and frequency statistics in CLOB and TORCH. Together, these two corpora serve as representative samples of present-day English and Chinese for the purpose of cross-linguistic contrastive analysis.

	Words (C)	Words (T)	Proportion (C)	Proportion (T)	No. of Docs(T)	No. Of Docs (C)
A: Press: reportage	100893	105918	8.68%	8.68%	103	136
B: Press: editorial	61328	66157	5.28%	5.42%	79	65
C: Press: reviews	39299	40473	3.38%	3.32%	30	54
D: Religion	38851	41800	3.34%	3.43%	24	23
E: Skills, trades and hobbies	81589	89302	7.02%	7.32%	61	50
F: Popular lore	111739	128251	9.61%	10.51%	48	55
G: Belles lettres, biography, essays	173989	180202	14.97%	14.77%	75	101
H: Miscellaneous (documents, reports, etc.)	67595	71933	5.82%	5.89%	30	31
J: Learned and scientific writings	182883	189510	15.73%	15.53%	80	100
K: General fiction	69356	70915	5.97%	5.81%	29	30
L: Mystery and detective fiction	57856	58681	4.98%	4.81%	24	24
M: Science fiction	14678	14672	1.26%	1.2%	6	6
N: Adventure and western fiction	69912	69627	6.01%	5.71%	29	31
P: Romance and love story	70808	70724	6.09%	5.8%	29	29
R: Humour	21521	22144	1.85%	1.81%	10	9

Table 2. Genre Categories and Frequency Statistics in CLOB and TORCH

3.2 Pronoun Extraction

For the extraction and classification of the various types of pronouns, each text in both corpora is parsed with the Stanford Parser, a state-of-the-art statistical parser for multiple languages including English and Chinese. Using the parser, Part-of-Speech (POS) tags as well as dependency relations between the words in a sentence are extracted. The dependency relations produced by the Stanford Parser are analyses of the grammatical structure of sentences based on an annotation framework called Universal Dependencies (UD). UD has been designed to

[†] Available at http://corpus.bfsu.edu.cn/content/torch2014

provide a cross-linguistically consistent inventory of categories that aim to make annotation and identification of similar constructions across different languages possible (i.e. capable of bringing out cross-linguistic parallelism), while at the same time permitting annotations of language-specific features when necessary.

Using the Python programming language, each text in each corpus is read and passed to the parser which produces annotations of the text. Following a number of predefined rules, pronouns from each corpus are extracted based on lexical (POS tags) as well as grammatical (dependency parses) information.

The processes of pronoun extraction are slightly different for Chinese and English. Chinese pronouns are relatively straightforward to extract as they have been unambiguously tagged by the parser. Words with the tag PN (pronoun) are tagged as pronouns: each Chinese pronoun extracted from the corpus is then classified into five subtypes: Personal, Reflexive, Demonstrative, Indefinite, Interrogative.

Personal	我/wo、我们/women、咱/zan、咱们/zanmen、俺/an、偶/ou、吾/wu、本人/benren、你/ni、您/nin、你们/nimen、他/ta、他们/tamen、她/ta、她们/tamen、其/qi、之/zhi、俩/liang、双方/shuangfang、两者/liangzhe、他人/taren、别人/bieren、对方/duifang、人家/renjia、它/ta、它们/tamen
Reflexive	自/zi、己/ji、自己/ziji、自我/ziwo、自身/zishen、本身/benshen、各自/gezi、亲自/qinzi
Demonstrative	这个/zhege、这儿/zheer、这天/zhetian、这样/zheyang、这边/zhebian、这里/zheli、此地/cidi、此处/cichu、如此/ruci、那个/nage、那位/nawei、那儿/naer、那边/nabian、那里/nali、以上/yishang、以下/yixia、后者/houzhe、这些/zhexie、这/zhe、那/na、此/ci
Indefinite	有些/youxie、有人/youren、有的/youde、一切/yiqie、每人/meiren、大伙/dahuo、大家/dajia、各位/gewei、个个/gege、人人/renren
Interrogative	哪/na、哪些/naxie、哪儿/naer、哪里/nali、什么/shenme、何/he、啥/sha、谁/shui、多久/duojiu、何处/hechu、何时/heshi

Table 3. Types of Chinese Pronouns extracted from the TORCH Corpus

The identification and classification of English pronouns are less straightforward since English has finer distinctions and a richer typology of pronouns. One difficulty in making a distinction between determiners and pronouns, which are functionally distinct since the former modifies a following noun and the latter are freestanding and identify or point to nouns. Some words (*this, those*) can serve as either determiners or pronouns depending on whether they are followed by a noun which they modify. However, relying on lexical information alone to resolve such ambiguity can be problematic since the grammatical parsers tend to tag the part-of-speech in both cases as *DT* (determiner). To address this problem, we consult the dependency role that the determiner/pronoun plays at the clause level, and filter out determiners by keeping tokens that serve as the nominal subject (UD relation: *nsubj* or *nsubjpass*) instead of determiner (UD relation: *det*) in the clause. Another difficulty lies in separating interrogative pronouns from relative pronouns, some of which share the same word forms (e.g. *who, which*). To distinguish the two, we specify that only those words with the POS tag of *WP* (wh-pronoun) serving the grammatical functions of either a subject or object (UD relation: *nsubj* or *nsubjpass*) in a clause are interrogative pronouns.

Based on a number of fine-tuned heuristics, we extracted from the English corpus the following subtypes of pronouns:

Personal	i, you, he, she, it, we, they, me, him, her, us, them, 'em
Reflexive	myself, yourself, himself, herself, itself, ourselves, yourselves, themselves
Demonstrative	this, these, that, those, former, latter
Indefinite	something, anything, everything, nothing, someone, anyone, everyone, no one, none, somebody, anybody, everybody, nobody
Possessive	my, mine, your, yours, his, her, hers, its, our, ours, your, yours, their, theirs
Relative	that, who, whom, which, when, where, whatever, whoever, whomever, whenever
Interrogative	who, what, which, whom, whose

Table 4. Types of Chinese Pronouns extracted from the CLOB Corpus

The results of the pronoun classification in both English and Chinese are investigated by the authors, who manually fine-tune the rules for more accurate extraction. Despite the authors' greatest efforts, however, the automatic nature of the extraction process means that there can still be minor errors in the identification and retrieval of different types of pronouns.

After all the pronouns in both corpora are classified into appropriate types, we compute the overall frequencies of each type of pronouns, both in raw numbers and in percentage terms. Then the frequency of each pronoun type across the genres in the corpora is computed. We compare the distribution of each pronoun type across genres between English and Chinese by checking potential correlations that might exist between them. Then we attempt to discuss and explain the similarities and differences between the distributions.

4. Results and Discussions

4.1 Overall Distribution

Pronoun Type	Percentage in CLOB	Percentage in TORCH
Personal	4.2%	2.45%
Reflexive	0.11%	0.23%
Demonstrative	0.29%	0.26%
Indefinite	0.19%	0.07%
Interrogative	0.16%	0.11%
Possessive	1.57%	N/A
Relative	1.37%	N/A

Table 5. Percentages of each pronoun type in CLOB and TORCH

Table 5. shows the percentage of each type of pronoun in the two corpora. In both corpora, the most common pronouns are personal pronouns, accounting for 4.2% and 2.45% of all corpus texts. Altogether the pronouns in English take up about 7.89% of all texts, which are similar to the 8% in an earlier report of pronoun percentage in the LOB corpus (Hudson, 1994), significantly more than Chinese pronouns (3.12%) in Chinese. These statistics confirm previous non-corpus-based research that pronouns are used much more frequently in English than in Chinese (Li & Liang, 1999). The reason, it is argued, is due to the omission of possessive and reflexive pronouns in Chinese. The same trend is seen in the use of personal pronouns in Chinese. Lv (Lv, 1999:8) for example, has pointed out that Chinese tends to omit personal pronouns as much as possible. (Tsao, 1982)suggests that English use pronouns to indicate coreferentiality while Chinese is much more tolerant of pronoun deletion.

Our categorized statistics suggest that since personal pronouns are the most common type of pronouns in both English and Chinese texts, the tendency to omit them in Chinese result in the significantly higher percentage in English. Three other types of pronouns, demonstrative, indefinite and interrogative, are also higher in percentage in English than in Chinese. However, contrary to(Li & Liang, 1999), a higher percentage of reflexive pronouns are found in Chinese than in English, a finding which we will further discuss in the Discussion section. The lower overall percentage of pronouns in Chinese is also partly due to the absence of two other types of pronouns: possessive and relative pronouns, which amount to nearly 3% in English texts.

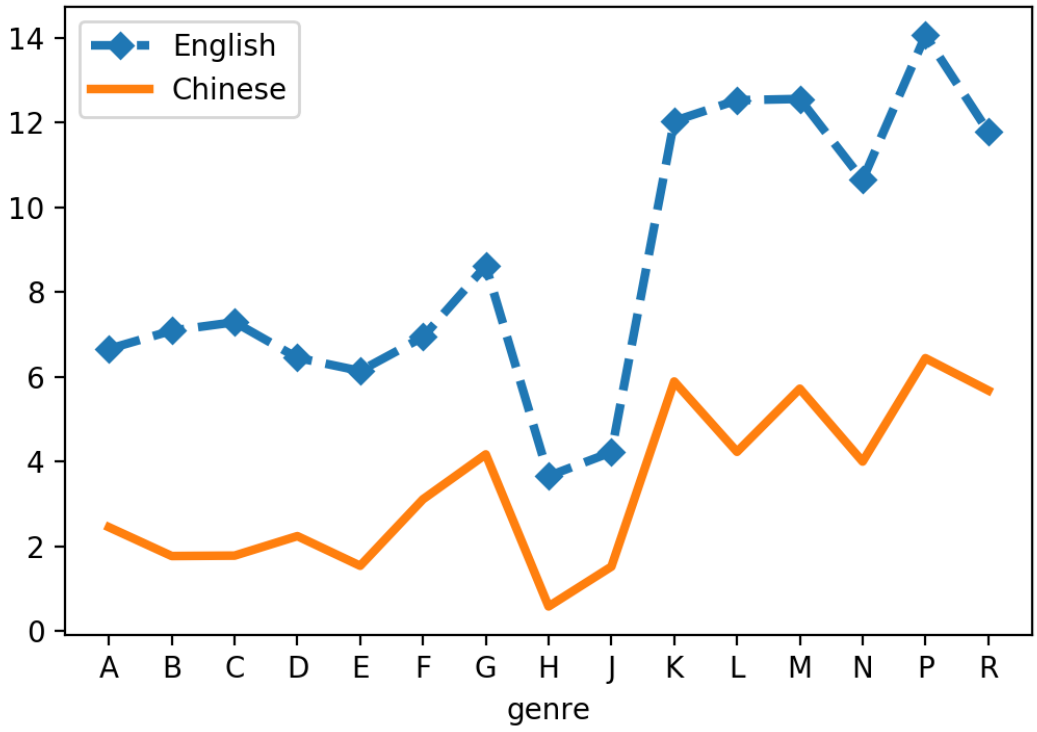

Figure 1. Distributions of pronouns across different genres in English (CLOB) and Chinese (TORCH)

Figure 1. shows the cross-genre distribution of pronouns in English and Chinese. The similarity in the distribution patterns is immediately obvious to the viewer. Low frequencies of pronouns are found in both English and Chinese in formal genres such as E (Skill & hobby), H (Miscellaneous government and house organs), J (Learned), followed by press-related genres (A-D). High frequencies are found in fiction-related genres such as (K, L, M, N, P) and humor (R).Popular lore (F) and Belles lettres, biography, essays (G) are somewhere in the middle. Comparing the cross-genre scores, a Pearson correlation coefficient of 0.88 ($p < 0.001$) is found between the two languages. This correlational patterning shows that while English and Chinese pronouns differ in raw percentages, their distribution across different genres are highly uniform. Since genres are "different ways of using language to achieve different culturally established tasks, and texts of different genres are texts which are achieving different purposes in the culture", it appears that, despite the significant differences between the two cultures leading to different traditions in the use of pronouns, pronouns in English and Chinese are used to achieve roughly the same purpose in the two cultures.

4.2 Distribution of Each Pronoun Type

We have discussed the overall distribution of pronouns in English and Chinese across genres. The next question to answer is, do the distribution of specific types of pronouns pattern in the same way as the overall distribution? To find out, we explore in this section the cross-genre distribution of each pronoun type. Our initial investigation indicates that the overall patterning of different kinds of pronouns is surprisingly similar across genres. Of the five directly comparable types of pronouns, four (personal, reflexive, indefinite and interrogative) have distribution across genres that are highly correlated with its counterparts in the other language. In the following, we survey the distributional statistics each of the pronoun types and discuss their implications.

4.2.1 Personal Pronouns

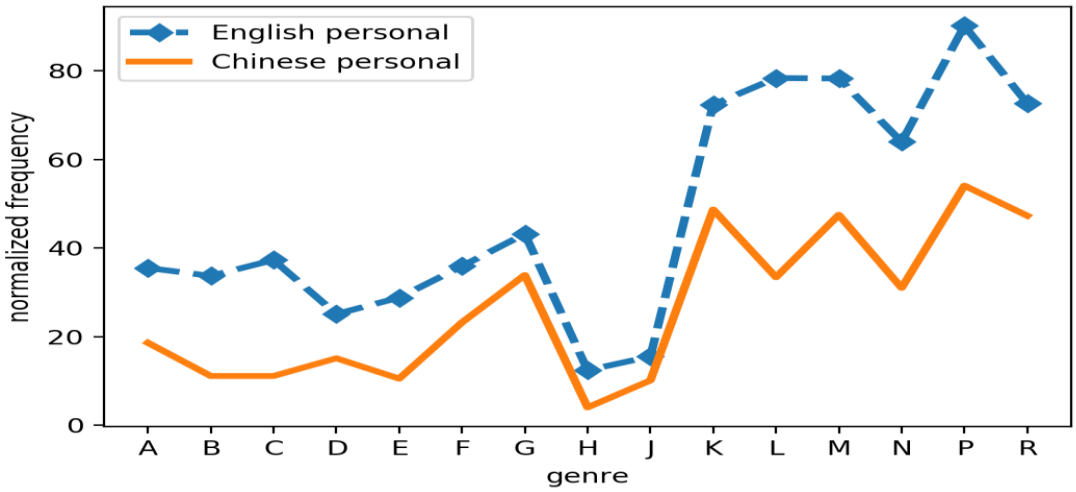

Figure 2. Cross-genre frequencies of personal pronouns in English and Chinese

Word (EN)	Percentage (EN)	Word (CN)	Percentage (CN)
it	20.12	我/wo	25.79
I	15.71	他/ta	20.06
he	14.84	她/ta	13.03
you	9.16	你/ni	11.78
they	7.67	我们/women	7.27
she	7.3	之/zhi	4.78
her	6.92	他们/tamen	4.31
we	6.22	其/qi	3.73
him	3.68	它/ta	2.97
them	3.3	您/nin	0.79
me	3.08	你们/nimen	0.78
us	1.94	别人/bieren	0.71
'em	0.07	它们/tamen	0.6
		对方/duifang	0.53
		双方/shuangfang	0.52
		她们/tamen	0.51
		人家/renjia	0.5
		他人/taren	0.32
		咱/zan	0.24
		俩/liang	0.21
		咱们/zanmen	0.2
		本人/benren	0.15
		两者/liangzhe	0.11
		偶/ou	0.04
		吾/wu	0.04
		俺/an	0.04

Table 6. Percentage of each type of personal pronouns in the two corpora

Personal pronouns are associated with the grammatical person. The first and second person pronouns are often used to refer to those directly involved in the discourse situation (e.g. the speaker and the listener), while the third person pronouns are used primarily to refer to persons/things inferable from context.

Figure 2. shows the normalized frequencies of personal pronouns across different genres in English and Chinese. In both English and Chinese, personal pronouns constitute the largest percentage of all pronouns investigated. It can be seen that English generally has more personal pronouns than Chinese (with the exception of Popular Lore, where the frequency in Chinese slightly exceeds English). There is a statistically significant correlation between English and Chinese across the genres (0.866, $p < 0.001$).

From Table 1., it can be seen that the personal pronoun with the highest frequency in English is the third person pronoun "it", which is considered the most neutral and semantically unmarked of the personal pronouns and is often used as an "empty" or anticipatory subject in denoting time, distance etc.(Quirk et al., 1985). *It* is followed by subjective pronouns such as *I, he, they, she*, which collectively are more frequent than objective ones (*her, him, them, me*). A rough Chinese equivalent of *it* is the third person pronoun 它/ta, which usually refers to non-human animate objects like animals. However, ta1 does not serve as the anticipatory subject, which explains its lower frequency percentage compared with *it* (2.97% vs. 20.12%).

The list of personal pronouns identified in the Chinese corpus is also significantly longer than English. This is because while Chinese does not have distinctions of case (the only exception being 之/zhi, which can only be used in the objective case) , it does have distinctions of number, gender, and in addition, politeness (你/ni/you vs 您/nin/respected you) and formality (他/ta/he, 他们/tamen/they/them etc. vs. 其/qi/formal third person pronoun, 之/zhi/formal third person pronoun, 我/wo/I vs. 咱/zan/colloquial I, 俺/an/colloquial I, 吾/wu/formal I).

Our statistics show that compared with English, Chinese personal pronouns are lower in frequency. This confirmsthe finding of (Hong, 1985), whopoints out that following Confucian teachings, second personal pronouns are discouraged in Chinese except in close relationships, and kin terms and titles etc. are used instead. This finding may also suggest that Chinese is similar to other Asian languages such as Korean, where similar patterns are found, with similar cultural roots. Kim (2009)for example, suggests that the lower frequency of personal pronouns in Korean may be due to agent omission which has roots in typical Asian indirectness and various socio-cultural differences between the two cultures.

4.2.2 Demonstrative Pronouns

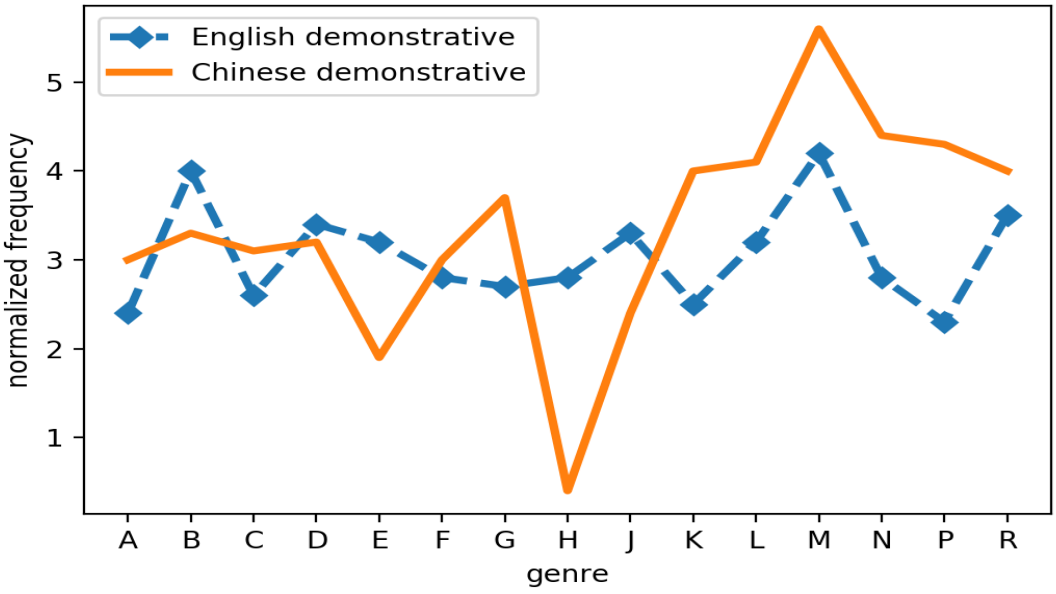

Figure 3. Cross-genre frequencies of demonstrative pronouns in English and Chinese

Word (EN)	Percentage (En)	Word (CN)	Percentage (CN)
this	45.73	这/zhe	45.37
that	38.56	这里/zheli	12.15
those	9.77	此/ci	9.48
these	5.24	那/na	8.83
latter	0.53	那里/nali	5.45
former	0.18	这样/zheyang	4.59
		这个/zhege	2.63
		这儿/zheer	1.64
		那儿/near	1.43
		那边/nabian	1.39
		这些/zhexie	1.3
		那位/nawei	1.24
		以上/yishang	0.9
		这边/zhebian	0.71

Table 7. Percentage of each type of demonstrative pronouns in the two corpora

Demonstrative pronouns mark something as known, and specify reference in terms of proximity (whether the referent is near or distant relative to the addressee). The words used as demonstratives are *this* and *that* as well as their plural forms (*these and those*). The Chinese equivalents are 这/zhe and 那/na, which serve as the prototype for compound demonstratives (e.g. 这样/zheyang, 那边/nabian). Chinese uses demonstrative pronouns significantly more frequently (up to four times as frequently) than English, except in genre H (Learned in Scientific Writing). Demonstrative pronouns are the only type of pronouns in English and Chinese that are not positively correlated ($r = 0.236$, $p > 0.05$). Compared with English, whose normalized frequencies are more or less stable across the genres (as indicated by the relatively flat line), Chinese has much more variation (as indicated by the fluctuating line) in the usage of demonstratives. Chinese demonstratives are used much more frequently in fictions than non-fictions.

Demonstrative pronouns are often studied in conjunction with personal pronouns. There has been some evidence that these two types of pronouns exhibit different patterns in terms of function and usage. (Gundel et al., 2004)for example, has found that demonstrative pronouns have more non-NP antecedents than NP antecedents, which is opposite to the case of personal pronouns. Different from other pronouns, many demonstratives also have no identifiable antecedents, and generally prefer complex entities as antecedents (Brown-Schmidt et al., 2005). These semantic and functional differences may explain the differences in the patterning of demonstrative pronouns from the other pronouns (which have shown similarity in patterning).

4.2.3 Reflexive Pronouns

Reflexive pronouns "reflect" other nominal element of the clause/sentence, forming with it a co referential relation. Functionally, it can serve as the object/complement of its antecedent, which is usually the subject. In addition, it can also be used for emphasis (as in *We ourselves couldn't come.*).

Figure 4. shows that Chinese has consistently more reflexive pronouns than English in all genres. The only exceptions are in the H (Miscellaneous (documents, reports, etc.), where the percentages are almost identical and R (Humor), where English tends to employ more reflexives. There is a fair degree of correlation between English and Chinese reflexive pronouns ($r = 0.617$, $p < 0.05$).

Both English and Chinese reflexive pronouns serve two functions: anaphoric and intensifying. However, they are manifested in different forms. The reflexive pronouns in English are formed by objective personal pronouns/possessive pronoun + self, distinguished by number (self vs selves). The different forms of English reflexives are more or less evenly distributed, with third person reflexive pronouns constituting the majority. In

Chinese, the prototypical reflexive pronoun ziji/自己, the equivalent of the English *self*, is predominantly frequent, accounting for nearly 80%, with a number of variations (e.g. zi/自, qinzi/亲自, benshen/本身).

Unlike English, Chinese has no distinction of person and number in the use of the variations of ziji, and the meanings of their English equivalents are either implied or expressed through the combination with another personal pronoun or noun (ta-ziji/他自己 or Zhangsan-ziji/张三自己).

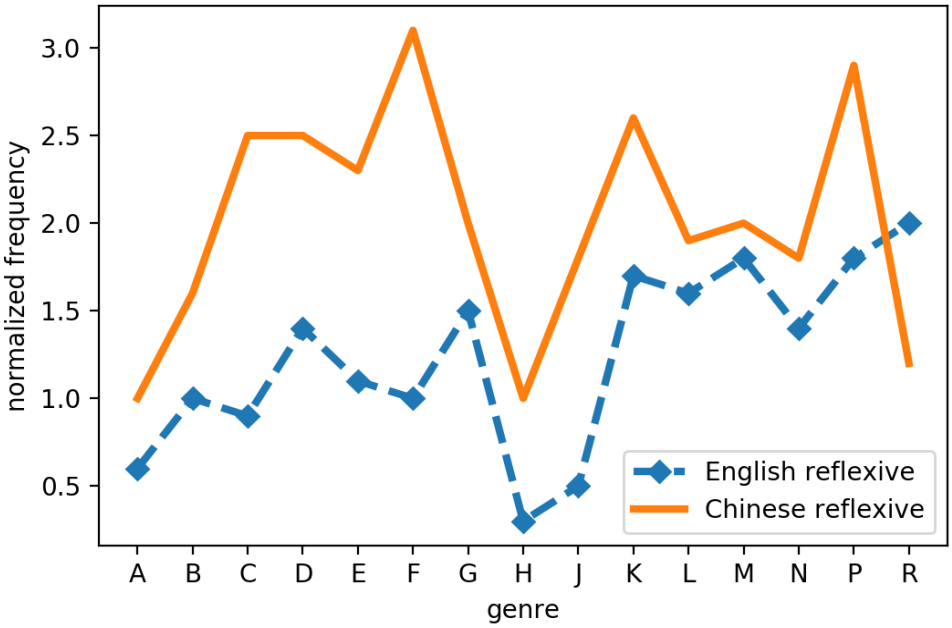

Figure 4. Cross-genre frequencies of reflexive pronouns in English and Chinese

Word (EN)	Percentage (En)	Word (CN)	Percentage (CN)
himself	26.12	自己/ziji	78.88
themselves	19.09	自身/zishen	6.05
itself	17.77	本身/benshen	4.78
myself	13.14	自我/ziwo	4.63
herself	10.82	各自/gezi	2.41
yourself	8.73	本人/benren	1.7
ourselves	3.79	己/ji	0.99
yourselves	0.54	自/zi	0.57

Table 8. Percentage of each type of reflexive pronouns in the two corpora

The difference in frequencies in English and Chinese pronouns, to our knowledge, has not been discussed in other studies. One reason that may account for the difference is that there are many usages of ziji in Chinese that equivalently translate into personal pronouns or determiners:

(1) 据野坂回忆，**自己**在延安得到与毛泽东等中共主要领导人同等的生活待遇。

Nosaka recalls that in Yanan **he** enjoyed the same living conditions as CPC leaders such as Mao Zedong.

(2) 作为记者，我们还要忠实于**自己**的记录和报道。

As a reporter, we have to be faithful to our **own** records and reporting.

(3) 梁思成尽了自己最大的努力想保护北京的古建。

Liang Sicheng has devoted **his** greatest efforts to the protection of the historical buildings in Beijing.

Because of the conflation of *ziji* with other common pronouns/determiners, it appears more frequently in the corpus than the English counterparts which are more restricted in meaning.

4.2.4 Indefinite Pronouns

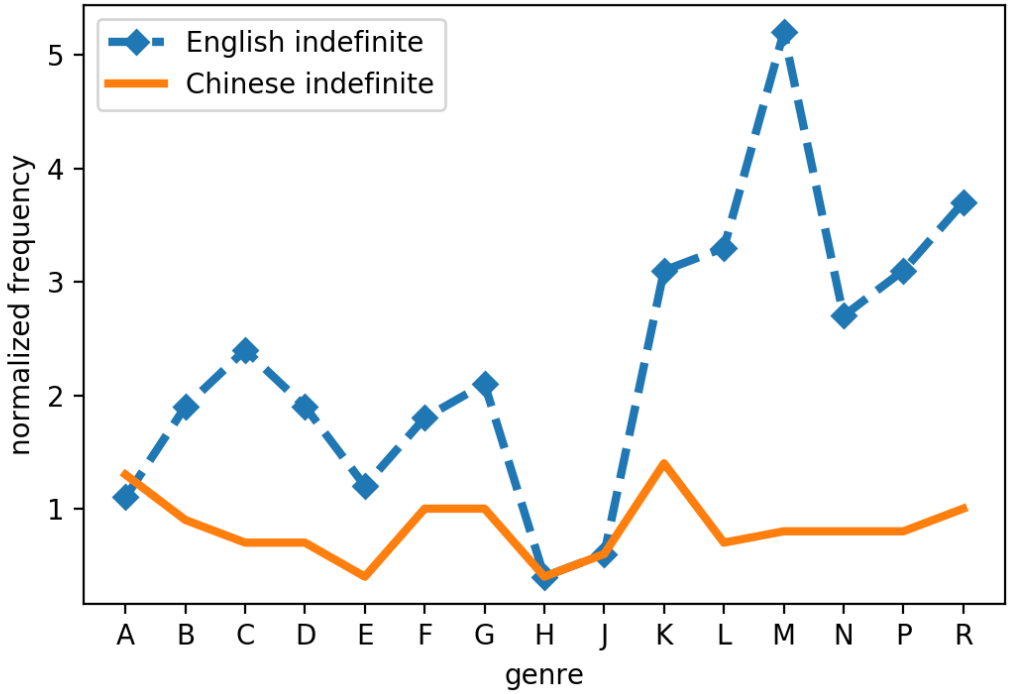

Figure 5. Cross-genre frequencies of indefinite pronouns in English and Chinese

Word (EN)	Percentage (En)	Word (CN)	Percentage (CN)
something	30.17	有些/youxie	23.49
nothing	18.53	大家/dajia	22.95
anything	14.84	一切/yiqie	20.23
someone	13.95	有人/youren	16.09
anyone	8.6	有的/youde	9.98
somewhere	4.41	人人/renren	2.17
anywhere	2.92	每人/meiren	2.04
nobody	2.81	各位/gewei	1.49
nowhere	1.77	个个/gege	0.88
somebody	1.1	大伙/dahuo	0.68
anybody	0.88	None	None

Table 9. Percentage of each type of indefinite pronouns in the two corpora

Indefinite pronouns refer to entities that the writer/speaker cannot or does not want to name more exactly. They are thus characterized by the lack of the element of definiteness in other types of pronouns. There is a fair degree of correlation between English and Chinese indefinite pronouns across genres ($r = 0.569$, $p < 0.05$). The most marked difference between the two languages is seen in Science fiction, where English uses many more indefinite pronouns than any other genres. Negative indefinite pronouns (e.g. nothing, nobody etc.) and universal pronouns (e.g. everyone, everybody) are missing in Chinese because they are treated as noun phrases (meiyou + ren) instead of a pronoun.

The high frequency of the word 大家/dajia is probably due to the fact that apart from its indefinite meaning, it is conflated with personal pronominal function when used just as another way to say *we* or *you*. For example, in (1), dajia can be used to refer to we, you, everybody, depending on the narrative context.

(1) 孙睿是什么样的人，大家心里都清楚。

We/You/Everybody know(s) what kind of person Sun is.Unfortunately, such conflation of function is difficult to disambiguate using automated approaches, and this calls for further research in pronoun classification.

4.2.5 Interrogative Pronouns

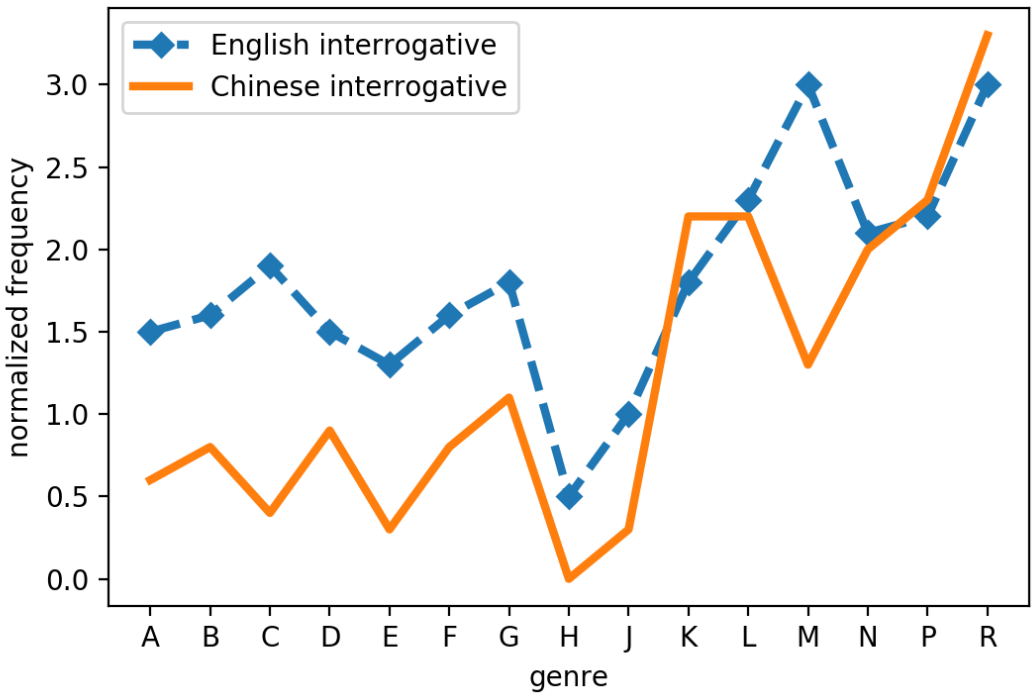

Figure 6. Cross-genre frequencies of interrogative pronouns in English and Chinese

what	86.19	什么/shenme	55.55
who	13.28	谁/shui	17.83
whom	0.48	哪/na	6.48
whose	0.05	哪里/nali	5.39
None	None	何/he	4.7
None	None	哪儿/nar	2.83
None	None	啥/sha	2.48
None	None	哪些/naxie	1.87
None	None	何时/heshi	1.09
None	None	多久/duojiu	1.0
None	None	何处/hechu	0.78

Table 10. Percentage of each type of interrogative pronouns in the two corpora

Interrogative pronouns are used to ask questions. They refer to entities that the speaker wants the addressee to specify. The English interrogative pronouns, apart from their prototypical use in direct questions as shown in (1), also appear in indirect questions (2) in the corpus

(1) **Who** gave you the name?
(2) He knew **what** the deal was, of course.

A comparison of the distribution of interrogative pronouns between English and Chinese paints a more complex picture than the other pronouns. While having a high correlation ($r = 0.719$, $p < 0.001$), neither of the languages has a consistent advantage in frequency across the genres. In A-C (Press), E (Skills, trades and hobbies), H (miscellaneous documents and reports), J (learned reports and writings), interrogative pronouns are found more frequently in Chinese than in English, while in less formal genres, such as K (general fiction), R (humor), the opposite is seen.The distribution of interrogatives in English is polarized. The vast majority of interrogatives are *what*, followed by a significant minority, while the remaining two account for less than one percent of the total count. It is notable that the interrogative pronoun *which* is not found in the corpus. Similarly in the Chinese corpus, albeit less extreme, 什么/shenme/what is the most used pronoun, accounting for more than half of all interrogatives.

Appearing most frequently in both English and Chinese are two words (English what and Chinese shenme) that are considered semantically equivalent in many contexts. Upon closer inspection, however, it is found that the function that these pronouns serve is distinct.

Similar to dajia, the classification of shenme is problematic here since, apart from its normal interrogative use, it also appears as indefinite pronoun with the meaning everything/anything.

(1)才进校，她对什么都还是一片茫然。

A freshman at school, she was at lost about **everything**.

(2)余成功只是淡淡扫了他们一眼，没说什么。

Yu Chenggong just glanced at them briefly without saying **anything**.

A rough estimate of the sampled corpus suggests that about one fourth of shenme should be regarded as indefinite pronouns. Unfortunately, without accurate word sense disambiguation techniques, this ambiguity cannot be resolved automatically.

5. Conclusions

In this study, we adopt a corpus-based approach to the analysis of the distributional patterns of different types of pronouns across different genres in two comparable balanced corpora in English and Chinese.We find considerable variation in the distribution of pronouns in different genres: formal documents have the fewest pronouns overall, fictions have the most, while press reports lie somewhere in the middle. While English tends to employ consistently more pronouns in every genre than Chinese, the distributional patterns of pronouns in the two languages across the genres are highly patterned and significantly correlated with one another, suggesting that pronouns can play similar functional roles in varying contextual situations in the two languages. Of the subtypes of pronouns in the two languages, five are found to be directly comparable. Personal pronouns are found to have the most similar (correlated) genre distribution in the two languages, while demonstrative pronouns share the least similarity. The distributional patterns for each pronoun type are investigated and their underlying functional and cultural implications discussed.

The identification of these classes of pronouns can in large part be automated with the help of state-of-the-art part-of-speech taggers and dependency parsers. There are special cases in automated classification of pronouns due to the inherent ambiguity in some of the words (e.g. Chinese 大家/dajia, 什么/shenme).

The results of this study can inform future research and application involving pronouns, with implications ranging from cross-linguistic studies of grammatical features to second language acquisition.

Sponsoring

This paper was supported by the MOE Project of Key Research Institute of Humanities and Social Sciences at Universities (Grant No. 13JJD740009),by Guangdong Planning Office of Philosophy and Social Science (Grant No. GD18YWW02), andthe Foundation for Distinguished Young Talents in Higher Education of Guangdong (Grant No. 2015WQNCX030), Department of Education of Guangdong Province, P. R. China.

References

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Longman.

Bramley, N. (2001). Pronouns of Politics: the use of pronouns in the construction of'self'and'other'in political interviews. Retrieved from https://openresearch-repository.anu.edu.au/handle/1885/46225

Brown-Schmidt, S., K.Byron, D., & K.Tanenhaus, M. (2005). Beyond salience: Interpretation of personal and demonstrative pronouns. *Journal of Memory and Language*, *53*(2), 292–313. https://doi.org/https://doi.org/10.1016/j.jml.2005.03.003

Campbell, R. S., & Pennebaker, J. W. (2003). The Secret Life of Pronouns. *Psychological Science*, *14*(1), 60–65. https://doi.org/10.1111/1467-9280.01419

Evans, G. (1980). Pronouns. *Linguistic Inquiry*, *11*(2), 337–362.

Gao, X. (2015). A Corpus-based Study of First Person Pronouns of Chinese and English Scientific Writing. *Foreign Language Teaching*, *36*(02), 30–34.

Gundel, J. K., Hedberg, N., & Zacharski., R. (2004). Demonstrative Pronouns in Natural Discourse. In *Proceedings of the 5th Discourse Anaphora and Anaphor Resolution Colloquium* (pp. 81–86). São Miguel, Portugal.

Hartung, F., Burke, M., Hagoort, P., & Willems, R. M. (2016). Taking Perspective: Personal Pronouns Affect Experiential Aspects of Literary Reading. *PLOS ONE*, *11*(5), e0154732. https://doi.org/10.1371/journal.pone.0154732

Hong, B. (1985). Politeness in Chinese: Impersonal Pronouns and Personal Greetings. *Anthropological Linguistics*, *27*(2), 204–213.

Hudson, R. (1994). About 37% of word-tokens are nouns. Language. *Language*, *70*(2), 331–339.

Kim, C.-K. (2009). Personal pronouns in English and Korean texts: A corpus-based study in terms of textual interaction. *Journal of Pragmatics*, *41*(10), 2086–2099.

Li, G., & Liang, M. (1999). Contrastive Analysis of English-Chinese Pronouns and Computer Assisted Translation, *6*, 22–24.

Lv, S. (1999). *Xiandai Hanyu Babai Ci*. Commercial Press.

Meissner, W. W. (2008). The role of language in the development of the self III: The significance of pronouns. *Psychoanalytic Psychology*, *25*(2), 242–256. https://doi.org/10.1037/0736-9735.25.2.242

Pennebaker, J. W. (2011). *The secret life of pronouns : what our words say about us*. Bloomsbury Press.

Quirk, R., Greenbaum, S., & Leech, G. (1985). *A Comprehensive Grammar of the English Language*. London and New York: Longman.

Thonney, T. (2013). "At first I thought… but I don't know for sure": The Use of First Person Pronouns in the Academic Writing of Novices. *Across the Disciplines*, *10*(1).

Tsao, F. (1982). English and Chinese (Mandarin). *Annual Review of Applied Linguistics*, *3*, 99–117. https://doi.org/10.1017/S0267190500000659

Xu, J., & Liang, M. (2012). A tale of two C's : Comparing English varieties with Crown and CLOB (the 2009 Brown family corpora). *ICAME Journal*, *37*(June), 175–184.

Zhang, M. (2008). A Comparative Study of First Person Pronouns in Abstracts in China and English Speaking Countries. *Shanghai Journal of Translators, 2*, 31–36.