The Dynamics of English Writing Development in Advanced Chinese Learners

Hou, Junping

Chapter 5

Coherence and cohesion: Development over time in advanced Chinese L2 learners of English

5.1 Introduction

In the study of Chapter 3, two texts at the beginning and end of an 18 month course of an advanced group of Chinese L2 learners were compared. Despite having spent at least four hours a week on their English skills, they did not improve on holistic scores nor on any typical linguistic complexity measures that show development such as sentence length (Wolfe-Quintero et al., 1998, Ortega, 2003, Lu, 2011). Because it was difficult to believe that these students did not improve at all, we assumed that the students might have reached a ceiling effect in these typical linguistic complexity measures and decided to explore whether improvement had taken place in other areas. A subsequent study in Chapter 4 (also see Hou, Loerts & Verspoor, 2016) showed that these students indeed improved over time in chunk coverage (the relative number of words in chunks divided by the number of words in the text) and one particular kind of chunk, collocations. In this chapter, we will go beyond purely linguistic measures and explore whether the students developed in aspects of text quality: coherence and cohesion. As the background literature section will show, both coherence and cohesion play a role in text quality, but how the two interact is not clear. One problem is that coherence is more difficult to operationalize than cohesion as coherence is an implicit and cohesion an explicit phenomenon. In this chapter, we will compare the texts written by a group of advanced Chinese L2 learners of English at the beginning of the course with those written at the end and explore whether there are differences in coherence and cohesion measures.

The main research questions are which coherence and cohesion measures correlate well with holistic proficiency scores and whether these advanced students of L2 English improve in text quality, operationalized as coherence and cohesion, over their 18 month course. The texts were holistically scored on general proficiency measures and then submitted to a hand-coded topic-based analysis based on Watson Todd, Thienpermpool and Keyuravong (2004) supposedly tapping into the construct of coherence. After that the texts were analyzed by means of an automated tool, Coh-Metrix, for 30 local, global and text cohesion measures (cf. Crossley, Kyle & McNamara, 2016a). To see to what extent these coherence and cohesion measures relate to proficiency, correlation tests were run. To measure to what extent the students changed over time, pre-post statistical analyses with all measures were conducted.

5.2 Text quality, cohesion and coherence

Text quality depends to some extent on whether the reader can easily make sense of it. Generally speaking, a reader's understanding of the meaning and quality of a text relies on the progression or connectedness of ideas in the discourse. A well-organized discourse structure is characterized by 'logical sequencing' (Weigle, 2002, p.116) with 'clear and consistent evidence of the ability to produce organized coherent and cohesive discourse' (Weir, 1990, p. 172). The connectedness refers to 'all of the links, both explicit and implicit, in a text that make it a unified whole, which is usually divided into cohesion and coherence' (Watson Todd, Khongput, & Darasawang, 2007, p.11). However, coherence between sentences is 'based not only on the sequential relation between expressed and interpolated propositions, but also on the topic of discourse of a particular passage' (Van Dijk, 1977, p. 93). In addition, Brown and Yule (1983) suggest that coherence depends primarily on the interpretation of linguistic messages and that coherence resides in 'how people interpret texts rather than in the text themselves' (Yule, 1996). Not surprisingly, McNamara, Graesser, McCarthy and Cai (2014), refer to coherence as 'a psychological construct' (p.194). Crossley et al. (2016a) summarize the literature by saying that coherence refers to the understanding that the reader derives from the text and depends on a number of factors including cohesion cues and non-linguistic factors such as prior knowledge and reading skill (McNamara, Kintsch, Songer & Kintsch, 1996; O'Reilly & McNamara, 2007).

While coherence refers to the general semantic relationships between sentences that are necessary for the reader to make sense of the entire text, cohesion refers more to explicit links between parts of the text by means of linguistic devices. Cohesion '[…] concerns the ways in which the components of the surface text, i.e., the actual words we hear or see, are mutually connected within a sequence' (De Beaugrande & Dressler 1981, p.3) and may be at the local, global and text level. At the local level, cohesion cues are explicit and include connectives such as *because, therefore*, and *consequently* (Halliday & Hasan, 1976) and overlapping words and concepts between sentences. At the global level, cohesion cues are more implicit and include semantic and lexical overlap between paragraphs in a text (Foltz, 2007). At the text level, cohesion cues are also

more implicit and based on, for example, 'givenness in which cohesion is measured across the text based on the number of words that are new (e.g., an initial noun referent) or given (noun referents that can be referred to pronominally)' (p. 2).

Most scholars would agree that a well-written text must have both coherence and cohesion, but how these aspects and their interrelationship can be operationalized is not clear. Carrell (1982) contends that cohesion does not lead to coherence and coherence by itself does not suffice to make a text coherent, as there must be some additional linguistic property (like cohesion) to make a text coherent. A passage can be coherent but may lack sufficient inter-sentence cohesion. In such a case, the reader is probably still able to get a general understanding of the meaning of the text. On the other hand, if a passage lacks coherence and thus continuity of meaning in situation and topic, the reader needs to use an internal representation of the topic and situation (e.g., cultural background) to obtain an effective judgment on the meaning of the text. Judgments on writing quality must thus concern both cohesion and overall coherence in content and organization.

Despite the fact that coherence is difficult to operationalize and may remain subjective, several researchers studied the quality of coherence in L2 writings by analyzing topical coherence in passages under the assumption that an essay that stays focused on a topic is more coherent than one that does not. In their study of exploring the differences in discourse competence level between English learners and German as foreign languages, and their influence on overall text quality, Medve and Takač (2013) found that 'the high rated essays contain a greater proportion of sequential progression than the medium- and low-rated essays ….the use of a greater number of sequential progressions enhances the text quality' (p.128). Somasundaran, Burstein and Chodorow (2014) found that coherence elements are found in adherence to the essay topic, elaboration, use of various vocabulary items and a sound organization of thought and ideas. One topic-based analysis method that meets most criteria in coherence quality analysis, i.e., the objectivity of the measures, the equivocality of the measures, and the typology of the measures, was developed by Watson Todd (1998, 2003). His original work (1998, 2003) on topic-based analysis used spoken classroom discourse, but later Watson Todd et al. (2004) applied the method to written discourse. The framework focuses on key concept frequency and their

relationship, which highly correlates with the propositional coherence of typology of the measures (Stubbs, 1983). Three relatively objective measures of topic-based coherence are average distance of the moves in a text, percentage of coherence breaks and the number of moves / per 10 T-units. Watson Todd et al. (2004) compared these measures with the teachers' marks of coherence on 28 texts written by Thai L2 writers of English and found that the numbers of moves between key-concepts per 10 T-units correlated with the teacher's holistic marks.

Few studies have explored the relation between objectively established coherence measures and cohesion. In most studies exploring the relationship between coherence and cohesion, coherence is established by human raters, the most informative predictor of text quality according to Crossley and McNamara (2010). Some of the findings are that more sophisticated L1 English writers use fewer cohesive devices than their less sophisticated peers, but this does not necessarily hold true for L2 writers. One common finding for L1 writers is that compositions that score high holistically contain more lexical (as opposed to referential and conjunctive) cohesive devices than those with lower scores (Hartnett, 1980; Witte & Faigley, 1981; Lieber, 1980; Anderson, 1980; Bereiter & Scardamalia, 1987; Hayes & Flower, 1980). McCulley (1985) investigated the connection between cohesion and writing quality in an analysis of 120 L1 argumentative essays composed by high school students with a coherence rating scale of the National Assessment of Educational Progress (NAEP) in 1978-1979. It was found that writing quality did not correlate with the total number of cohesive ties used in the essays, but specific cohesive ties (e.g., demonstratives, nominal substitution and repetition) contributed to the positive assessment of writing quality, suggesting that primarily lexical cohesive devices made a more important contribution to coherence. Fitzgerald and Spiegel (1986) examined L1 students' writings from two grade levels and investigated the degree to which coherence and cohesion relate with text quality and grade levels. Their findings suggest that the relationship between coherence and cohesion in children's writing varied according to text content but not to grade level. Crossley and McNamara (2010) found no evidence that cohesion cues and essay quality were related, but they rather found that linguistic sophistication characterized the essays that were rated higher. Tierney and Mosenthal

(1983) analyzed the correlation between holistic coherence scores and the number of cohesive ties used in two rhetoric class compositions, written by 12 year old writers of English and found no significant interaction effect regarding the use of cohesive devices. Although a significant interaction was gained for coherence rankings, cohesion analysis was considered to be a poor index of coherence or writing quality. However, the general finding was that good essays are not necessarily more cohesive than the weak ones, a finding supported by several other studies (cf. Lautamatti, 1990; Yule, 1996; Dueraman, 2007).

There are also some puzzling differences between L1 and L2 writers. In a comparison between L1 and advanced L2 writers, Connor (1984) found no difference between the L1 and L2 writers in cohesive density (reference or conjunction), but the L2 texts were not necessarily coherent. This is in line with Lindsay (1985), who found more instances of conjunctive and referential cohesion in essays that were rated lower in quality than those rated higher in quality for the L2 writers, but this was not the case for L1 writers of English. Scarcella (1984) found no overall difference in the use of lexical and referential cohesion compositions written by L1 and L2 writers of English, but weaker L1 writers used significantly more conjunctive cohesion than stronger L1 writers; this was not the case for L2 writers. Johnson (1992) carried out a study to investigate three types of cohesion categories (reference, conjunction and lexical cohesion) in good and weak essays written in L1 Malay, L1 English, and in L2 English by Malaysian writers. The conclusion was that good essays written in L1 Malay had more semantic ties through reiteration of words than in weak essays. In contrast, good essays in L1 English had more syntactic ties (conjunction and reference).

Several studies have examined different groups of L2 writers. In their L2 study on argumentative essays written by Chinese writers of L2 English, Yang and Sun (2012) found that more advanced learners used a greater number of cohesive devices and used them more accurately, and the use of cohesive devices strongly correlated with essay quality. In contrast, Guo, Crossley and McNamara (2013) reported that indices of local cohesion and text cohesion were negatively correlated with judgments of essay quality for essays written independently, but local cohesive indices were positively correlated with essays that were based on a source text. Crossley et al. (2016)

reported differences between local and global cohesive devices and their relation to L2 writing quality: local cohesion related negatively and global cohesion related positively with L2 writing quality. Especially for more advanced levels, the reliance on local cohesive devices is less frequent (Crossley & McNamara, 2011; Crossley, Roscoe & McNamara, 2011).

Crossley et al. (2016) explored how the use of cohesive devices (selected from both Coh-Metrix and TAACO tools) were related to human ratings and how the production of these devices changed over time in L2 writers. They found that several indices of cohesion were related to human judgments of text quality and that growth occurred for a number of cohesive devices. However, the growth did not occur in those devices that correlated with human ratings of text quality.

Few studies have thus far looked at how objectively established coherence and cohesion develops over time and how these measures are related to other L2 proficiency criteria. The current study will address this gap by investigating the relationship between human holistic ratings of L2 proficiency, the use of topic-based coherence measures as developed by Watson Todd et al. (2004), the use of cohesive devices (selected indices of Coh-Metrix too), and the development of these over 18 months. The main questions are:

(1) To what extent do coherence and cohesion measures relate to L2 proficiency?

(2) Are there significant changes in the use of topic-based coherence measures over time?

(3) Are there significant changes in the use of cohesion devices over time?

5.3 Method

This study concerns one pre and one posttest texts written by per 18 advanced Chinese learners of English as L2. In the previous study of Chapter 3, it was found that they had not improved in overall English proficiency nor in any typical linguistic complexity measures that show development. In this study, the holistically established proficiency scores are compared with topic-based coherence measures and with the use of cohesive devices as established by Coh-Metrix.

5.3.1 Participants

The participants consisted of 18 Chinese learners of English (average age 18) who were enrolled in the most advanced course for non-majors in English. As mentioned in Chapter 3, this group had already been admitted to a highly selective and prestigious university and had subsequently taken a proficiency exam that placed them in the higher-level class, aimed at passing the CET-6 exam (the highest level exam for non-majors). They were enrolled in an 18-month English program at the university.

5.3.2 Intervention

The intervention studied in the present chapter is identical to the one reported in Chapter 3. The students were exposed to English input with professional teachers with a high L2 proficiency and materials and assignments appropriate for their level as part of their training for the CET-6 exam in an 18-month course. Each week they were exposed to English for about 4-5 hours (totaling about 196 hours). The teachers at this university apply a learning-by-writing theory and spend time on writing practice following CET directions. They use and discuss model essays, sometimes those written by class members, in class to help the students get an idea of what constitutes a 'good essay' with respect to structure, content and language. In addition to grades, the teachers give rather detailed feedback on errors as well as brief evaluations and suggestions for improvement. Sometimes students are asked to give each other peer feedback.

The assignments followed the demands and purposes of written exercises as described in each curriculum unit, and students were asked to practice what they had learned in that specific unit following the specific structure of an argumentative essay. The constructs of coherence and cohesion were addressed in some of the units during the 18 months.

5.3.3 Texts

The two essays involved in the current study were both argumentative in nature. The pretest text concerned 'Diploma and Success / Knowledge'. The posttest text concerned 'The Pros and Cons of Exercise' (sample pre-posttest texts are shown in Appendix 7). There was no word limit and text length varied from 144 to 388 words. There was no time pressure, and the texts were written after class with possible access to dictionaries and other

resources, such as the internet. The learners were not aware that their texts would be examined for cohesion or coherence.

5.3.4 Holistic proficiency scores

In the study of Chapter 3, as part of the larger study, four texts (the first two and the last two) were holistically rated on general proficiency using a rubric consisting of 5 general indices of L2 proficiency--complexity, accuracy, fluency, idiomaticity and coherence (CAFIC). For scoring coherence, the following instructions were given:

> *Read the text as a whole without paying attention to the items in the rubric, but pay attention to 'flow'. Do sentences have a natural flow so that you can understand what is being said without having to reread sentences? Is there a focus and are all sentences in a paragraph related? If there are more paragraphs, are they connected?*

The 36 texts used in the present study were rated together with 160 other texts of learners of different proficiency levels after a two-hour training session with the rubric (details see Chapter 3). All 196 texts were randomly mixed and holistically rated by 8 trained raters (5 L1 and 3 L2 speakers of English) in groups of two. The raters were not aware of the background or purposes of the study.

Each text was scored for each of the five CAFIC rubrics on a Likert scale from 1 to 5, with a possible total CAFIC score of 25 points. As reported in Chapter 3, an excellent interrater reliability was found for the overall CAFIC score (ICC(1,3) = .799) and high reliability scores were obtained for the sub-scores complexity, fluency, idiomaticity, and coherence (all ICCs(1,3) > .636). The present study will concern only one of the pre and one of the posttest texts because the coding for coherence (see next section) and cohesion is extremely time consuming.

5.3.5 Topic-based coherence coding

In coding for topic-based coherence, we followed Watson Todd et al. (2004) precisely. The following is a brief summary of the steps involved in coding and includes illustrative tables, figures and examples taken from a sample

text written by one of the participants.

1. The number of T-units was established and marked with a dash '/'. The sample text contains 20 T-units.
2. The key-concepts were identified. Key concepts refer to nouns or noun-phrases that occur at least twice or are otherwise salient. Because of their saliency, titles are key concepts even if they do not occur twice. Function words and textual phrases such as 'people (think)', 'some persons (may agree)' are not considered key concepts. The sample text contains seven key concepts that are marked with superscript numbers from '1' to '7': *diploma and success, college, diploma, a higher diploma, success, job/work, skills* (Figure 5-1).
3. Within the text, each 'move' between two consecutive key-concepts is given a value, a '1' for a direct semantic link (such as a superordinate-subordinate relation such as *diploma* and *a higher diploma* or a cause-effect relation such as *college* and *diploma*) in the hierarchy or a '2' for no direct link. The flow of the moves between key-concepts are mapped and numbered according to their order of appearance in the text (see Figure 5-2). An immediate repetition of the same concept means no moves. Back and forth repeated moves between concepts are mapped and numbered separately (those direct-linked concepts are marked with dashed-lines and those no-direct-linked concepts are marked with dotted lines in Figure 5-2).
4. Two measures are calculated. The strength of relationships between consecutive key concepts, operationalized as the average distance of moves, is calculated by dividing the sum of all values '1' and '2' by the total number of moves. The lowest average distance in a whole text may be '1' and the highest may be '2'; hence, a lower average distance would suggest a more coherent text. We will refer to this measure as Strength of Moves (SoM) from now on (Table 5-1). The density of key concepts, operationalized as the number of moves per 10 T-units, is calculated by dividing the total number of moves by the total number of T-units and then multiplied by 10. Fewer moves per 10 T-units suggests a more coherent text in that the writer is

producing fewer different key-concepts or more repetitions than those with more moves /10 T-units. We will refer to this measure as Density of Moves (DoM) from now on (Table 5-2).

Each text was analyzed by the first author and discussed in detail with the other authors until agreement was reached on all key concepts, number of moves and types of moves.

Sample text

Diploma and success[1]. /

Nowadays, it seems necessary for everyone to go to college[2] to further their study and of course to get the diploma[3] when graduation. /

So they work in high gear to get access to better college[2] and have a higher diploma[4]. / All that they do are for the sake of success[5]. / They want to be successful, but is the diploma[3] equal to the success[5]? /

There are two points. / Someone thinks so. / The higher their diplomas[3] are, the more successful they will be. / They consider the government and many companies think highly of the diploma[3]. / The interviewer evaluates a person mostly by his diploma[3] and it stands for one's value. / So only when they have a good-looking diploma[4], can they get a great job[6] to achieve their emotions? / However, as a matter of fact, a high diploma[4] has already showed their success[5]. /

Then on the other hand, some others don't agree with this. / In their opinion, a person's success[5] depends on his all kinds of skills[7], multitasking ability and adapting skills[7]. / A successful person should be able to make full use of what he learns and transform his knowledge to something useful that could contribute to his work[6]. /

Moreover, it needs many valuable spirits like passion, patience and never giving up leading to success[5]. / So only a diploma[3] is not for more enough. /

Well, it' s time to put forward my point. / In my thought, the diploma[3] can provide a chance, a platform for you. / So it's necessary, and then as the second point, we should use all our intelligence and power to claw our way to the success[5]. /

Table 5-1. Calculation of average distance of Strength of Moves in sample text.

Values	1	2	Calculation	**SoM**
Moves	7	10	((7*1) + (10*2)) /17	**1.59**

Table 5-2. Calculation of Density of Moves in sample text.

T-units	Total number of moves	Calculation	**DoM**
20	17	(17/20)*10	**8.5**

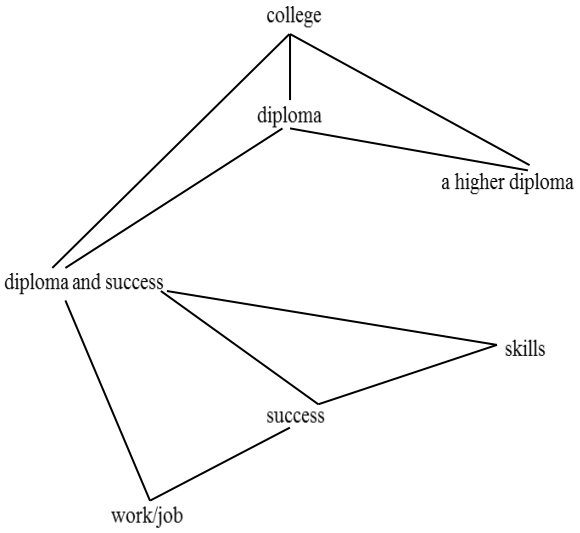

Figure 5-1. A hierarchy of the key concepts for the sample text.

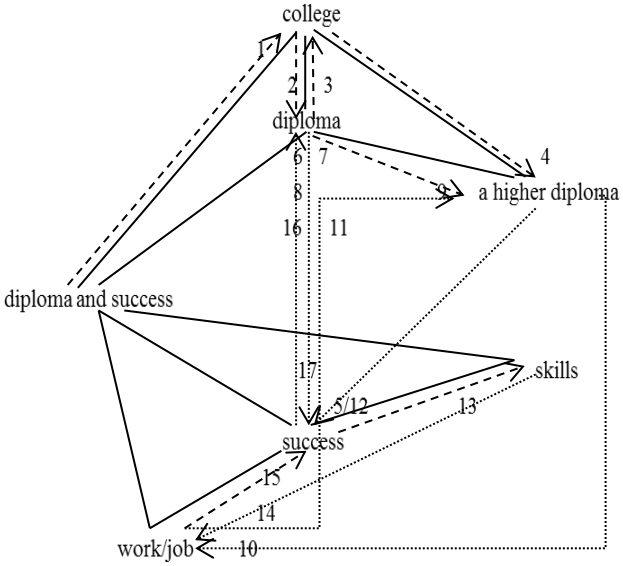

Figure 5-2. Moves between concepts mapped onto the hierarchy for the sample text.

5.3.6 Coh-Metrix analysis

Coh-Metrix is a computational tool that 'analyzes texts on multiple measures of language and discourse that are aligned with multilevel theoretical frameworks of comprehension' (Graesser, McNamara & Kulikowich, 2011, p. 223). It analyzes texts on over 50 types of cohesion relations and over 200 measures of language, text, and readability. The version (3.0) involved in the current study accounts for the variance in texts across grade levels and text categories and includes 11 major factors and 106 sub measures. Like Crossley et al. (2016a), however, the current paper will focus only on 6 features of local, global and text cohesion (with 30 sub-indices): referential cohesion, LSA (Latent Semantic Analysis), lexical diversity, connectives, situation model, syntactic pattern density.

Referential cohesion refers to overlap in content words between local sentences, or co-reference, which can aid readers in making connections between propositions, clauses, and sentences. Local cohesion is measured by assessing the overlap between consecutive, adjacent sentences, whereas global cohesion is assessed by

measuring the overlap between all of the sentences in a paragraph or text. The assumption is that greater referential cohesion is easier to process.

LSA, short for Latent Semantic Analysis (Landauer, McNamara, Dennis & Kintsch, 2013), provides measures of semantic overlap between sentences or between paragraphs. The assumption is that greater overlap is easier to process.

Lexical diversity refers to the variety of unique words (types) that occur in a text in relation to the total number of words (tokens). A high number of different words in a text indicates that new words need to be integrated into the discourse context (Graesser et al., 2011, p. 226). The assumption is that cohesion is higher when lexical diversity is lower and fewer words need to be integrated.

Connectives are words that explicitly state the connection between propositions. There are five general classes: causal (*because, so*), logical (*and, or*), adversative/contrastive (*although, whereas*), temporal (*first, until*), and additive, linking the ideas and clauses positively (*and, moreover*) or negatively (*however, but*). The assumption is connectives make a text easier to process.

Situational model applies to 'deeper cohesion features related to a text's temporal, causal, and spatial features (along with the intentionality of a texts) related to a reader's understanding of the text (Crossley et al., 2016a, p. 12). The assumption is that discontinuities will make the text more difficult to process (Zwaan, Magliano & Graesser, 1995; Zwaan & Radvansky, 1998).

Syntactic pattern density covers the density of word and phrase types. The assumption is that a relatively stronger density will be more difficult to process.

5.3.7 Design and analyses

To see to what extent topic-based coherence measures and cohesion measures correlate to general proficiency measures, correlation analyses were carried out between them and the CAFIC measures and length of text (also a great predictor of proficiency).

To see if any change had occurred within 18 months, paired sample *t*-tests were conducted to compare topic-based coherence measures and

cohesion measures in the first text with those of the last text. Non-parametric analyses were conducted for all variables that did not follow a normal distribution.

5.4 Results

5.4.1 Proficiency scores

In the current study, only one pre and one posttest text (rather than the first and last two texts as in the previous studies) written by the participants were used. In contrast to the findings in the previous study, some small differences were found between pre- and posttest in CAFIC sub-scores. As these data did not follow a normal distribution (*SW*s < .920; *df* = 36; *p* < .012), Wilcoxon Signed-Rank Tests were used and *Z*-scores are provided instead of *t*-values (See Table 5-3). All CAFIC measures and text length, had increased at the posttest but only accuracy showed a trend towards a significant increase ($Z(17) = -1.916$; $p = .055$).

Table 5-3. Differences between the use of CAFIC between pre- and posttest writings.

	***Z* statistic**	***p*-value**
Complexity	-.571[b]	.568
Accuracy	-1.916[b]	.055
Fluency	-1.638[b]	.101
Idiomaticity	-1.477[b]	.140
Coherence	-.281[b]	.779
CAFIC	-1.220[b]	.223
Length of text	-1.677[b]	.094

b: based on negative ranks;

*. Significant at $p < 0.05$ (2-tailed).

5.4.2 Topic based coherence measures

Table 5-4 presents the means and medians of the two topic based coherence measures at pre- and posttest times: the SoM ([f = (number of 1-value moves*1 + number of 2-value moves*2)/the total number of moves]) and the Dom (f = the total number of moves / the number of T-units *10).

Table 5-4. Descriptive of the use of topic-based coherence (pre-posttest).

		Number of moves valued '1'	Number of moves valued '2'	Number of T-units	Total number of moves	**SoM**	**DoM**
Pre	Mean	14.83	3.11	12.61	17.94	1.17	14.76
	Median	14.00	3.00	13.50	18.00	1.17	14.94
	SD	5.24	1.99	3.07	6.06	0.11	5.44
Post	Mean	15.61	2.61	15.67	18.22	1.11	11.14
	Median	12.50	1.50	15.00	16.00	1.09	10.94
	SD	8.95	3.24	4.23	10.93	0.11	4.52

Table 5-5 shows correlations between the CAFIC scores, length of text, SoM, and DoM for all 36 texts (pre and posttest) taken together.

Table 5-5. Correlations between CAFIC scores, coherence measures and text length of all 36 texts.

	Holistic-rating proficiency measures					**Topic-based Coherence measures**		
Measures	Accuracy[#]	Fluency	Idiomaticity	Coherence[#]	CAFIC[#]	SoM[#]	DoM	length of text[#]
Complexity	.468**	.660**	.581**	.701**	.808**	.117	.411*	.609**
Accuracy[#]	1	.374*	.750**	.671**	.719**	.129	.406*	.309^
Fluency		1	.478**	.597**	.754**	.087	.330*	.899**
Idiomaticity			1	.690**	.763**	.143	.400*	.432**
Coherence[#]				1	.909**	.187	.477**	.533**
CAFIC[#]					1	.197	.521**	.728**
SoM[#]						1	.327^	.215
DoM							1	.362*
length of text[#]								1

[#] As the data did not follow a normal distribution (*SW*s < .940, *df* = 36, *p*s < .049), Spearman's rho was used.

**. Correlation is significant at the 0.01 level (2-tailed); *. Correlation is significant at the 0.05 level (2-tailed);

^ Correlation failed to reach significance $p < 0.07$ (2-tailed).

As what we found in the study of Chapter 3, the CAFIC sub-measures were highly and significantly positively correlated with each other and with the total CAFIC score. SoM showed very weak correlations with other measures, but there was a trend towards a positive correlation between SoM and DoM ($r = .327$; $p = .051$). DoM correlated with all CAFIC measures and the length of text (all *r*s > .330; all *p*s < .05). Length of text showed strong positive correlations with all CAFIC measures (all *r*s > .432; all *p*s < .01), but only a weak relationship was found with accuracy ($r = .309$; $p = .066$).

Length of text did positively correlate with DoM (r = .362; p < .05).

Tables 5-6 and 5-7 show the same correlations as in Table 5-5, but for the pre- and posttest texts separately.

Table 5-6. Correlations between CAFIC scores, coherence measures and text length of the 18 pretest texts

	Holistic-rating proficiency measures[#]					Topic-based Coherence measures		
Measures	accuracy	Fluency	idiomaticity	coherence	CAFIC	SoM	DoM	length of text
complexity	.286	.641**	.579*	.566*	.751**	.013	.459^	.709**
accuracy	1	.244	.692**	.746**	.672**	-.022	.367	.074
fluency		1	.483*	.579*	.752**	-.119	.225	.799**
idiomaticity			1	.697**	.753**	.063	.658**	.393
coherence				1	.908**	-.033	.495*	.450^
CAFIC					1	-.004	.559*	.693**
SoM						1	-.085	.171
DoM							1	.187
Length of text								1

[#] As CAFIC data did not follow a normal distribution (*SW*s < .889, *df* = 18, *p*s < .036), Spearman's rho was used.

**. Correlation is significant at the 0.01level (2-tailed); *. Correlation is significant at the 0.05 level (2-tailed).

^ Correlation failed to reach significance p < 0.07 (2-tailed).

Table 5-7. Correlations between CAFIC scores, coherence measures and text length of the 18 posttest texts

	Holistic-rating proficiency measures					Topic-based Coherence measures		
Measures	Accuracy[#]	Fluency	Idiomaticity	coherence	CAFIC	SoM[#]	DoM	length of text[#]
Complexity[#]	.617**	.604**	.544*	.812**	.882**	.371	.541*	.557*
Accuracy[#]	1	.512*	.785**	.649**	.784**	.345	.560*	.506*
fluency		1	.417^	.718**	.824**	.341	.731**	.950**
idiomaticity			1	.711**	.761**	.306	.445^	.473*
coherence				1	.933**	.395	.653*	.615**
CAFIC					1	.462^	.732**	.752**
SoM[#]						1	.593**	.385
DoM							1	.717**
length of text[#]								1

[#] As these data did not follow a normal distribution (*SW*s < .894; *df* = 18; *p*s < .045), Spearman's rho was used.

**. Correlation is significant at the 0.01level (2-tailed); *. Correlation is significant at the 0.05 level (2-tailed);

^ Correlation failed to reach significance p < 0.09 (2-tailed).

A comparison of Table 5-6 and 5-7 shows that on the whole there are more and stronger correlations at the posttest. SoM does not correlate with many CAFIC measures in either set. DoM showed several positive relationships with CAFIC scores at the pretest, but more so at the posttest.

Table 5-8 shows the changes in SoM and DoM over 18 months; both decrease over time.

Table 5-8. Differences between the use of topic-based coherence between pre- and posttest writings.

	Z / T	*p*-value
SoM[#]	-1.448[b]	.148
DoM	-2.466	.025*

[#] As the data did not follow a normal distribution (*SW*= .927, *df* = 18, *p* = .021), Wilcoxon Signed-Rank Test was used and the *Z*-Statistic is provided instead of T-value;

b: based on negative ranks.

* Significant at $p < 0.05$ (2-tailed).

A Wilcoxon Signed Ranks test showed that the decrease of SoM at posttest time (*Mdn* = 1.091; *SD* = .112) as compared to SoM at pretest time (*Mdn* = 1.170; *SD* = .110) was not significant. A paired sample *t*-test showed that DoM was significantly lower at posttest time (*M* = 11.139; *SD* = 4.516) than at pretest time (*M* = 14.761; *SD* = 5.435), ($t(17) = -2.47$; $p < .05$).

5.4.3 Coh-Metrix measures

Since only DoM showed strong correlations with holistic ratings and indicates an increase in coherence over time, our subsequent analyses did not include SoM. A correlation analysis between the CAFIC measures, DoM, and the six Coh-Metrix categories was carried out. Nonparametric correlations showed that 10 indices from 3 categories (LSA, connectives and situation model) reached significance or showed trends with CAFIC, human-rated coherence and DoM (Table 5-9).

Table 5-9 shows that CAFIC correlates negatively with LSASS1d ($r = -.390$, $p < .05$) and positively with CNCCaus ($r = .339$, $p < .05$), and shows a trend of a positive relationship with LSAGNd ($r = .321$, $p = .057$). Coherence also correlates negatively with LSASS1d ($r = .343$, $p < .05$) and

Table 5-9. Correlations between CAFIC measures, number of moves /10 T-units, and Coh-Metrix measures.

Categories	Full Descriptions of measures	**Labels of Measures**	**CAFIC**	**Coherence**	**DoM**
LSA	LSA overlap, adjacent sentences, mean	**LSASS1**			-.280^
	LSA overlap, adjacent sentences, standard deviation	**LSASS1d**[#]	-.390*	-.343*	
	LSA given/new, sentences, standard deviation	**LSAGNd**	.321^		
Connectives	All connectives incidence	**CNCAII**		.297^	
	Causal connectives incidence	**CNCCaus**	.339*	.347*	
	Adversative and contrastive connectives incidence	**CNCADC**			.391*
Situation Model	Causal verb incidence	**SMCAUSv**			.328*
	Causal verbs and causal particles incidence	**SMCAUSvp**		.404*	.410*
	Ratio of casual particles to causal verbs	**SMCAUSr**		.282^	
	Ratio of intentional particles to intentional verbs	**SMINTEr**[#]		.283^	

[#]As these data did not follow a normal distribution (*SW*s < .936, *df* = 36, *p* < .05), spearman's rho was used.

*. Correlation is significant at the 0.05 level (2-tailed);

^. Correlation failed to reach significance p < 0.10 (2-tailed).

positively with CNCCaus and SMCAUSvp (*r*s > .347, *p* < .05). DoM correlated positively with CNCADC , SMCAUSvp and (r*s* > .328, *p* <.097).

Table 5-10 shows changes over time in 7 of the 30 Coh-Metrix indices we tested. Wilcoxon Signed Ranks test showed that referential coherence CRFCWO1 (*Z*(17) = -2.004; *p* < .05) and CRFCWOa (*Z*(17) = -2.069; *p* < .05) had increased significantly from pre to posttest; Paired sample *t*-tests showed that LSA indices of LSASS1 (*t*(17) = 2.817; *p* < .05) and LSASSp (*t*(17) = 2.223; *p* < .05) increased significantly from pre to posttest and a trend towards an increase in LSAPP1 (*t*(17) = 1.886; *p* = .077) was found. Referential cohesion CRFCWOad (*t*(17) = 1.913; *p* = .073) and Syntactic

Table 5-10. Differences in Coh-Metrix measures between pre- and posttest writings.

Categories	Full Descriptions of measures	**Labels of Measures**	***Z/T*-score**	***p*-value**
Referential cohesion	Content word overlap, adjacent sentences, proportional, mean	**CRFCWO1**[#]	-2.004[b]	.045
	Content word overlap, all sentences, proportional, mean	**CRFCWOa**[#]	-2.069[b]	.039*
	Content word overlap, all sentences, proportional, standard deviation	**CRFCWOad**	1.913	.073
LSA	LSA overlap, adjacent sentences, mean	**LSASS1**	2.817	.012*
	LSA overlap, all sentences in paragraph, mean	**LSASSp**	2.223	.040*
	LSA overlap, adjacent paragraphs, mean	**LSAPP1**	1.883	.077
Syntactic pattern density	Adverbial phrase density, incidence	**DRAP**	2.077	.053

[#] As these data did not follow a normal distribution (*SW*s < .917, *df* = 18, *p*s < .011), Wilcoxon Signed-Rank Test was used and *Z*-statistic is provided;

b: based on positive ranks.

*. Significant at $p < 0.05$ (2-tailed).

Pattern Density item DRAP ($t(17) = 2.077$; $p = .053$) showed tendencies of a significant increase over time.

5.5 Discussion and conclusion

The current study examined whether 18 Chinese learners of L2 English had made some progress in writing over the 18-month advanced English course preparing them for the CET-6 test. A previous study had shown that they had made no progress whatsoever when proficiency was measured holistically and an automated analysis had shown that they had progressed only in a few isolated syntactic measures (see Chapter 3). As the outcome was difficult to believe, a subsequent study was conducted to see if the learners had improved in their use of formulaic language and it was found that the overall chunk coverage and the number of collocations had increased (see Chapter 4). The current study explored to see if the learners might have improved in coherence or cohesion. Coherence was operationalized as topic-based

coherence in which the number of related concepts are counted and the density of moves (DoM), operationalized as the average number of moves in 10 T-units, and strength of moves (SoM), operationalized as the average distance between these concepts, is calculated. Cohesion was operationalized as the explicit mention of cohesive devices (local cohesion) and the semantic overlap between sentences and paragraphs (global cohesion).

As far as coherence is concerned, the statistical analysis revealed that SoM showed very few correlations with the CAFIC proficiency measures and showed no significant difference between the pre and posttest. However, DoM correlated with almost all CAFIC proficiency measures not only when all the data were taken together but also with the pretest and posttest data separately. Moreover, there was a significant difference between the pre and posttest. We may conclude that DoM is a good measure of text quality and that the students improved significantly in coherence, operationalized as making fewer moves per 10 T-units. These findings on the relationship between topic-based coherence and holistic proficiency are consistent with Watson Todd et al.'s (2004) findings that the strength of relationships between key concepts does not correlate with the teachers' score for coherence, but density does. The question arises why strength does not correlate and density does. The strength is related to the type of relationship between concepts, those that are related in a conceptual space or not. It is conceivable that the degree to which concepts are related has no direct relation to coherence, a question that will need further research. Density, however, is probably a good indication of topic continuity and focus. The fewer different key concepts that appear on average, the more the writer remains on focus. As this measure is relatively easy to establish as only T-units and key concepts have to be established, it may be a useful measure in assessing coherence.

As far as cohesive devices are concerned, counted automatically with the Coh-Metrix tool, only three measures, two global and one local one, correlated with human judgments of overall text quality as measured with the CAFIC rubric. As far as global measures is concerned, the 'LSA overlap, adjacent sentences, standard deviation' measure correlated negatively. This may be explained as follows. The higher the standard deviation is, the more variability there is in overlapping concepts in adjacent sentences; in other

words, the writer is less consistent in using overlapping concepts. The 'LSA given/new, sentences, standard deviation' showed a positive trend. Since givenness refers specifically to the closeness of semantic similarity between sentences. A low standard deviation would mean that the givenness throughout the text is fairly stable; while a high standard deviation would suggest that it moves around a lot. The local measure that correlated positively with human ratings of text quality was the incidence of causal connectives correlated. A few more measures correlated with the coherence sub-score of the CAFIC rubric, and also some in the situation model involving causal verbs and particles. Somewhat different Coh-Metrix scores correlated with DoM: adversative connectives and causal verb incidence. We conclude from these isolated and varied correlations that there is only a vague correlation between human ratings and Coh-Metrix measures in these texts. However, the pre-posttest differences do show clear improvement for our 18 writers in global cohesion with 3 content word overlap measures and 3 LSA overlap measures. It may be argued that these six measures and DoM, which is based on the subjective identification of key concepts, partially tap into the same construct. They indicate that the text stays more on topic and therefore has more focus.

Because the study by Crossley et al. (2016a) is the most similar to ours, it is interesting to see to what extent our findings converge or not, but we also have to keep in mind that there were a few major differences. Crossley et al. examined a heterogeneous group of L2 students in that they had different L1s; their students were probably more advanced and had more exposure to the target language. Crossley et al. looked at changes over a six month period and we looked at changes over an 18 month period. Crossley et al. used two automated tools, Coh-Metrix and the Tool for the Automatic Analysis of Cohesion (TAACO; Crossley, Kyle & McNamara, 2016b), and we used a hand-counted measure and Coh-Metrix. Also the human ratings were based on different rubrics: Crossley et al. focused in their rubric on cohesion and we focused on general proficiency, based on a complexity, accuracy, fluency, idiomaticity and coherence (CAFIC) rubric. Despite these differences, the main similarity is that in both studies, strong growth occurred in global cohesion measures related to content words or key concepts, suggesting that like more mature L1 writers these more advanced

L2 writers have begun to work 'toward composing at the global level by developing coherence between paragraphs' (Crossley et al., 2016a, p. 13).

A difference between the studies may be found in the use of local cohesive devices. Crossley et al. found quite a few correlations between the use of such devices and human judgements and we did not, but we did find some significant increase in the use of some of these devices. Crossley et al. (2016a) speculate that the use of these devices may be related to topic, which could also have been the case in our study. Still, it is interesting that the use of local cohesive devices is not usually associated with more advanced L1 writing. Crossley et al. speculate that there may be some differences in which L1 and L2 writings are judged. We would agree that it is a possible cause, but an additional explanation could be that L2 writers may use formulations or words that are not target like, which may make it more difficult for readers to process the text. Explicit local cohesive devices can then help to make the logical connections between the propositions more clear and are therefore appreciated by human judges.

Our findings have some implications for research and teaching. First of all, the findings in this study support Verspoor et al. (2012) in the idea that L2 learners progress non-linearly. At different phases in their development, different aspects of language may develop. The learners in the current study had made no progress in any typical CAF measures, nor in any holistic scores, but they did improve in coherence, which may be argued to be a higher language skill. However, we have to realize that coherence is more than continuity of topic (Indrasuta, 1988). Still, we would like to argue that it is possible to operationalize coherence, although a subjective construct in itself, somewhat objectively. Both automated global cohesion measures and the hand counted DoM showed changes over time. However, in our study the hand-counted measure correlated better with human judgments. The reason could be that human judges are still better in judging whether words or concepts are related to each other or not than computational tools.

For teachers especially in the Chinese context, it may be useful to know that coherence does develop at the most advanced stages. Applying the key concept method, they might be able to show their learners why one piece of writing is more coherent than the other or show where breaks in continuity occur. However, as Watson Todd et al. (2004) point out, because of the

complex and laborious work involved in coding for topic continuity, there is no doubt that the method may have more value for researchers than for teachers.

This study has its limitations. First, we investigated only a small group of learners in China who were following an advanced course in English, and second, they wrote on a limited genre; therefore, we should not generalize beyond this group.