診斷學習者英語寫作篇章結構：以篇章連接副詞爲例

Diagnosing discoursal organization in learner writing via conjunctive adverbials

高東榆 Tung-yu Kao
國立成功大學外國語文學系（所）
Department of Foreign Languages & Literature
National Cheng Kung University
dodofishletter@hotmail.com

陳麗美 Li-mei Chen
國立成功大學外國語文學系（所）
Department of Foreign Languages & Literature
National Cheng Kung University
leemay@mail.ncku.edu.tw

Abstract

The present study aims to investigate genre influence on the use and misuse of conjunctive adverbials (hereafter CAs) by compiling a learner corpus annotated with discoursal information on CAs. To do so, an online interface is constructed to collect and annotate data, and an annotating system for identifying the use and misuse of CAs is developed. The results show that genre difference has no impact on the use and misuse of CAs, but that there does exist a norm distribution of textual relations performed by CAs, indicating a preference preset in human cognition. Statistic analysis also shows that the proposed misuse patterns do significantly differ from one another in terms of appropriateness and necessity, ratifying the need to differentiate these misuse patterns. The results in the present study have three possible applications. First, the annotate data can serve as training data for developing technology that automatically diagnoses learner writing on the discoursal level. Second, the founding that textual relations performed by CAs form a distribution norm can be used as a principle to evaluate discoursal organization in learner writing. Lastly, the misuse framework not only identifies the location of misuse of CAs but also indicates direction for correction.

Keywords: conjunctive adverbial, textual relation, misuse pattern, learner corpus.

1. Introduction

Due to much interest in learning English around the globe, many tools are developed, or wanted to be developed, to facilitate learners to learn English better. One of many wanted tools is probably a tool that can automatically diagnose a piece of learner writing and provide direction for improvement of the writing. The need results from the fact that only

by constantly revising process can learners keep polishing their writing skill but that there is just not enough manpower to help learners recognize the defects in their writing. Therefore, much software is developed to satisfy the need, such as the two famous online writing platforms, *My Access!* and *Criterion*, and the two popular writing software packages, *StyleWriter* and *White Smoke*.

However, after evaluating the above mentioned tools aiming to automatically diagnose learner writing, it is found that the diagnosis is mainly a grammar check at the sentence level yet fails to generate revising suggestions on the discourse level. In other words, the existing tools may help learners compose a piece of writing free from grammatical mistakes, but poor organization of sentences and anomaly in coherence may still lead to failure in comprehension. Therefore, a writing-facilitating tool that can automatically diagnose learner writing on the discourse level is further wanted. To do so, a further investigation of existing learner corpora is made to seek if they fit as training data for developing such tools in question. The result shows that all the three corpora under investigation, Taiwanese Learner Corpus of English (TLCE) [1], Chinese Learner English Corpus (CLEC) [2], and International Corpus of Learner English (ICLE) [3], are only annotated with linguistic information at the sentence level, which limits further development on the discourse level. In light of the investigation, the first goal of the present study is to construct a learner corpus that provides annotated discoursal information as a basis for developing technology that can automatically diagnose learner writing in terms of discoursal organization.

With the goal in mind, the correct use and misuse of conjunctive adverbials are selected as the discoursal information that is used to annotate the targeted learner corpus. In terms of correct use, many writing textbooks introduce conjunctive adverbials (hereafter CA) as explicit linguistic features that organize textual relation among sentences in a coherent order, and contend that CAs performing certain textual relation would be more prominent in certain genre [4] [5] [6] [7] [8]. For instance, the words or phrases, such as *firstly*, *next*, and *in addition*, are thought to appear more in the process genre, indicating progressive relations in the text. Yet, after reviewing literature [9] [10] [11] [12] [13] [14], it is found that the textual relations performed by CAs present a norm distribution no matter which genre the writing belongs to, which is contrary to what writing textbooks usually suggest.

In terms of misuse of CAs, [15] regulates three common misuse patterns, *non-equivalent exchange*, *connective overuse*, and *surface logicality*, that often occur in learner writing. However, after trying applying the misuse framework of CAs to classify the mistakes found in learner writing, the framework is found insufficient in doing so. Based on the review of literature on CAs, the second goal of the present study aims to empirically examine if writing genres play a role in the use of CAs, and to propose a framework that can better describe the misuse patterns of CAs found in learner writing.

In short, the present study is two-fold. One is to compile a learner corpus annotated

with discoursal information, to be specific, information on CAs, which can serve as training data of developing technology that automatically diagnoses learner writing on the discoursal level, while the other is to investigate genre influence on use and misuse of CAs and to construct a misuse framework for CAs.

2. Annotating system of CAs

The annotating system developed in the present study is used to annotate learner writing in terms of the use and misuse of CAs that organize textual relation among sentences. The set of annotations that indicates textual relations performed by CAs is based on the taxonomy in [16], whereas the set concerning misuse patterns of CAs is on the classification in [15].

2.1 Annotation for textual relations by CAs

According to the taxonomy in [16], there are seven types of CAs that organize seven textual relations among sentences, which include *Listing*, *Transitional*, *Appositive*, *Summative*, *Resultive*, *Inferential*, and *Contrastive*. In the present study, two textual relations, *Resultive* and *Inferential*, are collapsed into one since both indicate the cause-effect textual relation, and one additional textual relation, *Corroborative*, is supplemented. As a result, seven types of textual relations performed by CAs are used to annotate learner writing, which are *Listing*, *Transitional*, *Appositive*, *Summative*, *Resultive/Inferential*, *Contrastive*, and *Corroborative*. Table 1 lists all the textual relations with their definitions and the possible language items performing these relations. Notice that Table 1 also shows that one language item may serve more than one textual relation, for example, the language item *then* is in both *Listing* and *Resultive/Inferential* relations. In other words, the semantic annotation must depend on the relation performed by the CA, not on certain fixed language items.

Table 1. The Set of Textual Relations Indicated by CAs

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

Another issue regarding the annotation of the textual relations is register. Register refers to the fact that CAs performing the same textual relation are further classified into written register and spoken register, with the latter is considered informal and suggested to be avoided in formal writing. Take *moreover* and *plus* for example. While both CAs indicate the *Listing* textual relation, the use of the latter is sometimes seen as a misuse for its informal nature in writing. Given the distinction in CA register use, the annotating system also differentiates CAs performing the same textual relation in terms of register to examine the influence genre difference has on register use in CAs.

2.2 Misuse Patterns of Conjunctive Adverbials (CAs)

In contrast with the set that annotates learner writing with textual relations performed by CAs, the other set in the annotating system is to indicate the misuse of CAs when they fail to logically connect sentences or do not appropriately fit the context. With the three misuse patterns proposed in [15], there are six misuse patterns in total generalized in the present study, which are *Non-equivalent Exchange*, *Connective Overuse*, *Surface Logicality*, *Wrong Relation*, *Semantic Incompletion,* and *Distraction*. Table 2 showcases the six misuse patterns with their definitions and examples.

Table 2. The Set of Misuse Patterns of CAs

Misuse Pattern	Definition & Example
Non-equivalent Exchange	Use CAs conveying the same textual relation in an interchangeable manner when they are not • Those are the images of the UK that the Communists want to impose on the local Chinese. *<u>On the contrary</u>*, they describe the communists as patriotic Chinese who did not show the slightest fear.

Connective Overuse	Use CAs with high density in short texts, making texts fragmental and readers unable to expect where texts are going to lead. • The communicative approach proves not only practicable for juniors, but also for senior. ***However***, only the junior forms were observed. ***Nevertheless***, the study in juniors is essential for this is the stage when students establish the right ways of learning English.
Surface Logicality	Use CAs to impose logicality to texts or bridge the gap among propositions when there exists no deep logicality in texts. • This question means the same as 'Evaluate the degree to which Japanese imperialism was a result of militarism.' ***So*** this question requires an independent argument about them. ***So*** the student must think critically if Japanese imperialism was a result of militarism.
Wrong Relation	Use a CA to express certain textual relation that it does not express. • Many studies have showed that it would be better for the hearing disabled to have the cochlear implant at an early age. ***Also***, if implanted the cochlear implant at the age one to two, their language learning could come out of great improvement.
Semantic Incompletion	The context where CAs are used needs more elaboration to make the CAs functional. • After finishing the competitive entrance exam, you enter the college. ***However***, nowadays, graduating from college not necessarily guarantees you future.
Distraction	The context would be coherent itself without the use of the conjunctive adverbial or that the use is redundant. • Statistics that four countries had higher averages of education than Taiwan. ***For example***, the percentage to get admitted to college of Finland and South Korea is 90 percent, New Zealand with 86 percent and Sweden with 84 percent.

2.3 Annotating system in electronic format

In total, there are 20 labels, 14 for identifying textual relations and 6 for recording misuse patterns, in the developed coding scheme, as presented in Table 3.

Table 3. The Complete annotating system

Textual relations			Misuse Patterns	
Register	Type	Abbreviation	Type	Abbreviation
Y / R	Listing	Y/R Lis	Non-equivalent Exchange	N NE
Y / R	Transitional	Y/R Tra	Connective Overuse	N CO

Y / R	Appositive	Y/R App	Surface Logicality	N SL
Y / R	Summative	Y/R Sum	Wrong Relation	N WR
Y / R	Resultive/Inferential	Y/R Res	Semantic Incompletion	N SI
Y / R	Contrastive	Y/R Con	Distraction	N DI
Y / R	Corroborative	Y/R Cor		

Then, to make the annotating system applicable to computational development, the system is converted into digital tags that preserve the linguistic information on the text. Table 4 presents the 20 digital tags.

Table 4. The digital tags in the annotating system

Textual relations (Written Register)	＜tag Y Lis anno=" "＞＜/tag＞	＜tag Y Res anno=" "＞＜/tag＞
	＜tag Y Tra anno=" "＞＜/tag＞	＜tag Y Con anno=" "＞＜/tag＞
	＜tag Y App anno=" "＞＜/tag＞	＜tag Y Cor anno=" "＞＜/tag＞
	＜tag Y Sum anno=" "＞＜/tag＞	
Textual relations (Spoken Register)	＜tag R Lis anno=" "＞＜/tag＞	＜tag R Res anno=" "＞＜/tag＞
	＜tag R Tra anno=" "＞＜/tag＞	＜tag R Con anno=" "＞＜/tag＞
	＜tag R App anno=" "＞＜/tag＞	＜tag R Cor anno=" "＞＜/tag＞
	＜tag R Sum anno=" "＞＜/tag＞	
Misuse Patterns	＜tag N NE anno=" "＞＜/tag＞	＜tag N WR anno=" "＞＜/tag＞
	＜tag N CO anno=" "＞＜/tag＞	＜tag N SI anno=" "＞＜/tag＞
	＜tag N SL anno=" "＞＜/tag＞	＜tag N DI anno=" "＞＜/tag＞

The tag design include a pair of pointed brackets delimit the text it annotates. In the tag, there are four layers separated by space. The first layer uses the word *tag* to ratify the other words in the first bracket as supplemented linguistic information. The letters, Y, R, and N, on the second layer refer to written register use, spoken register use and misuse pattern. The third layer specifies which use of misuse of the enclosed CA is. The last layer, shown as anno=" ", allows researchers to supplement other information if necessary. The following is an illustrative example.

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

3. Corpus compiling and application with CA annotation

The learner corpus compiled in the present study is based on the OLAC Metadata Set via the developed Perl-based online interface, accessible at http://awta.csie.ncku.edu.tw/. In

total, 2290 pieces of English compositions by Chinese speakers, approximately one million words, are collected over three years. These compositions belong to 13 different genres, including *process*, *summary*, *essay question*, *cause-effect*, *comparison-contrast*, *definition*, *description*, *narration*, *classification*, *multiple strategies*, *argumentation*, *problem solving*, and *research article*.

Among the collected data, 65 pieces of writing, with 5 pieces a genre and 28941 words in total, are further selected for the investigation into the use and misuse of CAs. The selected data are annotated via the online tagger, as seen in Figure 1. Part A is the raw text, Part B shows the annotating system, and Part C presents how the raw text is annotated with CA information.

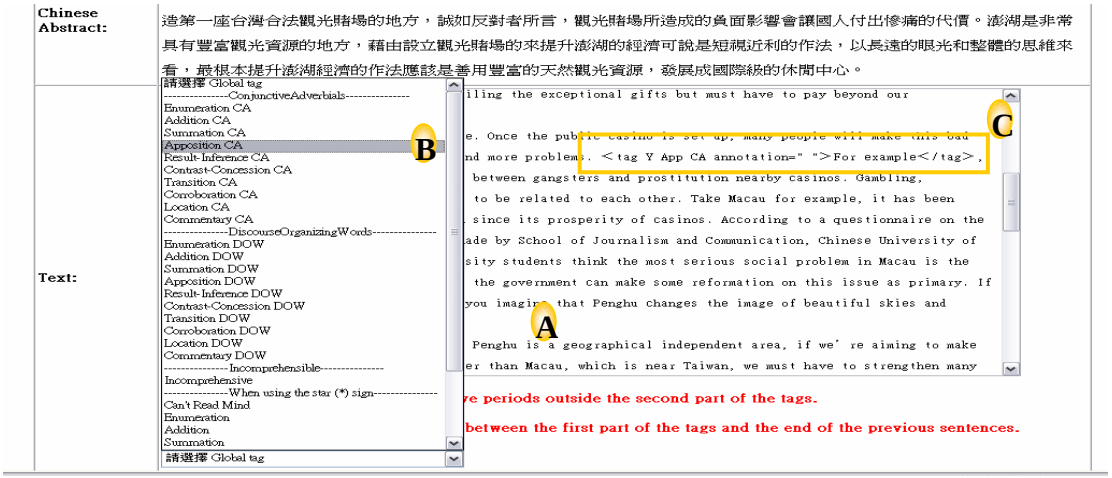

Figure 1. Tagger page

Each annotation of CA is made through a four-step procedure, shown in Figure 2. The first step identifies the CA. The second step is to judge whether or not the CA is correctly used. If the CA logically connects the context, the use of the CA is viewed as correct and the judgment goes to Yes. If not, the use is incorrect and the judgment goes to No. Lastly, if the use is correct yet the language form is stylistically improper, the judgment goes to Spoken Register. The third and last steps complete the annotation. If the judgment of the procedure goes to Yes or Spoken Register, then decide which textual relation the CA conveys, and select a tag from the bottom list. Likewise, if the judgment goes to No, select a misuse pattern tag from the bottom list to annotate the CA.

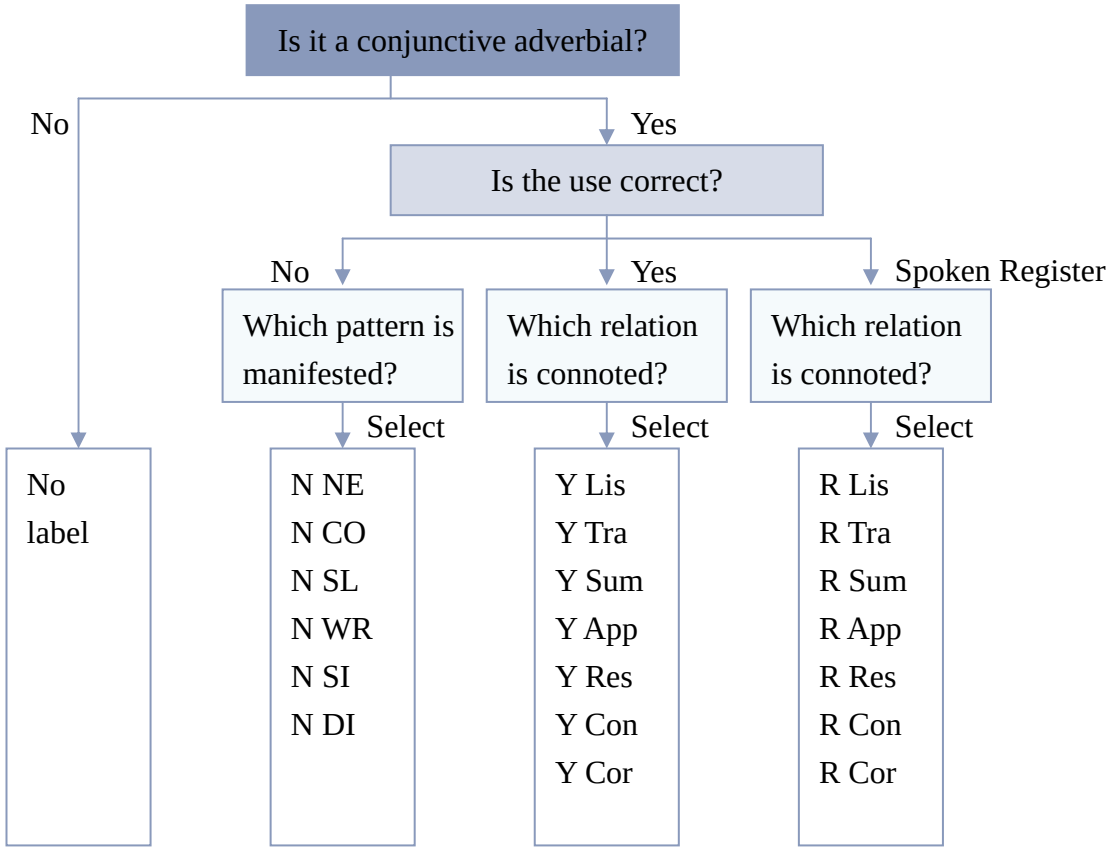

Figure 2. The annotating procedure

After annotating the selected data and tallying the counts, all the obtained figures were further analyzed via inferential statistical measurements on SPSS to investigate on the use and misuse distribution of CAs across genres.

To investigate the use distribution of CAs across genres, a two-way within-subjects analysis of variance (hereafter ANOVA) is designed, with two independent variables being textual relation and genre while the dependent variable is the counts of CAs. To further examine the effect of register, the ANOVA design would be calculated again, with the independent variable, textual relation, replaced with textual relation performed by CAs in written register. Lastly, to investigate the misuse distribution of CAs across genres, a two-way within-subjects ANOVA is employed again, with the two independent variables being misuse pattern and genre while the dependent variable is the counts of CAs. A significant level of $p<.05$ was chosen.

4. Results

In the investigation of the use distribution of CAs across genres, the raw counts of CAs show that regardless of genre difference, there is a tendency that the listing and contrastive relations are the two most frequently occurring types performed by CAs while the summative and transitional relations are the two least frequently occurring types. The rest of the textual relations are in the middle. In addition, the ANOVA analysis to examine the

effect of textual relation and genre shows 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 ($F(6, 24)=10.476$, $p<0.05$). Table 5 shows the statistic results.

Table 5. ANOVA results for the effect of textual relation and genre

Source	DF	SS	MS	F	P
Textual Relation	6	3005.681	500.947	10.476	0.000*
Genre	8	111.234	13.904	1.697	0.137
Textual Relation×Genre	48	64.977	1.354	1.070	0.366

*p<.05

In the follow-up investigation of the use distribution of written-register CAs across genres, the raw counts of written-register CAs show that regardless of genre difference, the listing and contrastive relations are the two most frequently occurring types performed by written-register CAs while the summative and transitional relations are the two least frequently occurring types. The rest of the textual relations are in the middle. After applying ANOVA analysis, as presented in Table 6, it is found that there is no interaction between textual relation performed by written-register CAs and genre ($F(48, 144)=0.969$, $p=0.537$). However, there does exist the main effect from textual relation ($F(6, 18)=8.585$, $p<0.05$). Meanwhile, Due to the scarce occurrence of spoken-register CAs, the row counts of spoken-register CAs are too small to decide the frequency order of occurrence and to run an ANOVA analysis.

Table 6. ANOVA results for the effect of textual relation via written-register CAs and genre

Source	DF	SS	MS	F	P
Textual Relation	6	1968.256	328.043	8.585	0.000*
Genre	8	99.374	12.422	2.062	0.082
Textual Relation×Genre	48	49.831	1.038	0.969	0.537

*p<.05

Lastly, in the examination of the misuse distribution of CAs across genres, ANOVA analysis shows, as seen in Table 7, that there is no interaction between misuse pattern and genre ($F(40, 160)=1.031$, $p=0.432$) as well as no main effect from genre ($F(8, 32)=1.857$, $p=0.102$) while there exists the main effect from misuse pattern ($F(5, 20)=3.210$, $p<0.05$). Although the raw counts of misuse patterns seem to show a norm distribution, it is just coincidence for most misuse patterns have no significant difference with others. However, some misuse patterns do differ from each other on a significant level. The misuse pattern, *Wrong Relation*, significantly differs from *Semantic Incompletion* and *Non-equivalent Exchange*, whereas *Surface Logicality* differs from *Conjunctive Overuse* and *Distraction*.

Table 7. ANOVA results for the effect of CA misuse and genre

Source	DF	SS	MS	F	P
CA Misuse	5	49.857	9.971	3.210	0.027*
Genre	8	25.239	3.155	1.857	0.102
Misuse Pattern×Genre	40	24.100	0.603	1.031	0.432

*p<.05

5. Discussion

The present study aims to achieve two goals. The first goal is to compile a learner corpus annotated with linguistic information on textual relation performed by CAs, which can serve as training data of developing technology that automatically diagnoses learner writing on the discoursal level. The second goal is to investigate genre influence on use and misuse of CAs and to construct a misuse framework for CAs.

In terms of the first goal, the compiled learner corpus fulfills the expectation. Researchers can use the annotated data as training data to develop automatically discourse-diagnosing technology and conduct pilot studies based on the rest of the corpus. Meanwhile, the annotated data are based on XML format, which bestows the data with great compatibility for all operating systems and extensibility to other possible alteration [17] for other application.

In terms of the second goal, it is found that, contrary to what most writing textbooks suggest that CAs performing certain textual relation are more prominent in certain genre, genre has no role in impacting the distribution of textual relations performed by CAs. In effect, the textual relations performed by CAs form a norm distribution regardless of genre difference. That is, *Listing* and *Contrastive* are the most frequent. *Resultive/Inferential*, *Appositive* and *Corroborative* are the second most frequent. *Summative* and *Transitional* are the least frequent, which corresponds to what is found in [9] [10] [11] as well as [12] [13] [14]. The same is true of the distribution norm performed by written-register CAs.

The lack of genre influence may result from the fact the genres are not mutually exclusive. That is, different genres may share many similar characteristics, which, in some sense, makes different genres one general superordinate genre without distinct differences. Consequently, a distribution norm of textual relations would be discovered, because the distribution norm of textual relations performed by CAs across genres, in fact, is the distribution of textual relations of the general superordinate genre.

To account for the formation of the distribution norm among textual relations performed by CAs, two explanations are proposed. One lies in the nature of different textual relations. For example, the transitional and summative relations occur least frequently at a significant level. This is understandable in that the two relations serve opening and closing functions which only appear at the beginning and at the end no matter how long a textual unit is. The other explanation is that there is a preference preset in human cognition for employing CAs to convey certain textual relations. Take the contrastive relation as example. The relation is relatively complicated because it requires the action to analyze two events and to locate the contrastive points, which would take more energy to describe the relation compared with writing in the common temporal sequence. Due to the extra energy required, Economy Principle, to minimize the energy consumption [18], is applied in human cognition, which is to use CAs to convey the contrastive relation explicitly, rather than describe the relation in context. Ultimately, the contrastive relation becomes one of the textual relations most frequently performed by CAs.

Lastly, although no genre influence on the misuse patterns of CAs is found, nor is a distribution norm of CA misuse, the proposed misuse framework is proved meaningful in differentiating CA misuse patterns. According to ANOVA analysis, *Wrong Relation* significantly differs from *Non-equivalent Exchange* and *Semantic Incompletion*, while *Surface Logicality* from *Connective Overuse* and *Distraction*. The results suggest that the causes of these misuse patterns are fundamentally different, and that the distinction and recognition of them are necessary. Also, the six misuse patterns can be divided into two groups based on the significant difference among them, with one group being *Wrong Relation, Non-equivalent Exchange*, *Semantic Incompletion*, while the other group being *Surface Logicality*, *Connective Overuse* and *Distraction*. To explain the division, the principles of appropriateness and necessity are proposed. The former group refers to the situation in which the use of the CA is required to signify the textual relation between sentences but the use is not correct, or inappropriate. In contrast, the latter group refers to the situation in which the use of the CA is not necessary and sentences themselves can form a unit of text with the CA.

6. Conclusion

The present study contributes in three aspects. First, a learner corpus annotated with

textual relations via CAs is compiled, which can serve as training data for developing technology that automatically diagnoses learner writing on the discoursal level. Second, it is found that genre difference plays no role in impacting either textual relations via CAs or the misuse of CAs, and that there exists a norm distribution of textual relations performed by CAs across genres. The found norm distribution can be used to examine whether or not a piece of learner writing conforms to proper discoursal organization. Deviation from the norm distribution may be a signal, suggesting learners to re-organize their text. Third, the proposed misuse framework can help learners locate the misuse of CAs, and provide direction for correction by evaluating whether the misuse is inappropriate or not necessary.

Nevertheless, there is still room for further research. For a starter, the annotated data only account for a small amount of the compiled corpus. More data are expected to be annotated in the future, which can further validate the study and provide more training data to develop automatized technology. Moreover, although no genre influence is found in textual relations performed by CAs or in the misuse of CAs, as the anonymous reviewer suggests, the results may be still subject to other factors, such as age, educational background, English proficiency, or even L1 transfer. If the interaction between the use of CA and these factors can be made clear in future studies, non-native writers can receive a different angle in terms of learning CA use and organizing their English writing.

References

[1] H. H. Shih, Compiling Taiwanese Learner Corpus of English. *Computational Linguistics and Chinese Language Processing, 5*(2), 87-100, 2000.

[2] S. C. Gui, and H. Z. Yang, *Chinese Learner English Corpus*, Shanghai: Shanghai Foreign Language Education Press, 2003.

[3] S. Granger, International Corpus of Learner English - ICLE. Retrieved November 23, 2008, from http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.htm#heading1

[4] M. Connelly, *Get writing: Sentences and paragraphs*, Australia: Thomson Higher Education, 2006.

[5] J. M. Lannon, *The writing process: A concise rhetoric, reader, and handbook*, New York: Longman, 2007.

[6] M. Morenberg, and J. Sommers, *The writer's options: Lessons in style and arrangement*, New York: Pearson Longman, 2008.

[7] J. M. Reid, *The process of composition*, White Plains, NY: Longman, 2000.

[8] R. L. Smalley, M. K. Ruetten, and J. R. Kozyrev, *Refining composition skills: Rhetoric and grammar*, Boston: Heinle & Heinle, 2001.

[9] Y. Field, and L. Yip, A comparison of Internal conjunctive cohesion in the English essay writing of Cantonese speakers and native speakers of English. *RELC Journal,*

23, 15-28, 1992.

[10] M. Liu, and G. Braine, Cohesive features in argumentative writing produced by Chinese undergraduates. *System, 33*, 623-636, 2005.

[11] W. Y. C. Chen, The use of conjunctive adverbials in the academic papers of advanced Taiwanese EFL learners. *International Journal of Corpus Linguistics, 11*(1), 113-130, 2006.

[12] G. Tankó, The use of adverbial connectors in Hungarian university students' argumentative essays. In J. M. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (pp. 157-181). Amsterdam: John Benjamins B. V, 2004.

[13] B. Altenberg, and M. Tapper, The use of adverbial connectors in advanced Swedish learners' written English. In S. Granger (Ed.), *Learner English on Computer* (pp. 80-93). London: Longman, 1998.

[14] T. C. Shen, *Advanced EFL Learners' Use of Conjunctive Adverbials in Academic Writing*. MA Thesis, Taiwan: National Taiwan Normal University, 2006.

[15] W. J. Crewe, The illogic of logical connectives. *ELT Journal, 44*(4), 316-325, 1990.

[16] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik, *A Comprehensive Grammar of the English Language*, London: Longman, 1985.

[17] A. Møller, and M. I. Schwartzbach, *An introduction to XML and Web technologies*, New York: Addison-Wesley, 2006.

[18] F. Ungerer, and H. J. Schmid, *An Introduction To Cognitive Linguistics*, UK: Pearson Education Limited, 2006.