Computational Linguistics and Chinese Language Processing
Vol. 26, No. 1, June 2021, pp. 43-78

Textual Relations with Conjunctive Adverbials in English Writing by Chinese Speakers: A corpus-based Approach

Tung-Yu Kao*[*] and Li-Mei Chen*[*]

Abstract

The study aims to investigate the use of conjunctive adverbials (CA, hereafter) performing various textual relations in the English writing by Chinese speakers across genres and over time. To begin with, a corpus of one million word was compiled and the corpus interface was constructed. Later, 45 pieces of writing by 5 college students during 4 semesters were selected for data annotation and analysis, with each student contributing 9 pieces for 9 text genres. The results show that there exists a distribution norm of CA-performed textual relations based on CA occurrence frequency and that the distribution is independent of genre and time effects. Compared with literature, the found distribution is also considered free from the first language influence. This suggests that the found distribution is a mental representation of mature human cognition, underlying English writing on global and coherent levels. Therefore, the found distribution is of great potential for developing automatic tools of discourse diagnosis.

Keywords: Conjunctive Adverbial, Textual Relation, Text Genre, English Writing, Corpus Compilation, Automatic Discourse Diagnosis.

1. Introduction

Implemented in August 2019, Curriculum Guidelines of 12-Year Basic Education (MOE, 2014) has ushered in a new era of the English education in Taiwan. In the past, the English education in school was under severe criticism for overemphasizing sentential grammaticality and understating discoursal and pragmatic aspects. The concern has received attention and been addressed. In the spirits of the curriculum reform, the Guidelines for the English education (MOE, 2018) specifies that students should be taught to identify text genres through

[*] Department of Foreign Languages and Literature, National Cheng Kung University
E-mail: dodofishletter@hotmail.com; leemay@mail.ncku.edu.tw
The author for correspondence is Li-Mei Chen.

text features, pay attention to cohesion and coherence across sentences, and employ reading strategies, such as skimming and inferring, to comprehend the text as a whole. In other words, the focus of the English education in Taiwan evolves to direct students to the importance of textual relations on the discourse level.

In fact, in higher education, the emphasis of textual relations in classes for English writing had been well addressed. It is found that most writing textbooks (Connelly, 2013; Langan, 2010; Lannon, 2007; Morenberg & Sommer, 2008; Reid, 2000; Smalley *et al*., 2012; Wyrick, 2008) introduce various genres by indicating that some conjunctive adverbials (CAs, hereafter) are more prominent in certain genres than other CAs based on the textual relations they perform. For instance, CAs, such as *firstly*, *next*, and *in addition*, are thought to appear more in the process genre to show progressive relations across sentences or paraphrases. All things considered, learning to recognize and employ CAs that signal textual relations in genres can be said to become a highlighted component throughout the English education in Taiwan.

Also, to respond to the enormous demand of English writing by native speakers (NS, hereafter) and by non-native speakers (NNS, hereafter), writing-assisting online platforms and software packages, such as *My Access!*, *Grammarly*, *StyleWriter*, and *WhiteSmoke*, have been developed for the purpose of automatically facilitating and advising writers to compose better. As seen in Table 1, *My Access!* is a process-oriented writing platform where people are guided through scaffolds, like brainstorming charts and revision requests, for contents development, and provided with human-graded scores and feedback as reference. The other three are product-checking writing software to proofread writing forms, including punctuation, spelling, and sentential grammaticality, for correctness, and to offer possible alternatives for more concise sentence rephrasing and better word choice.

Table 1. Functions offered by the four writing-assisting software.

Functions Tools	Form Checker	Style Options	Writing Scaffold	Score & Feedback	Unique Feature
My Access! [1]	+	–	+	+	Previous writing pieces analyzed based on grammar mistakes
Grammarly[2]	+	+	–	–	Different options provided based on desired tones
StyleWriter[3]	+	+	–	–	Statistics, such as word count and sentence length, provided
WhiteSmoke[4]	+	+	–	–	Full-text and word-to-word translation available

Note: [1]https://www.myaccess.com/myaccess/do/log
[2]https://www.grammarly.com/
[3]https://www.editorsoftware.com/stylewriter.html
[4]https://www.whitesmoke.com/

With these textual foci in the English education in Taiwan and technology advances in the English writing, however, few empirical studies has been conducted after Chen's (2006) seminal work to examine whether the difference in the occurrence of CAs across different genres is on a significant level, and existing writing tools are still inadequate to automatically diagnose a writing text and generate comments on the discourse level. In a preliminary database search on Linguistics and Language Behavior Abstracts (LLBA), the results for peer-reviewed articles after 2007 with *conjunctive adverbial* and *writing* in the abstract produced 92 entries. On closer inspection, none investigated various textual relations across different genres. Although there were studies explored the use of CAs, the CA items were pre-selected and belonged to the same textual relation, such as Phoocharoensil (2017) exploring resultive CAs *THUS*, *THEREFORE*, *HENCE*, and *SO* in written academic English. In another database search on Academic Search Complete (EBSCOhost), the results for peer-reviewed articles after 2007 with *automatic*, *evaluation*, and *discourse* in the abstract produced 88 entries. On closer inspection, the only study dedicated to computational text-level discourse analysis is Morey, Muller, and Asher' (2018) study. Yet, their study was based on Rhetorical Structure Theory (RST), rather than textual relations performed by CAs.

Therefore, the present study aims to take a corpus approach to investigate (1) whether the distribution of CA-performed textual relations in the English writing by Chinese speakers is significantly subject to text genres, and (2) whether it significantly varies over time with learning and practicing. Ultimately, the data collected and the results are hoped to serve as training data and calculating principles for developing automated discourse-evaluating application.

2. Literature Review

This section consists of two parts, with the first part focusing on the working definition of CAs, short for conjunctive adverbials, employed in the present study and the other reviewing previous studies on their distribution in the English writing by native and non-native speakers.

2.1 Working Definition of CAs (Conjunctive Adverbials)

In Halliday and Hasan's (1976) cohesion framework, CAs are one type of cohesion to achieve textual coherence by which sentences are grouped together and considered an integrated discourse unit. The type of cohesion differs from other types in the way it functions to make connections among sentences. CAs relate sentences by providing one possible interpretation to confine the effect of sentences on one another, rather than using anaphoric relation to ensure the involvement of the same topic in sentences.

For example, the two sentences in (1) are regarded as one unit for the pronoun in the second sentence refers back to the subject in the first sentence and establishes a link between the two sentences. Unlike both sentences in (1) staying with the topic of the person, sentences in (2) deal with different topics but are still viewed as a unit because the CA, *however*, offers one kind of textual relation to denote how the two propositions in (2) are related to each other. From the discussion, CA, therefore, can be defined as one text-creating mechanism indicating the inter-sentential textual relation, and accordingly exclude the discussion of coordinators or subordinators, which signal the intra-sentential textual relation.

(1) *Barack Obama* was inaugurated as the President of the United States on January 20, 2009. *He* is the first African American president in the history of the country.

(2) May is the plum rains season in Taiwan. *However*, the rainfall this year reaches a historic minimum.

Nevertheless, the definition is not exclusive enough. As seen in (3a) and (4a), both *however* and *later* indicate how the second sentence is related to the first sentence in respective examples, with the former yielding a contrastive effect and the latter designating the temporal order. Yet, not both are considered CAs. According to Quirk, Greenbaum, Leech, and Svartvik (1985), one common feature shared by CAs, sentence adverbials in Quirk *et al*'s term, is that the type of cohesion cannot occupy the focus of a cleft sentence. In this sense, after tested in (3b) and (4b), *however* remains a CA while *later* would be excluded from the scope of the present study.

(3) a. When the Tae Kwon Do contestant Li-wen Su sprained her knee in Olympics, people thought she would quit the contest. *However*, she continued fighting to the end.

b. *It is however that she continued fighting to the end.

(4) a. Landing on the moon was first ridiculed as an impossible mission. *Later*, people realized this could really work.

b. It is later that people realized this could really work.

Given the semantic and syntactic criteria for defining CAs, an additional criterion is employed in the present study. According to Halliday and Hasan (1976), CAs fall into three kinds of language form: adverbs, prepositional phrases, and prepositional expressions with reference items, as presented in (5a) to (5c), respectively. It is clear that (5a) and (5b) are linked to the first sentence because of the adverb in (5a) and the prepositional phrase in (5b) specifying textual relations between sentences. However, the link between (5c) and the first sentence is established based mainly on the presence of the reference in the prepositional expression. The language form that Halliday and Hasan also regard as CAs works more like lexical cohesions than CAs. Therefore, the third criterion supplemented in the present study is that a CA must be lexicalized and self-contained. In other words, the present study only investigates CAs in the form of adverbs and prepositional phrases.

(5) The captain had steered a course close in to the shore.

a. *Therefore*, they avoided the worst of the storm.

b. *As a result*, they avoided the worst of the storm.

c. *As a result of this*, they avoided the worst of the storm.

Ultimately, the working definition of CAs in the present study is as follows, and Table 2 shows how the three criteria delimit the investigating scope in the present study.

Criterion 1: A conjunctive adverbial must semantically indicate the relation between the sentences before and after it.

Criterion 2: A conjunctive adverbial must be syntactically forbidden to be the focus in a cleft sentence.

Criterion 3: A conjunctive adverbial must be lexicalized and self-contained.

Table 2. Working definition delimiting the investigated CAs.

The CA forms in Halliday and Hasan (1976)	Examples	Tested based on the three criteria	To be examined in the present study
Adverbs	therefore	Satisfying the criteria	Yes
Prepositional phrases	as a result	Satisfying the criteria	Yes
Prepositional expressions with reference items	as a result of that	Violating criterion 3	No

2.2 Previous Studies on Textual Relations Performed by CAs

The significance of CAs lies in the fact that they direct the interpretation among sentences in text, which leads to the attempt to classify textual relations explicitly indicated by conjunctive adverbials. According to Halliday and Hasan (1976), *Additive*, *Adversative*, *Causal*, and *Temporal* were the four types of textual relations regulated, with various subdivisions in each type. Later, the various subdivisions were collapsed, and the taxonomy was simplified by Celce-Murcia & Larsen-Freeman (1999).

Table 3. CA taxonomy in the literature.

Systems	Researchers	Types
Four-type classifying system	Halliday & Hasan (1976)	Additive, Adversative, Causal, Temporal
	Celce-Murcia & Larsen-Freeman (1999)	Additive, Adversative, Causal, Temporal
More-type classifying system	Quirk *et al*. (1985)	Listing, Summative, Appositional, Resultive, Inferential, Contrastive, Transitional
	Biber *et al*. (1999)	Enumeration, Addition, Summative, Appositional, Result/Inference, Contrast/Concession, Transitional

Compared with Celce-Murcia and Larsen-Freeman collapsing subdivisions, Quirk *et al*. (1985) revised Halliday and Hasan's four-type system as a system of seven types, namely, *Listing*, *Summative*, *Appositional*, *Resultive*, *Inferential*, *Contrastive*, and *Transitional*. Similarly, Biber *et al*. (1999) developed their own classifying version, which was very much the same with Quirk *et al*.'s except separating *Listing* in Quirk *et al.* as *Enumeration* and *Addition*, changing *Contrastive* as *Contrast/Concession*, and combing *Resultive* and *Inferential* into *Result/Inference*. Table 3 summarizes CA taxonomy based on the four-type and the more-type classifying systems.

With the four-type and the more-type classifying systems, empirical studies involving investigation into the distribution of CA-performed textual relations in the English writing are reviewed. Table 4 and 5 present studies that employed the four-type framework of CAs for analysis, with the former based on the English writing by NS and the latter based on the English writing by NNS. As shown in Table 4, the distribution patterns of CA-performed textual relations in Field and Yip (1992) and in Chen (2006) are identical. CAs of *Adversative* occur the most frequently, followed by *Additive*, *Causal*, and *Temporal* in a descending order. This may suggest that there exists a distribution norm in NS cognition, and that the norm is independent of genre and time influences.

In contrast, the writing by NNS does not present such a distribution norm of textual relations carried out by CAs. As seen in Table 5, the five studies present four distribution patterns; in other words, the results in these studies differ from one another. In addition, genre and time influences fail to account for the lack of distribution pattern consistency. Consider genre influence. Field and Yip (1992) as well as Liu and Braine (2005) investigated the same writing genre, and so did Xie (2014) and Huang (2018). Yet, neither sets of studies found the same distribution pattern. Now, consider time influence. Field and Yip (1992) as well as Xie (2014) collected data from the same time period, and so did Liu and Braine (2005) as well as Huang (2018). Again, neither sets of studies found the same distribution pattern.

Table 4. Distribution of textual relations in the NS writing based on four-type framework.

Researchers	Field & Yip (1992)	Chen (2006)
Framework	H & H (1976)[1]	C & L (1999)[2]
Genre	Argumentation	Research articles
Data Source	High school students in Sydney	Journal articles on TESOL
Distribution of Textual Relations	Adversative >[3] Additive > Causal > Temporal	Adversative > Additive > Causal > Temporal

Not e: [1] Halliday & Hasan (1976)
[2] Celce-Murcia & Larsen-Freeman (1999), the simplified framework of Halliday and Hasan
[3] The symbol ">" means "more occurring frequencies than."

To account for the lack of distribution pattern consistency, first language background is suggested to be the cause. After comparison, it is found that the distribution pattern exhibited in the NNS writing in Field and Yip (1992) is the same as that in the NS writing shown both in Field and Yip (1992) and in Chen (2006). The consistency may be explained by the fact that the so-called NNS group in Field and Yip (1992) should be considered Chinese-English bilingual natives. They lived in once-UK-colonized Hong Kong and were immersed in the English-speaking environment growing up. Consequently, they may share the same distribution pattern with NS in cognition in terms of English writing. The language background also account for why the NNS writing in Liu and Braine (2005) exhibits the same distribution pattern as that in Chen (2006) since Mandarin is the first language for both data sources.

Table 5. Distribution of textual relations in the NNS writing based on four-type framework.

Researchers	Field & Yip (1992)	Liu & Braine (2005)	Chen (2006)	Xie (2014)	Huang (2018)
Framework	H & H (1976)	H & H (1976)	C & L (1999)	H & H (1976)	H & H (1976)
Genre	Argumentation	Argumentation	Various kinds	Exposition	Exposition
Data Source	High school students in Hong Kong	College students in Beijing	MA TESOL students in Taiwan	High school students in Taiwan	College students in Taiwan
Distribution of Textual Relations	Adversative> Additive> Causal> Temporal	Additive> Causal> Temporal> Adversative	Additive> Causal> Temporal> Adversative	Additive> Temporal> Causal> Adversative	Temporal> Additive> Adversative> Causal

However, for the four studies with data sources taking Chinese as the mother tongue, the distribution patterns in Xie (2014) and in Huang (2018) still differ from the pattern exhibited in Liu and Braine (2005) and in Chen (2006). The difference may lie in writing length. The length of one piece of writing collected in Xie (2014) is 120 to 170 words, and that in Huang (2018) is 150 to 200 words. Since the writing collected in Chen (2006) is research-related, including literature reviews, research proposals, and pedagogical "how-to" papers, the average length is much longer, between 3000 and 4000 words. Given the above observation and discussion, it might be inferred that while time and genre differences do not impact the distribution pattern of textual relations, first language and writing length might have a role in affecting it.

Table 6. Distribution of textual relations based on more-type framework.

Researchers	Tankó (2004)		Altenberg & Tapper (1998)		Shen (2006)	
Framework	Quirk *et al*. (1985)		Quirk *et al*. (1985)		Biber *et al*. (1999)	
Genre	Argumentation		Argumentation		Research articles	
Data Source	NS: No data	NNS: Hungarian college students	NS: College students	NNS: Swedish college students	NS: Journal articles on TESOL	NNS: Conference papers by Taiwanese
Distribution of Textual Relations		Listing> Resultive> Contrastive> Summative> Appositive> Inferential> Transitional> Corroborative (0 case)	Contrastive> Resultive> Listing> Appositive> *Corroborative*> Summative> Transitional (0) Inferential (0)	Contrastive> Resultive> Appositive> Listing> Corroborative> Summative> Transitional> Inferential (0)	Contrastive> Appositive> *Result/Inference*> Listing*> Corroborative> Transitional> Summative	Listing> Contrastive> *Result/Inference*> Appositive> Summative> Transitional> Corroborative

*The *Additive* and *Enumerative* textual relations in Shen (2006) were collapsed into *Listing* in comparison with the other studies.

Tankó (2004), Altenberg and Tapper (1998), and Shen (2006) are the studies taking the more-type system as the framework for analysis. Although the former two studies were based on Quirk *et al*.'s (1985) classification and the last was on Biber et al.'s (1999), the frameworks they employed were much the same. The only difference is the use of category names without substantial contents changes. For example, *Additive* and *Enumerative* in Shen (2006) could be collapsed and equate *Listing* in the other two studies. Moreover, the three studies all referred to Granger and Tyson's (1996) classification and designated one more textual relation, *Corroborative*, conveying writers' attitudes toward and comments on the text, in their frameworks. In terms of text genre and time period, Tankó (2004) as well as Altenberg and Tapper (1998) limited their data to the argumentative writing by college students with the former taking the latter's NS data as benchmark, while the data in Shen (2006) were research papers from academic journals as benchmark and conference papers by Taiwanese postgraduates as the NNS samples. The research design and results of the three studies are summarized in Table 6.

Based on Table 6, despite genre, age, and first language differences in data sources, a distribution norm of CA-performed textual relations in English writing can be identified, when textual relations are considered in groups. In the distribution norm, *Contrastive*, *Resultive*, *Listing* occur most often, *Appositive* and *Summative* ranks moderate in the order, and *Transitive* and *Inferential* are seldom used.

From the observed distribution norm, three points are induced as well. Firstly, fine classification of textual relations might be able to clarify nuances in occurrence frequency better than generic classification, and to manifest the underlying distribution norm. The inference is made due to the difference in the observed results based on the four-type and more-type frameworks. While the distribution patterns of CA-performed textual relations based on both frameworks might be not susceptible to text genre and time period, first language might have an influence on the distribution in the former framework but not in the latter. Secondly, it makes sense that Biber *et al*. (1999) combined *Resultive* and *Inferential* in Quirk *et al*.'s (1985) classification for the inferential relation has the lowest occurring ratio, e.g., the zero occurrence in Altenberg and Tapper (1998). Lastly, it is found that NS use the *Corroborative* relation in their writing more frequently than NNS do. This might originate in the fact that writers would exhibit a higher level of authority when writing in their first languages than in other languages (Chen, 2006). With corroborative adverbials serving to express writers' opinions, the kind of conjunctive device, therefore, is used more often in the NS writing for establishing authority.

3. Methodology

The present study takes a corpus-based approach to explore the distribution of CA-performed textual relations in the English writing by Chinese speakers across genres and over time. For the research goal, the section first reports how the corpus in present study was compiled, and then elaborates how the selected data were annotated with the coding scheme through the coding procedure. The section is wrapped up with an introduction to statistical measures for data analysis.

3.1 Corpus Compilation

A corpus-based approach was employed in the present study. The corpus compilation was based on the first four stages in Atkins, Clear, and Ostler’s (1992) corpus building, which are specifications and design, hardware and software, data capture and mark-up, as well as corpus processing.

In the Specifications and Design stage, the corpus formation in the present study was designed based on the OLAC Metadata Set. OLAC Metadata Set is an exclusive protocol framed by the Open Language Archives Community, regulating the information for digitally archiving language resources and basing its digital storage on the XML format (Simons & Bird, 2008). It was XML format’s extensible features that allowed the present study to tailor its own format for the research purpose.

Three kinds of specification were formatted in the present study. They are *Informant Background*, *Article Message*, and *Text Annotation*. *Informant Background* offers a basis for possible research directions, while *Article Message* helps select suitable materials for research analysis. Table 7 lists the complete specifications for *Informant Background* and *Article Message*.

Table 7. Specifications for Informant Background and Article Message.

Informant Background		*Article Message*	
Aspects	Specifications	Aspects	Specifications
Basic Information	Account	Author	Account
	Chinese name		Chinese name
	English name		English name
	Gender	Attribute	Academic year
	Age		Genre
Education	Vocation / Speciality		Draft
	Level of education		Title
	University		Word count
	Department	Text	Outline
Language Use	Mother tongue		English abstract
	Known languages		Chinese abstract
	Learning experience		Text body
		Revising Process	Teacher’s feedback
			Author’s response

Unlike the first two formations dealing with the sources of text, the third kind of formation, Text Annotation, copes with text itself and preserves the linguistic information annotating the text it is attached to. As illustrated in the following instance, the CA is tagged within a pair of pointed brackets, and the metalinguistic coding information is annotated in the first pointed brackets.

> ＜tag **Y Enu CA annotation=" "**＞First＜/tag＞, children who have nasal allergy always have some mental problems to some extent.

In the Hardware and Software stage, the programming language Perl was chosen to develop the corpus interface because Perl is well known for text processing, such as dealing with files, strings, and regular expressions (Suehring, 2006). The construction of the corpus interface consisted of two phases, which were requirements analysis and system implementation. The former analyzed functions the interface should offer to serve the purpose of the present study, whereas the latter used program modules to assemble the required functions.

Based on requirements analysis, researchers, students and teachers were the three identities involved. Figure 1 visualizes the layout of the interface, where functions required by different identities are specified. Note that the block highlights the functions directly related to the purpose of the present study, and only the functions in the block are further depicted in Table 8.

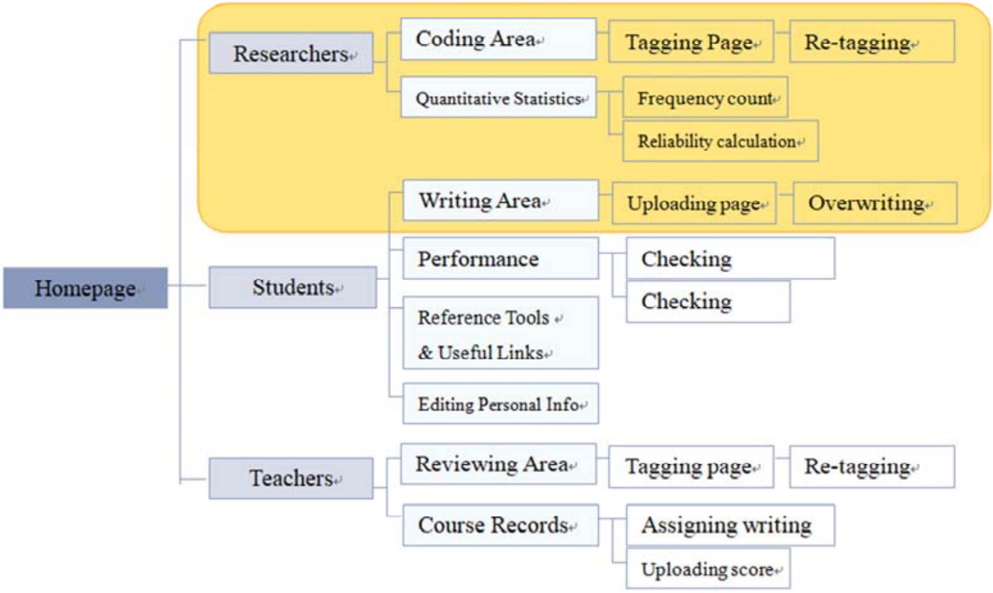


Figure 1. Interface layout.

Table 8. Description of each function.

Interface Function	Description
Tagging Page	Annotate the data with meta-information
Re-tagging	Retrieve annotated data for reviewing and revising earlier annotation
Frequency Count	Search the corpus by an analyzing code or a specific word, show all the matched cases and tally the total occurring frequency.
Reliability Calculation	Present two researchers' annotations of the same text in parallel and calculate all the combination situations of agreement and disagreement.
Uploading Page	Upload compositions.
Overwriting	Retrieve uploaded compositions for overwriting earlier drafts.

Following the requirements analysis was the system implementation of the interface. Table 9 presents and defines eleven modules that help execute functions required.

Table 9. Execution of each program module.

Program Module	Abbreviation	Execution
Highlight	High	Distinguish annotated information from raw data
Input	In	Receive the input data
Hash	Hash	Calculate agreement frequency of coders' annotation
Match	Mat	Compare annotations of coders based on units
Output	Out	Present the retrieved data on screen
Import	Imp	Import the enquired data from the corpus
Save	Sav	Save the uploaded data in the corpus
Filter	Fil	Sift data entries that match the search instruct
Login	Log	Secure the legitimacy of users
Split	Spl	Break down the text into units (sentences or words)
Tagger	Tag	Annotate the raw data with meta-information

Given the eleven program modules, Figure 2 visualizes how each function is modulized. In the figure, squares and cambers, respectively, represent the desired functions and the assembling modules. The correspondence between abbreviations and modules is presented in Table 9.

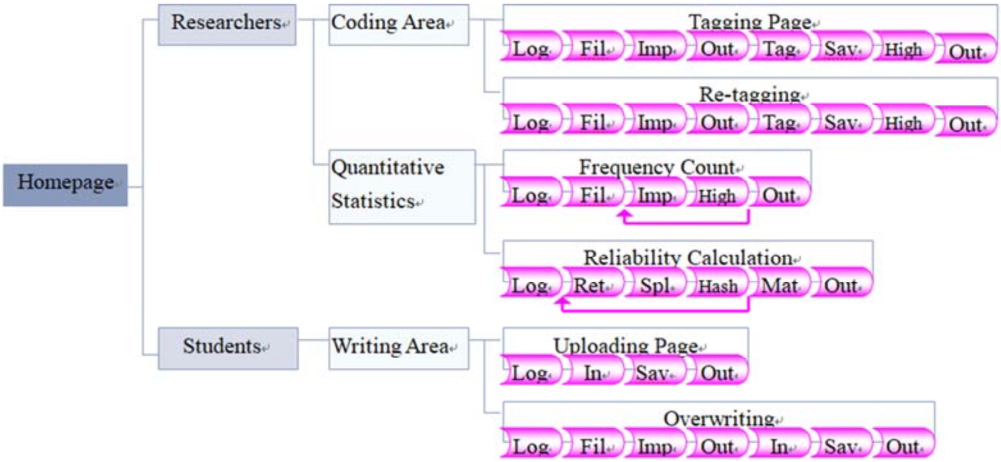


Figure 2. Assembly of modules.

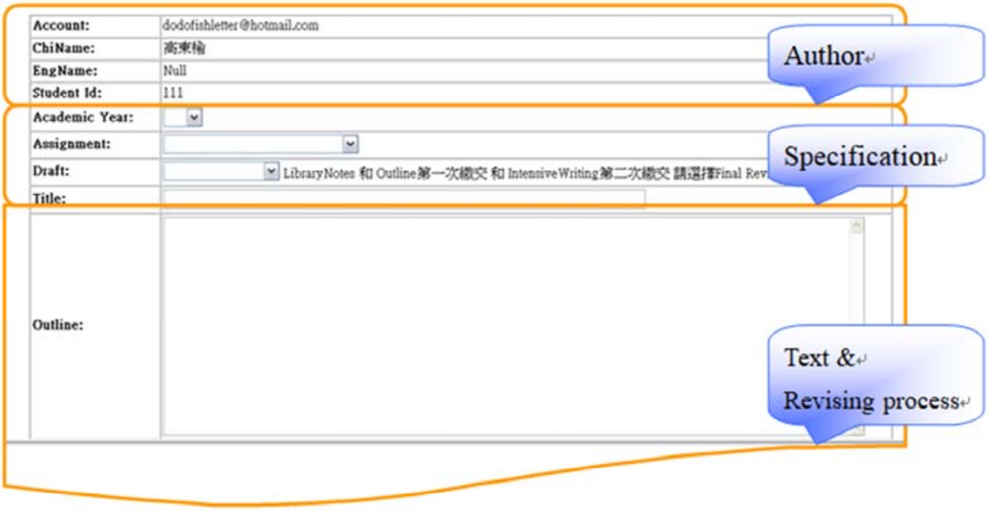
To this point, the construction of the corpus interface was completed. Table 10 demonstrates how the eleven program modules assemble all the functions.

Table 10. Module Assemblage of Functions.

Function Module	*Tagging page*	*Re-tagging*	*Frequency count*	*Reliability calculation*	*Uploading page*	*Over-writing*
Highlight	+	+	+			
Input					+	+
Match				+		
Hash				+		
Output	+	+	+	+	+	+
Import	+	+	+	+		+
Save	+	+			+	+
Filter	+	+	+			+
Login	+	+	+	+	+	+
Split				+		
Tagger	+	+				

With the corpus interface was constructed, the Data Capture and Mark-up stage ensued. Students in the Department of Foreign Languages & Literature at National Cheng Kung University in Taiwan agreed to participate in the present study, contributing their writing pieces to compile the corpus of English academic writing by Chinese NS. The genres of all the writing pieces collected belonged in 13 types, including *process*, *summary*, *essay question writing*, *cause-effect*, *comparison-contrast*, *definition*, *description*, *narration*, *classification*, *multiple strategies*, *argumentation*, *problem solving*, and *research article*. For each genre, each student wrote 3 or 4 pieces of composition, which could be independent of one another or revised drafts for the prior draft.

Figure 3 presents the *Uploading page* where students upload their writing pieces as raw data. While the system would automatically import information on *Author*, students needed to manually select labels for *Attribute*, and typed their writing in the text-editing area. Once students clicked the bottom Save and all the information would be automatically marked up and converted into machine-readable text.

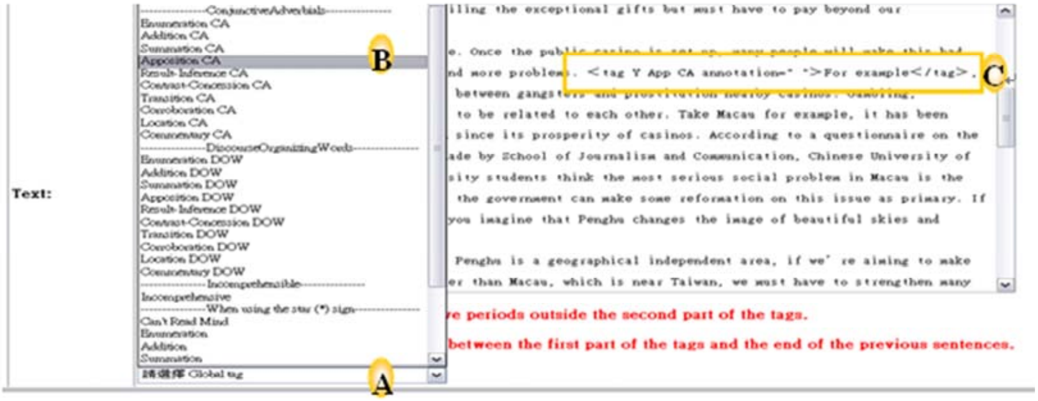


Account:	dodofishletter@hotmail.com
ChiName:	高東榆
EngName:	Null
Student Id:	111
Academic Year:	
Assignment:	
Draft:	LibraryNotes 和 Outline第一次繳交 和 IntensiveWriting第二次繳交 請選擇Final Rev
Title:	
Outline:	

Figure 3. Uploading page.

After collecting the raw data came the Corpus Processing stage. Two most relevant data-processing functions provided on the corpus interface are *Tagging page* and *Frequency count*.

Figure 4 presents Tagging page, where researchers proceed to analyze the data. Part A is the tagger. Part B shows the coding scheme to attach to the language form in text. It is worth mentioning that the coding scheme shown on the tagger is replaceable, and can be changed according to different research purposes. Part C demonstrates the annotated text after tagging.



Text:

--------------ConjunctiveAdverbials--------------
Enumeration CA
Addition CA
Summation CA
Apposition CA
Result-Inference CA
Contrast-Concession CA
Transition CA
Corroboration CA
Location CA
Commentary CA
--------------DiscourseOrganizingWords--------------
Enumeration DOW
Addition DOW
Summation DOW
Apposition DOW
Result-Inference DOW
Contrast-Concession DOW
Transition DOW
Corroboration DOW
Location DOW
Commentary DOW
--------------Incomprehensible--------------
Incomprehensive
--------------When using the star (*) sign--------------
Can't Read Mind
Enumeration
Addition
Summation
請選擇 Global tag

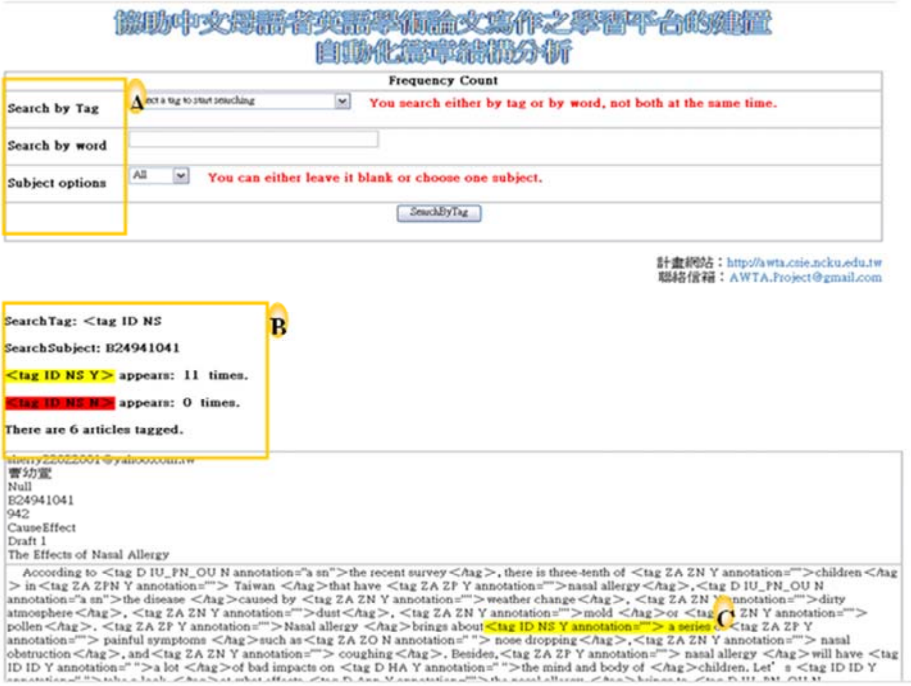
```
illing the exceptional gifts but must have to pay beyond our

e. Once the public casino is set up, many people will make this bad
nd more problems. <tag Y App CA annotation=" ">For example</tag>.
 between gangsters and prostitution nearby casinos. Gambling,
 to be related to each other. Take Macau for example, it has been
 since its prosperity of casinos. According to a questionnaire on the
ade by School of Journalism and Communication, Chinese University of
sity students think the most serious social problem in Macau is the
 the government can make some reformation on this issue as primary. If
you imagine that Penghu changes the image of beautiful skies and

 Penghu is a geographical independent area, if we' re aiming to make
er than Macau, which is near Taiwan, we must have to strengthen many
```

re periods outside the second part of the tags.

between the first part of the tags and the end of the previous sentences.

Figure 4. Tagging page.

Frequency count helps researchers obtain descriptive statistics for study results, as shown in Figure 5. Part A provides two ways to get the statistics. The frequency count can be either based on a specific tag or based on a specific word. *Subject options* aims to limit the search scope. The default search scope is the whole corpus, but researchers could narrow the search by clicking the column. Part B reports the frequency count, whereas Part C presents and highlights all the data matching the search instruction.



協助中文母語者英語學術論文寫作之學習平台的建置
自動化篇章結構分析

Frequency Count

Search by Tag	...ect a tag to start searching You search either by tag or by word, not both at the same time.
Search by word	
Subject options	All You can either leave it blank or choose one subject.

計畫網站：http://awta.csie.ncku.edu.tw
聯絡信箱：AWTA.Project@gmail.com

SearchTag: <tag ID NS

SearchSubject: B24941041

<tag ID NS Y> appears: 11 times.

<tag ID NS N> appears: 0 times.

There are 6 articles tagged.

sherry22022001@yahoo.com.tw
曹幼萱
Null
B24941041
942
CauseEffect
Draft 1
The Effects of Nasal Allergy

According to <tag D IU_PN_OU N annotation="a sn">the recent survey</tag>, there is three-tenth of <tag ZA ZN Y annotation="">children</tag> in<tag ZA ZPN Y annotation=""> Taiwan </tag>that have <tag ZA ZP Y annotation="">nasal allergy</tag>,<tag D IU_PN_OU N annotation="a sn">the disease </tag>caused by <tag ZA ZN Y annotation="">weather change</tag>, <tag ZA ZN Y annotation="">dirty atmosphere</tag>, <tag ZA ZN Y annotation="">dust</tag>, <tag ZA ZN Y annotation="">mold </tag>or <tag ZA ZN Y annotation=""> pollen</tag>. <tag ZA ZP Y annotation="">Nasal allergy </tag>brings about<tag ID NS Y annotation=""> a series of<tag ZA ZP Y annotation=""> painful symptoms </tag>such as<tag ZA ZO N annotation=" "> nose dropping</tag>,<tag ZA ZN Y annotation=""> nasal obstruction</tag>, and<tag ZA ZN Y annotation=""> coughing</tag>. Besides,<tag ZA ZP Y annotation=""> nasal allergy </tag>will have <tag ID ID Y annotation=" ">a lot </tag>of bad impacts on <tag D HA Y annotation=" ">the mind and body of </tag>children. Let' s <tag ID ID Y

Figure 5. Frequency count.

3.2 Data Selection

After three years of data collection, the compiled corpus reached one million words. It consists of 2290 pieces of English compositions by Chinese speakers, belonging to 13 different genres. The total word count is 1429397 words.

Since the present study was targeted at exploring whether text genre and time period play a role in the distribution of textual relations manifested by CAs in the English writing by Chinese speakers, it was better that data provided by students covered all the genres and time periods for the purpose of minimizing individual differences and various writing instructions they received during different time periods. As a result, 45 writing pieces by 5 college students were selected to annotate and analyze. Each student contributed 9 writing pieces for 9 genres, with one piece dedicated to one genre, over 4 semesters. Table 11 shows nine genres collected, expected writing length, data word count, data sentence count, and the time period when the piece was written.

Table 11. Selected data for analysis in the present study.

Genre	Abbre-viation	Piece number	Length of a piece	Total word number	Total sentence number	Semester for data collection
Comparison-contrast	Com-Con	5	450-500	2380	114	2nd
Cause-effect	Cau-Eff	5	450-500	2494	132	2nd
Description	Des	5	450-500	2206	117	3rd
Definition	Def	5	450-500	2512	123	3rd
Narration	Nar	5	450-500	2742	177	3rd
Classification	Cla	5	700-750	3474	173	4th
Multiple-strategies	Mul-Str	5	700-750	3595	168	4th
Argumentation	Arg	5	900-1000	4682	213	5th
Problem-solving	Pro-Sol	5	900-1000	4857	232	5th

3.3 Coding Scheme and Coding Procedure

A coding scheme was developed to annotate the selected data with linguistic information. The classification of textual relations adopted in the present study followed Quirk *et al.*'s (1985) taxonomy with modification. In Quirk *et al.*'s taxonomy, there were seven types distinguished, including *Listing*, *Transitional*, *Appositive*, *Summative*, *Resultive*, *Inferential*, and *Contrastive*, whereas the present study collapsed *Resultive* and *Inferential* into one class as well as supplemented one additional semantic class, *Corroborative*. As a result, in the present study, textual relations indicated by the CAs encompasses seven types, which are *Listing*, *Transitional*, *Appositive*, *Summative*, *Resultive/Inferential*, *Contrastive*, and *Corroborative*. Table 12 lists all the textual relations with their definitions and the possible language items performing these relations.

Apart from presenting the textual relations, Table 12 also shows that one language item may serve more than one textual relation. For example, the language item *then* may perform either the *Listing* relation or *Resultive/Inferential*. In other words, the semantic coding must depend on the relation performed by the CA, not on certain fixed language items.

Table 12. Textual relations and their definitions.

Textual relation	Definition	Example
Listing	Mark the next unit of discourse with or without relative priority or temporal sequence.	first, moreover, then, in addition
Transitional	Serve to shift attention to another topic that does not follow directly from the preceding event.	meanwhile, in the meantime, now
Appositive	Provide an example or an equivalent of the preceding text.	in other words, for example
Summative	Conclude or sum up the information in the preceding discourse.	in conclusion, to summarize
Resultive/ Inferential	Mark the second part of the discourse as the result or consequence of the preceding discourse.	accordingly, then, as a result, so
Contrastive	Show incompatibility between information.	however, on the contrary, anyhow
Corroborative	Express writers' attitudes toward and comments on the text.	in fact, of course, actually

Note: The classification is based on Quirk *et al.*'s (1985) taxonomy with modification.

Another issue regarding the annotation of the textual relations is register use. Take the CA *besides* as example. While *besides* performs the *Additive* relation between sentences, it is considered spoken register and should be avoided in formal writing. Therefore, the coding scheme designed also takes register differences into account, and marks CAs as *Written* or *Spoken*. Later, to make the coding scheme applicable to the computerized interface, the scheme needs converting into the *Text Annotation* format as mentioned previously.

Table 13 presents the complete coding scheme and its electronic format. With CAs enclosed in the two pairs of pointed brackets, the word *tag* signals the beginning of a piece of annotated linguistic information while /tag the end. As for a piece of the annotated linguistic information, it constitutes three layers, separated by space. The layers specify register difference, textual relation, and supplementary annotation if necessary.

Table 13. The Complete Coding Scheme.

Textual relation		Eletronic Format	
Type	Abbreviation	Written Register	Spoken Register
Listing	Lis	＜tag W Lis annotation=" "＞ ＜/tag＞	＜tag S Lis annotation=" "＞ ＜/tag＞
Transitional	Tra	＜tag W Tra annotation=" "＞ ＜/tag＞	＜tag S Tra annotation=" "＞ ＜/tag＞
Appositive	App	＜tag W App annotation=" "＞ ＜/tag＞	＜tag S App annotation=" "＞ ＜/tag＞
Summative	Sum	＜tag W Sum annotation=" "＞ ＜/tag＞	＜tag S Sum annotation=" "＞ ＜/tag＞
Resultive/ Inferential	Res	＜tag W Res annotation=" "＞ ＜/tag＞	＜tag S Res annotation=" "＞ ＜/tag＞
Contrastive	Con	＜tag W Con annotation=" "＞ ＜/tag＞	＜tag S Con annotation=" "＞ ＜/tag＞
Corroborative	Cor	＜tag W Cor annotation=" "＞ ＜/tag＞	＜tag S Cor annotation=" "＞ ＜/tag＞

In addition to the design of the coding scheme, two pitfalls need to be tackled before data analysis as well. One is misuse of CAs, and the other is ill-formed sentences in learner writing.

Since the data sources are Chinese NS learning to write in English, misusing CAs to wrongly indicate textual relations among sentences is inevitable. When a CA misuse happens, textual coherence breaches, the reading flow is interrupted, and the text becomes difficult to comprehend. The pitfall is how to code the misused CA. The use of the CA is incorrect, so it cannot be coded with the textual relation it usually designates. To code the linguistic item with the actual textual relation between sentences is not reasonable, because the coding of the item would be researcher’s interpretation. Due to the fact that there is no way to know what textual relation the writer intended to construct between sentences, the misuse occurrence of CAs is excluded from the investigation scope.

The other pitfall is concerning ill-formed sentences in learner writing. CAs indicate textual relations across sentences, but it is found that the environment where CAs occur varies a lot. For instance, a period is used to end not only sentences but also natural constituents of language, say, a noun phrase. Sometimes it is not a natural constituent of language at all, but a grammatical mistake, such as a pseudo sentence without a finite verb. As opposed to fragmental strings of words, it is also found that sentences may not be separated properly. For example, semicolons are purposefully used to juxtapose a series of unrelated sentences which, in fact, should be severed by periods, or run-on sentences are made without proper punctuation or conjunction (Tseng & Liou, 2006). Due to this ubiquitous structural deficiency, the CAs examined in the present study are those that function and indicate textual relations across units, and a unit is decided as a group of words delimited by a period no matter whether the unit is a complete sentence, a fragment, or a multi-sentence compound.

After taking care of the pitfalls, Figure 6 visualizes the four-filter procedure of coding. Each filter identifies CAs with a linguistic label. The top filter ratifies a CA based on the working definition proposed previously. The second filter judges whether or not the CA is correctly used. The third and last filters identify its register and the textual relation it performs.

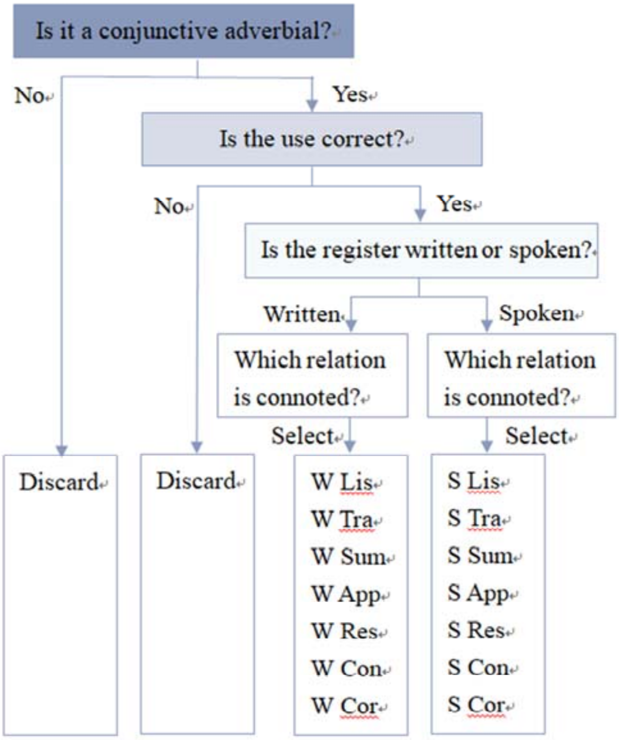


Figure 6. The coding procedure.

One researcher of the present study was the primary data annotator, responsible for annotating all the selected data for analysis. The selected data for analysis were 45 writing pieces by 5 college students. Each student contributed 9 writing pieces for 9 genres, with one piece dedicated to one genre, over 4 semesters. To ensure the reliability of data annotation, one native speaker was recruited as the inter-annotator and annotated 10% of the selected data.

The 10% of the selected data consisted of 5 pieces, with 3 pieced randomly selected from the first three semesters and 2 from the last semester. Following the coding procedure presented in Figure 6, both annotators first decided whether a lexical item satisfied the three criteria of defining a CA, proposed in Section 2.1. Then, annotators elicited their own knowledge to decide whether the CA item was used correctly and what its register was. Lastly, annotators selected a suitable tag to annotate the CA item.

The interface provides a function called Reliability Calculation, as visualized in Figure 1, to automatically report interrator agreement. The function automatically compares both annotators' annotations by examining whether both annotators tag one CA with the same textual relation. When both annotators tag one CA with the same textual relation, it is considered one match. The result shows that the matching agreement is 92.7%. Two reasons might account for the high agreement. First, the working definition of CAs provides clear semantic and syntactic criteria for CA identification. Second, since most CAs convey only one textual relation, both annotators are destined to tag most CAs with the same textual relations. Mismatch may happen only when one CA conveys more than one textual relation, yet the kind of CAs are few.

3.4 Statistical Analysis

After coding the selected data and tallying the counts, all the obtained figures were further analyzed via inferential statistical measures on SPSS to answer the two proposed research questions in the present study.

To begin with, all the raw counts of different textual relations were transformed into the-same-denominator figures to avoid the influence of writing length of the collected data in the different genres. However, while most corpus-based studies obtain the occurrence frequency ratio by using the total word count as the denominator and the CA use count as numerator, the calculation is criticized to be "fundamentally flawed" (Bolton et al., 2002) because CAs function at the discourse level. Therefore, the present study employed the unit count as denominator to normalize the occurrence frequency. As previously defined, a unit is delimited by a period regardless of the sentence structure of the unit. The unit may be a complete sentence, a cluster of words, or a multi-sentence compound.

After attaining normalized occurrence frequencies, various statistical measures were performed to answer two research questions raised. To answer research question one concerning whether genre plays a significant role in the distribution of textual relations expressed by CAs, a two-way within-subjects ANOVA was conducted, with two independent variables being textual relation and genre while the dependent variable being the CA occurrence frequency. To further examine the effect of register, the ANOVA design was calculated again, with the dependent variable becoming the written-register CA occurrence frequency. To answer research question two regarding whether time has a significant influence on the distribution of CA-performed textual relations, a two-way within-subjects ANOVA was carried out. Textual relation and genre were the two independent variables, whereas the CA occurrence frequency was the dependent variable. The ANOVA design was implemented once more to further explore the register effect. Textual relation and genre were still the two independent variables, yet the dependent variable was replaced with the written-register CA occurrence frequency. A significant level of *p*<.05 was chosen.

4. Results

This section reports whether the distribution of CA-performed textual relations varies across genres and time.

4.1 Distribution of Textual Relations Performed by CAs Across Genres

Since the use of CAs can be characterized by written and spoken registers, the distribution of CA-performed textual relations across genres is presented in three conditions, including CAs without register differentiation, written-register CAs, and spoken-register CAs.

4.1.1 Distribution of Textual Relations Through All CAs Across Genres

Table 14 presents the occurring counts of the seven CA-performed textual relations in each of the 9 genres. The raw occurring counts are signaled by *n*, and to evade the influence stemming from various writing lengths of the 9 genres, the raw occurring counts are transformed into occurring counts per 1,000 units.

Table 14. Occurring counts of textual relations performed by all CAs across genres.

Textual Relation	Com-Con		Cau-Eff		Des		Def		Nar		Cla		Mul-Str		Arg		Pro-Sol	
	n	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000
Lis	18	157.9	16	121.2	15	128.2	13	105.7	12	67.8	18	104.1	18	107.1	29	136.2	33	142.2
Tra	0	0.0	0	0.0	2	17.1	1	8.1	2	11.3	0	0.0	0	0.0	2	9.4	1	4.3
App	0	0.0	3	22.7	3	25.6	8	65.0	0	0.0	2	11.6	5	29.8	8	37.6	9	38.8
Sum	2	17.5	1	7.5	1	8.5	1	8.1	0	0.0	1	5.8	1	5.9	1	4.7	1	4.3
Res	3	26.3	7	53.0	1	8.5	4	32.5	3	16.9	4	23.1	6	35.7	12	56.3	14	60.3
Con	14	122.8	5	37.8	7	59.8	13	105.7	7	39.6	14	80.9	13	77.4	18	84.5	21	90.5
Cor	1	8.7	3	22.7	2	17.1	4	32.52	5	28.3	5	28.9	0	0.0	5	23.5	5	21.6
Unit Count	114		132		117		123		177		173		168		213		232	

Based on Table 14, a two-way within-subjects ANOVA is calculated to examine the effects of textual relation and genre. The two independent variables are textual relation and genre, while the dependent variable is the CA occurring counts per 1,000 units. The results show that there is no interaction between textual relation and genre (F(48, 192)=1.070, p=0.366) as well as no main effect from genre (F(8, 32)=1.697, p=0.137). However, there does exist a main effect from textual relation (F(6, 24)=10.476, p<0.05).

Given a main effect from textual relation, Table 15 pinpoints pairs of textual relations with significant differences, and presents related descriptive statistics. It is shown that *Listing* and *Contrastive* have the highest occurrence frequency while *Summative* and *Transitional* have the lowest. The occurrence frequencies of *Resultive/Inferential*, *Appositive* and *Corroborative* are between the two groups.

The reason of presenting the distribution with different compartments rather than in a linear sequence lies in the fact that the 3 textual-relation compartments significantly differ from one another but that there is no significant difference between textual relations within the same compartment. For example, in terms of occurring frequency, *Listing* and *Contrastive* are the highest and second highest, and both textual relations are significantly different from the other textual relations. Yet, the two are not significantly different from each other.

Table 15. Significant differences between textual relations performed by all CAs and related descriptive statistics.

Lis	Tra	App	Sum	Res	Con	Cor		Mean	SD
---	0.022*	0.035*	0.025*	0.060	0.228	0.037*	Lis	23.596	6.625
0.022*	---	0.037*	0.937	0.001*	0.005*	0.038*	Tra	1.147	0.512
0.035*	0.037*	---	0.035*	0.183	0.024*	0.526	App	4.849	1.369
0.025*	0.937	0.035*	---	0.000*	0.002*	0.002*	Sum	1.200	0.565
0.060	0.001*	0.183	0.000*	---	0.014*	0.004*	Res	6.953	0.610
0.228	0.005*	0.024*	0.002*	0.014*	---	0.003*	Con	14.153	2.199
0.037*	0.038*	0.526	0.002*	0.004*	0.003*	---	Cor	4.073	0.917

4.1.2 Distribution of Textual Relations Through Written-Register CAs Across Genres

Table 16 presents the occurring counts of the seven textual relations performed by written-register CAs in each genre. *n* designates the raw occurring counts, which are then transformed into the occurring counts per 1,000 units.

Table 16. Occurring counts of textual relations through written-register CAs across genres.

Textual Relation	Com-Con		Cau-Eff		Des		Def		Nar		Cla		Mul-Str		Arg		Pro-Sol	
	n	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000
Lis	16	140.4	11	83.3	11	94.0	10	81.3	9	50.9	18	104.1	12	71.4	24	112.7	27	116.4
Tra	0	0.0	0	0.0	0	0.0	1	8.1	0	0.0	0	0.0	0	0.0	1	4.7	1	4.3
App	0	0.0	3	22.7	3	25.6	7	56.9	0	0.0	2	11.6	5	29.8	8	37.6	9	38.8
Sum	2	17.5	1	7.6	1	8.6	1	8.1	0	0.0	0	0.0	1	5.9	1	4.7	1	4.3
Res	2	17.5	6	45.5	1	8.6	4	32.5	3	16.9	4	23.1	6	35.7	12	56.3	14	60.3
Con	13	114.0	5	37.9	6	51.3	10	81.3	5	28.3	14	80.9	13	77.4	17	79.8	20	86.2
Cor	1	8.8	3	22.7	2	17.1	4	32.5	5	28.3	2	11.6	0	0.0	5	23.5	5	21.6
Unit Count	114		132		117		123		177		173		168		213		232	

The transformed figures in Table 16 are calculated through a two-way within-subjects ANOVA to examine the distribution of textual relation performed by written-register CAs across genres. The two independent variables are textual relation and genre, while the dependent variable is the normalized occurring counts per 1,000 units. The results show that there is no interaction between textual relation and genre (F(48, 144)=0.969, p=0.537) as well as no main effect from genre (F(8, 24)=2.062, p=0.082). However, there does exist the main effect from textual relation (F(6, 18)=8.585, p<0.05).

Table 17 presents related descriptive statistics and pinpoints pairs of textual relations with significant differences. At first glance, there seems to be a distribution norm of textual relations performed by written-register CAs, much similar to that based on all CAs. Nevertheless, the seeming distribution norm is an illusion, because most textual relations do not differ from one another on a significant level. Take *Listing* and *Transitional* as an example. The former does not significantly differ from any textual relations while the latter only differs from *Contrastive* and *Resultive/Inferential*, which is very different from what happens in the distribution of textual relations performed by CAs without differentiating registers.

Table 17. Significant differences between textual relations through written-register CAs across genres and related descriptive statistics.

Lis	Tra	App	Sum	Res	Con	Cor		Mean	SD
---	0.050	0.065	0.050	0.111	0.393	0.080	Lis	20.465	6.458
0.050	---	0.084	0.759	0.008*	0.012*	0.075	Tra	0.514	0.299
0.068	0.084	---	0.056	0.283	0.045*	0.519	App	4.974	1.659
0.050	0.759	0.056	---	0.003*	0.009*	0.033*	Sum	0.391	0.231
0.111	0.008*	0.283	0.003*	---	0.024*	0.005*	Res	6.672	0.906
0.393	0.012*	0.045*	0.009*	0.024*	---	0.009*	Con	13.966	2.445
0.080	0.075	0.519	0.033*	0.005*	0.009*	---	Cor	4.028	1.169

4.1.3 Distribution of Textual Relations Through Spoken-Register CAs Across Genres

Table 18 presents the occurring counts of seven textual relations through spoken-register CAs in each genre, with *n* referring to the raw occurring counts and its transformed counts per 1,000 units. Based on Table 18, in most genres, few textual relations are performed by spoken-register CAs except for *Listing* and *Contrastive*. Moreover, the occurrence of *Contrastive* is limited, with only one or two cases. It is *Listing* that appears most frequently in the spoken CA form. Due to the scarce occurrence of spoken-register CAs, statistical analysis is not employed in this part.

Table 18. Occurring counts of textual relations through spoken-register CAs across genres.

Textual Relation	Com-Con		Cau-Eff		Des		Def		Nar		Cla		Mul-Str		Arg		Pro-Sol	
	n	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000
Lis	2	17.5	5	37.9	4	34.2	3	24.4	3	17.0	0	0.0	6	35.7	5	23.5	6	25.7
Tra	0	0.0	0	0.0	2	17.1	0	0.0	2	11.3	0	0.0	0	0.0	1	4.7	0	0.0
App	0	0.0	0	0.0	0	0.0	1	8.1	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Sum	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	1	5.8	0	0.0	0	0.0	0	0.0
Res	1	8.8	1	7.6	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Con	1	8.8	0	0.0	1	8.6	3	24.4	2	11.3	0	0.0	0	0.0	1	4.7	1	4.3
Cor	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	3	17.3	0	0.0	0	0.0	0	0.0
Unit Count	114		132		117		123		177		173		168		213		232	

4.2 Distribution of Textual Relations Performed by CAs over Time

This part seeks to answer whether or not the distribution of CA-performed textual relations would vary in accordance to different time periods on a significant level. First, the results are presented without differentiating CAs in register. Then, the results are shown in terms of CAs in written register and in spoken register.

4.2.1 Distribution of Textual Relations Through All CAs over Time

To investigate the distribution of CA-performed textual relations over time, a temporal unit is set to be a semester. The data collected are sorted according to the semester when participants wrote the piece of writing. Table 19 presents the occurring counts of the seven textual relations performed by CAs in each of the four semesters. Again, the raw occurring counts are signaled by n, and transformed into the occurring counts per 1,000 units.

The results of examining the effects of textual relation and time on the distribution are attained via a two-way within-subjects ANOVA, with the two independent variables being textual relation and time as well as the dependent variable the occurring counts per 1,000 units. No interaction between textual relation and time (F(18, 72)=0.912, p=0.567) is found, nor is any main effect from time (F(3, 12)=2.147, p=0.147). Nevertheless, there exists a main effect from textual relation (F(6, 24)=11.318, p<0.05).

Table 19. Occurring counts of textual relations performed by all CAs over time.

Textual Relation	Time							
	Semester 2		Semester 3		Semester 4		Semester 5	
	n	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000
Lis	34	138.21	39	93.53	36	105.60	62	139.33
Tra	0	0	6	14.39	0	0	2	4.49
App	3	12.20	10	23.98	7	20.53	16	35.96
Sum	3	12.20	1	2.398	2	5.87	2	4.49
Res	7	28.46	8	19.18	10	29.33	26	58.43
Con	18	73.17	23	55.16	26	76.25	35	78.65
Cor	4	16.26	11	26.38	5	14.66	10	22.47
Unit Count	246		417		341		445	

Table 20 presents descriptive statistics of the occurring counts of CAs signaling 7 textual relations, and locates pairs of textual relations with significant differences. The results show that the distribution of CA-performed textual relations is the same as that found in the previous section. *Listing* and *Contrastive* occur the most frequently, *Summative* and *Transitional* appear the least frequently, and the occurring frequencies of *Resultive/Inferential*, *Appositive* and *Corroborative* rank in the middle. The results also show that while there is a significant difference among different compartments, no significant difference between textual relations within the same compartment is found.

Table 20. Significant differences between textual relations through all CAs over time and related descriptive statistics.

Lis	Tra	App	Sum	Res	Con	Cor		Mean	SD
---	0.019*	0.029*	0.022*	0.055	0.218	0.033*	Lis	23.832	6.405
0.019*	---	0.033*	0.676	0.003*	0.003*	0.028*	Tra	0.944	0.399
0.029*	0.033*	---	0.033*	0.136	0.017*	0.546	App	4.633	1.267
0.022*	0.676	0.033*	---	0.000*	0.001*	0.002*	Sum	1.248	0.589
0.055	0.003*	0.136	0.000*	---	0.010*	0.004*	Res	6.770	0.761
0.218	0.003*	0.017*	0.001*	0.010*	---	0.003*	Con	14.161	2.023
0.033*	0.028*	0.546	0.002*	0.004*	0.003*	---	Cor	3.989	0.839

4.2.2 Distribution of Textual Relations Through Written-Register CAs over Time

Table 21 presents the occurring counts of textual relations conveyed by written-register CAs in each semester. The raw occurring counts are symbolized by *n*, and then transformed into the occurring counts per 1,000 units.

After calculating a two-way within-subjects ANOVA, with the two independent variables being textual relation and time and the dependent variable the occurring counts per 1,000 units, it is found that no interaction between textual relation and time (F(18, 72)=1.154, p=0.322). However, there exist main effects from textual relation (F(6, 24)=11.344, p<0.05) and from time (F(3, 12)=4.524, p<0.05).

Table 22 presents related descriptive statistics and shows pairs of textual relations with significant differences in occurrence frequency. The results show that there exists a distribution norm, with *Listing* and *Contrastive* being the most frequent textual relations performed by written-register CAs. *Resultive/Inferential*, *Appositive,* and *Corroborative* are the second most. Finally, *Summative* and *Transitional* are the least frequent.

Table 23 presents related descriptive statistics and indicates the occurrence of significant difference among semesters in terms of written-register CA use. The results show that, except for semester 2, the use of written-register CAs significantly grows semester after semester.

Table 21. Occurring counts of textual relations through written-register CAs over time.

Textual Relations	Time							
	Semester 2		Semester 3		Semester 4		Semester 5	
	n	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000	*n*	*n per* 1,000
Lis	26	105.69	28	67.10	30	88.00	52	116.90
Tra	0	0	2	4.80	0	0	1	2.25
App	3	12.20	9	21.60	7	20.53	16	35.96
Sum	1	4.07	1	2.40	1	2.93	0	0
Res	6	24.39	7	16.80	10	29.33	26	58.43
Con	17	69.11	20	48	26	76.25	33	74.16
Cor	4	16.26	11	26.40	2	5.87	10	22.47
Unit Count	246		417		341		445	

Table 22. Significant differences between textual relations through written-register CAs over time and related descriptive statistics.

Lis	Tra	App	Sum	Res	Con	Cor		Mean	SD
---	0.023*	0.033*	0.025*	0.074	0.355	0.040*	Lis	18.883	5.260
0.023*	---	0.036*	0.771	0.003*	0.002*	0.036*	Tra	0.352	0.233
0.033*	0.036*	---	0.028*	0.123	0.011*	0.373	App	4.513	1.244
0.025*	0.771	0.028*	---	0.123	0.011*	0.373	Sum	0.470	0.203
0.074	0.003*	0.123	0.001*	---	0.005*	0.001*	Res	6.446	0.844
0.355	0.002*	0.011*	0.002*	0.005*	---	0.001*	Con	13.374	1.774
0.040*	0.036*	0.373	0.026*	0.001*	0.001*	---	Cor	3.549	0.923

Table 23. Significant differences between semesters through written-register CAs over time and related descriptive statistics.

Semester 2	Semester 3	Semester 4	Semester 5		Mean	SD
---	0.290	0.817	0.124	Semester 2	6.620	1.226
0.290	---	0.390	0.007*	Semester 3	5.344	1.553
0.817	0.390	---	0.045*	Semester 4	6.368	0.959
0.124	0.007*	0.045*	---	Semester 5	8.860	1.270

4.2.3 Distribution of Textual Relations Through Spoken-Register CAs over Time

Table 24 presents the occurring counts of seven textual relations performed by spoken-register CAs in each semester, with *n* referring to the raw occurring counts and then transformed into the occurring counts per 1,000 units. Owing to the zero occurrences of many textual relations performed by spoken-register CAs, no statistical analysis is performed. Nevertheless, it is found that the use of spoken-register CAs and the types of textual relations performed by spoken-register CAs are diminishing in a steady fashion.

Table 24. Occurring counts of textual relations through spoken-register CAs over time.

Textual Relations	Time							
	Semester2		Semester3		Semester4		Semester5	
	n	*n per* 1,000	*n*	*n per* 1,000	*N*	*n per* 1,000	*n*	*n per* 1,000
Lis	8	32.520	11	26.379	6	17.595	0	0
Tra	0	0	4	9.592	0	0	1	2.247
App	0	0	1	2.398	0	0	0	0
Sum	2	8.130	0	0	1	2.932	2	4.494
Res	1	4.065	1	2.398	0	0	0	0
Con	1	4.065	3	7.194	0	0	2	4.494
Cor	0	0	0	0	3	8.798	0	0
Unit Count	246		417		341		445	

5. Discussion

This section respectively discusses the attained results concerning the distribution of CA-performed textual relations in the English writing by Chinese NS across genres and over time, and is wrapped up by a general discussion.

5.1 Discussion on Distribution of Textual Relations Through CAs Across Genres

A two-way ANOVA is calculated to examine the influence of genre and textual relation on the distribution of textual relations manifested by CAs. It is found that while genre has no role in impacting the distribution, there exists a norm distribution of textual relations across genres. In the norm distribution, textual relations are compartmented into three groups based on occurrence frequency. *Listing* and *Contrastive* occur most frequently. *Resultive/Inferential*, *Appositive*, and *Corroborative* have the second most occurrence frequency. *Summative* and

Transitional are the least frequent. The occurrence frequencies of the three groups significantly differ from one another, but for textual relations within any group, their occurrence frequencies do not differ from each other on a significant level. In addition, no significant difference among the three groups of textual relations is found when CA counts are differentiated by register. That is, the use of written-register CAs does not present a norm distribution of textual relations.

Table 25 presents a comparison between the norm distribution found in the present study and that in the previous studies. It is found that the two distributions highly resemble each other. Despite some differences in classification, *Listing* and *Contrastive* still occur most frequently, and *Transitional* is still the least frequent. Yet, *Resultive* and *Summative* in the present study occur less frequently than those in the previous studies.

Table 25. Comparison between the distributions of the textual relations.

Distribution	Most frequent		Moderate frequent		Least frequent
Distribution of the present study	Listing Contrastive	>	Resultive/Inferential Appositive Corroborative	>	Summative Transitional
Distribution of the previous studies*	Listing Contrastive Resultive	>	Appositive Summative	>	Transitive Inferential

*The distribution is based on Table xx in Section 2.2.

The lack of genre influence might be in relation to textual-relation-creating mechanism. Textual relations can be manifested by various mechanisms. For instance, CAs can be easily replaced with and rephrased by discourse-organizing words (McCarthy, 1991; Winter, 1977; Yu, 2007). As shown in the following example, the CA in (1) can be rephrased as the noun phrase in (2), and the listing relation is still conveyed. Thus, while various genres may have their own distinct discoursal characteristics, the CA use alone may not be sufficient to distinguish genres.

Taiwan officially becomes an aged society.

1) Firstly, the birth rate is substantially declining.

2) The first reason is that the birth rate is substantially declining.

Another reason to explain the lack of genre influence is that genres are not mutually exclusive. Even though different genres may be constructed with different discoursal characteristics, they may also incorporate characteristics from one another, especially when writing length becomes longer. Considering that some of the writing pieces are twice the length of the others in the present study, the nine genres investigated may share many characteristics, which, in some sense, makes the nine genres a general superordinate genre. Therefore, no genre influence is found.

As opposed to lack of genre influence, CA-performed textual relations present a constant distribution norm across genres. Two explanations are proposed to account for it. The first explanation is that the nature of different textual relations has forecast their occurrence frequencies. For example, it is understandable that *Summative* is in the group where textual relations occur least frequently, because *Summative* indicates a conclusion which only appears at the end of a text no matter how long the text is. In contrast, *Listing* occurs most frequently, because writers can always employ *Listing* CAs to indicate more ideas to come without limitation.

The second explanation for a norm distribution of textual relations is that there exists a preference programmed in human cognition for employing CAs to convey certain textual relations. Take *Contrastive* as example. The textual relation is relatively complicated because it requires an effort to analyze two events and to locate the contrastive points. Therefore, it would take more energy to describe the relation in text compared with writing in the common temporal sequence. Due to the extra energy required, Economy Principle is applied in order to minimize the energy consumption while to successfully achieve communication (Ungerer & Schmid, 2006). Based on the rationale, it is inferred that the tendency to select CAs, rather than other textual-relation-creating devices, to convey the contrastive relation is programmed in human cognition since one or two words of CAs are enough to minimize the energy consumption and to express the textual relation clearly. Ultimately, *Contrastive* becomes one of the textual relations most frequently performed by CAs.

Two pieces of evidence may support the preset preference of certain textual relations manifested by CAs in human cognition. The observational evidence is that regardless of genre difference, *Contrastive* has the highest occurrence frequency in all NS data and some NNS data from previous studies (Chen, 2006; Altenberg & Tapper, 1998; Field & Yip, 1992; Shen, 2006) reviewed in section 2. The statistic evidence is that a distribution norm of textual relations is found through all CAs, but not through written-register CAs. Because register is a manmade literary concept, neither written-register CAs nor spoken-register ones can completely reflect human cognition. Therefore, deliberately exploring written-register CAs alone and ignoring spoken-register CAs fails to construct a distribution norm of textual relations on a significant level. Only when all CAs are considered without register differentiation can human cognition be fully represented by a norm distribution of textual relations across genres.

5.2 Discussion on Distribution of Textual Relations Through CAs over Time

To explore whether time and textual relation affect the distribution of textual relations conveyed by CAs, a two-way ANOVA is calculated, and the results are similar to those found when exploring the effects of genre and textual relation on the distribution. A distribution norm of CA-performed textual relations free from time influence is found. The distribution norm is divided into three groups. *Listing* and *Contrastive* are the most frequent textual relations manifested by CAs, *Resultive/Inferential*, *Appositive* and *Corroborative* are the second most frequent, and *Summative* and *Transitional* are the least frequent. The three groups are significantly different from one another in terms of occurrence frequency, and textual relations within any group do not. However, when CAs are separated from written register from spoken register, time is found to have a main effect on the CA use, with more use of written-register CAs in later semesters.

The lack of time influence on the distribution of textual relations might be in relation to cognitive development. According to Inhelder and Piaget (1999), the development of logical thinking reaches maturation after adolescence. Therefore, it may be assumed that once the logical thinking becomes less variable, an innate distribution norm of textual relations to express ideas in human cognition may emerge accordingly. Since the data sources in the present study are college students with mature cognition, time is no longer a factor affecting their use of CAs to perform various textual relations. Instead, a distribution preference is reflected when textual relations are performed by CA.

In contrast, according to the statistic results, time has a main effect on the register use of CAs. Over time, the use of written-register CAs is significantly increasing while that of spoken-register CAs is decreasing. The result is understandable. Since register is often taught and then acquired by writing learners through education, the more time students stay in school and receive writing training, the more skillful students are to write in written register and avoid spoken register. Moreover, the fact that register use can be taught over time while the distribution of textual relations via CAs is immune to time highlights the possibility of a norm distribution existing in human cognition, because the norm distribution cannot be taught and changed over time.

5.3 General Discussion

According to the results, neither genre nor time has an effect on the occurrence frequencies of CAs manifesting various textual relations. Instead, a distribution norm of CA-performed textual relations across genres and over time is found. In the distribution norm, *Listing* and *Contrastive* have the highest occurrence frequency. *Summative* and *Transitional* have the least. *Resultive/Inferential, Appositive* and *Corroborative* rank in the middle of the frequency order.

As reviewed in section 2, previous studies also reported similar results. A distribution independent of genre and time was found in the NS writing in studies based on four CA-performed textual relations (Chen, 2006; Field & Yip, 1992) as well as in both NS and NNS writing in studies based on more CA-performed textual relations (Altenberg & Tapper, 1998; Shen, 2006; Tankó, 2004). Moreover, the formations of these found distribution patterns are very similar as well. *Contrastive* which is based on the more-type framework and equates *Adversative* in the four-type framework, usually occurs most frequently. *Summative* and *Transitional* happen least frequently.

The striking similarity between the results in the present study and those in previous studies is in support of the contention that there exists a preset distribution preference of using CAs to manifest various textual relations in human cognition. It reflects how the human mind perceives these textual relations in terms of logical complex. In other words, the distribution norm of CA-performed textual relations based on CA occurrence frequency is a mental representation of human cognition. The register factor provides more evidence to support the contention. When CAs are divided by register, no distribution pattern of textual relations performed by written-register CAs can reach a significant level. This is because, without elements embodied by spoken-register CAs in real world, the mental representation becomes flawed, and it is this incomplete mental representation that no distribution pattern exactly reflects.

Since the data source in the present study is the English writing by Chinese NS, another relevant issue is whether the found distribution norm also underlies the English writing by its NS. In light of the fact that the found distribution norm in the present study is very similar to the distribution identified in both NS and NNS writing in studies with a framework of fine classification (Altenberg & Tapper, 1998; Shen, 2006; Tankó, 2004), it is suggested that the found distribution pattern may be insusceptible to the first language influence. That is to say, the found distribution is universal in the English writing in spite of writers' language background. Any English writing pieces where the found distribution cannot be extracted may be regarded as ill-composed whether the writers are English NS or NNS.

The finding has great application potential in automation of discourse diagnosis. Up to date, researchers has attempted to develop automatic tools to diagnose English writing on a discourse level based on sentence length, syllable counts, and difficulty levels of vocabulary (Chall & Dale, 1995; Klare, 1984). The outcome has not been satisfactory for these linguistic characteristics are on local and shallow levels (Bailin & Grafstein, 2001). Benjamin (2012) points out that only by taking into account factors on global and deeper levels can automatic tools judge whether a writing piece constructs a coherence mental representation and produce reliable discourse diagnosis. On that note, the distribution found in the present study can serve as a crucial criterion for automatic tool development. Such a possible tool can extract CAs in a piece of writing and compare the distribution pattern of textual relations performed by these extracted CAs to the found distribution norm. Based on the matching degree, how well-constructed the piece is can be evaluated accordingly. What's better, since the found distribution is not subject to genre, time, and the first language influence, tools featuring the distribution criterion can be available to an all-inclusive variety of users.

6. Conclusion

The present study began with an investigation into the use of CAs performing various textual relations, and discovered that a distribution norm of CA-performed textual relations based on CA occurrence frequency persists across genres and over time. The found distribution can serve as an indicator of discoursal coherence. For a piece of English writing presenting the found distribution during discourse analysis, it may be considered potentially coherent. Instructors can also point out the incoherence in learners' writing by referring to the deviation from the found distribution. The study ended up suggesting using the found distribution as an evaluating criterion for developing automatic tools of discourse diagnosis.

For further research, two possibilities await. Firstly, *Listing* and *Contrastive* are two textual relations with highest occurring frequencies. While the high frequency of *Contrastive* can be explained by the cognitively economic reason, that of *Listing* is said be rooted in teaching instructions (Shen, 2006). Due to the insufficient English proficiency of NNS, NNS might be encouraged to employ more *Listing* CAs because it is quick to construct the textual structure, which leads to *Listing* ubiquity in the NNS writing. Thus, even though genre and time have no influence on the found distribution, whether teaching instructions plays a role in the distribution remains unknown. Secondly, the mechanisms to realize textual relations are not limited to the CA use. Whether the found distribution still holds after including the derived and paraphrased forms of CAs, such as preposition expressions with references and discourse-organizing words, is also worth pursuing.

References

Altenberg, B. & Tapper, M. (1998). *The use of adverbial connectors in advanced Swedish learners' written English.* (First edition ed.). In S. Granger (Ed.), Learner English on Computer (pp. 80-93). Longman.

Atkins, S., Clear, J. & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing*, *7*(1), 1-16. https://doi.org/10.1093/llc/7.1.1

Bailin, A., & Grafstein, A. (2001). The linguistic assumptions underlying readability formula: A critique. *Language & Communication*, *21*(3), 285-301. https://doi.org/10.1016/S0271-5309(01)00005-2

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, *24,* 63-88. https://doi.org/10.1007/s10648-011-9181-8

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Pearson Education.

Bolton, K., Nelson, G. & Hung, J. (2002). A corpus-based study of connectors in student writing: Research from the International Corpus of English in Hong Kong (ICE-HK). *International Journal of Corpus Linguistics*, *7*(2), 165-182. http://dx.doi.org/10.1075/ijcl.7.2.02bol

Celce-Murcia, M. & Larsen-Freeman, D. (1999). *The Grammar Book: An ESL/EFL Teacher's Course*. (Second Edition ed.). Heinle & Heinle Publishers.

Chall, J. S., & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.

Chen, C. W.-y. (2006). The use of conjunctive adverbials in the academic papers of advanced Taiwanese EFL learners. *International Journal of Corpus Linguistics*, *11*(1), 113-130. http://dx.doi.org/10.1075/ijcl.11.1.05che

Connelly, M. (2013). *Get writing: Sentences and paragraphs*. Thomson Higher Education.

Field, Y. & Yip, L. (1992). A comparison of Internal conjunctive cohesion in the English essay writing of Cantonese speakers and native speakers of English. *RELC Journal, 23*(1), 15-28. https://doi.org/10.1177/003368829202300102

Granger, S. & Tyson, S. (1996). Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes*, *15*(1), 17-27. https://doi.org/10.1111/j.1467-971X.1996.tb00089.x

Halliday, M.A.K. & Hasan, R. (1976). *Cohesion in English*. Routledge.

Huang, K.-H. (2018). *Textual Cohesion and Cohesive Errors in English Writing of Taiwan University Students* (Master’s thesis). Retrieved from https://hdl.handle.net/11296/35mj85

Inhelder, B. & Piaget, J. (1999). *The Growth of Logical Thinking from Childhood to Adolescence*. Routledge.

Klare, G. (1984). Readability. In Pearson, P. *et al.* (Eds), *Handbook of Reading Research*, (pp. 681-744). Longman.

Langan, J. (2010). *Paragraph Skills*. McGraw-Hill.

Lannon, J. M. (2007). *The writing process: A concise rhetoric, reader, and handbook.* Longman.

Liu, M. & Braine, G. (2005) Cohesive features in argumentative writing produced by Chinese undergraduates. *System*, *33*(4), 623-636. https://doi.org/10.1016/j.system.2005.02.002

McCarthy, M. (1991). *Discourse Analysis for Language Teachers*. Cambridge University Press.

Ministry of Education (MOE). (2014). *Curriculum Guidelines of 12-Year Basic Education: General Guidelines*. Retrieved May 03, 2020, from https://www.naer.edu.tw/files/15-1000-14113,c639-1.php?Lang=zh-tw

Ministry of Education (MOE). (2018). *Curriculum Guidelines of 12-Year Basic Education: Guidelines for English.* Retrieved May 03, 2020, from https://www.naer.edu.tw/files/15-1000-14113,c639-1.php?Lang=zh-tw

Morenberg, M. & Sommers, J. (2008). *The writer's options: Lessons in style and arrangement*. Pearson Longman.

Morey, M., Muller, P. & Asher, N. (2018). A Dependency Perspective on RST Discourse Parsing and Evaluation. *Computational Linguistics*, *44*(2), 197-235. https://doi.org/10.1162/COLI_a_00314

Phoocharoensil, S. (2017). Corpus-based exploration of linking adverbials of result: Discovering what ELT writing coursebooks lack. *The Southeast Asian Journal of English Language Studies*, *23*(1), 150-167. http://doi.org/10.17576/3L-2017-2301-11

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman.

Reid, J. M. (2000). *The process of composition*. Longman.

Shen, T.-C. (2006). *Advanced EFL Learners’ Use of Conjunctive Adverbials in Academic Writing* (Master’s Thesis). Taiwan: National Taiwan Normal University.

Simons, G. & Bird, S. (2008, May 31). *OLAC Metadata. OLAC: Open Language Archives Community*. Retrieved May 14, 2020, from http://www.language-archives.org/OLAC/metadata.html

Smalley, R. L., Ruetten, M. K. & Kozyrev, J. R. (2012). *Refining composition skills: Academic writing and grammar*. Cengage.

Suehring, S. (2006). Beginning web development with Perl: from novice to professional. Springer.

Tankó, G. (2004). *The use of adverbial connectors in Hungarian university students' argumentative essays*. In Sinclair, J. M. (Ed.), How to Use Corpora in Language Teaching (pp. 157-181). John Benjamins. https://doi.org/10.1075/scl.12.13tan

Tseng, Y.-C. & Liou, H.-C. (2006). The effects of online conjunction materials on college EFL students' writing. *System*, *34*(2), 270-283. https://doi.org/10.1016/j.system.2006.01.006

Ungerer, F. & Schmid, H.-J. (2006). *An Introduction to Cognitive Linguistics*. Routledge.

Winter, E. O. (1977). A clause-relational approach to Eenglish texts: A study of some predictive lexical items in written discourse. *Instructional Science*, 6, 1-92. https://doi.org/10.1007/BF00125597

Wyrick, J. (2008). *Steps to writing well: With additional readings*. Wadsworth.

Yu, H.-y. (2007). Discourse grammar for academic reading: Textual relationships. *English Teaching & Learning*, *31*(2), 159-197. https://doi.org/10.6330/ETL.2007.31.2.05

Xie, J.-Z. (2014). *Textual Cohesion in Senior High School Students' Expository Writings in Southern Taiwan* (Master's thesis). Taiwan, National Pingtung University of Education.