Systemic Functional Approach to Cohesion and Its Application to Automated Cohesion Analysis

Yu Tian, Minkyung Kim, Scott Crossley and Qian Wan

August 1, 2020

Systemic Functional Approach to Cohesion and Its Application to Automated Cohesion Analysis

Yu Tian[1], Minkyung Kim[2], Scott Crossley[1], and Qian Wan[1]

Department of Applied Linguistics and ESL, Georgia State University[1]

Department of International Studies, Nagoya University of Commerce and Business[2]

Author Note

The authors declare that there no conflicts of interest with respect to this preprint.

Correspondence should be addressed to Yu Tian (Email: ytian9@gsu.edu)

Abstract

This study investigated how the use of cohesive devices predicts writing fluency for second language (L2) undergraduate students (N = 99). Linear mixed effects models were built to predict writing fluency using cohesion indices. Results showed that the use of semantic overlap between adjacent sentences negatively predicted fluency in process. Furthermore, the use of more unattended demonstratives related to higher fluency in process but greater revisions, whereas more attended demonstratives associated with fewer revisions.

Keywords: writing fluency, cohesion, second language learners

The Use of Cohesive Devices as An Indicator of Writing Fluency for L2 Undergraduate Students

Introduction

Writing fluency (i.e., degree of language control during writing often expressed in terms of pausing times, revisions, and production rates; Ellis, 2003) has been recognized as an integral construct of learners' language proficiency. Research in recent years has documented a set of non-linguistic factors associated with second language (L2) learners' writing fluency such as the length of immersion (Chenoweth & Hayes, 2003), first language (Dustmann, 1994), writing prompt (Way, Joiner, & Seaman, 2002), and genres of writing (Olive, Favart, Beauvais, & Beauvais, 2009). However, more research is needed to examine how linguistic features in writing may interact with L2 learners' writing fluency as this perspective is important for understanding writing fluency (Leijten, Horenbeeck, & Van Waes, 2019).

The present study investigates how the use of cohesive devices predicts writing fluency using longitudinal data collected from a heterogeneous group of L2 undergraduate students in the U.S. We operationalize writing fluency via keystroke log measures including production rate, burst (inscription between pauses), pauses and revisions. Following Halliday and Hasan's (1976)'s framework on text cohesion, we conceptualize cohesive features in five categories: reference, substitution, ellipsis, conjunction, and lexical cohesion. The research question that guides this study is:

1. To what extent do cohesive features in L2 undergraduate students' essays predict writing fluency?

Method

Participants

L2 undergraduate students (N = 99) at a U.S. university participated in the study. These participants were from a variety of linguistic backgrounds among which Chinese (n = 15) and French (n = 15) were the most common. Their mean age was 20.384 years (*SD* = 2.629) and they had studied English for 12.869 years on average (*SD* = 4.082).

Design and Procedures

The writing data were collected over two days at a five-month interval. On each occasion, individual participant was seated in a quiet language laboratory and was given 25 minutes to complete an essay on one of the two SAT-based prompts (*Competition* and *Appearance*) in English. The order of the two prompts was counterbalanced among the participants, and the same data collection procedure was repeated during the second session. Participants' keystroke and mouse activities in both writing tasks were logged and time stamped via Inputlog 7 (Leijten & Van Waes, 2015).

Data Analyses

Writing fluency measures. Writing process information in terms of text production rate, bursts, pauses, and revision behaviors was extracted from the resulting log files using general analysis, pause analysis, and fluency analysis. In this study, when pause-related indices were calculated, a minimum pause threshold was set at 2000 ms, so that pauses represented participants' higher-level cognitive processes (e.g., planning and generating ideas), rather than lower-level cognitive processes, such as those related with lexical issues and spelling (Limpo and Alvès, 2017). Accordingly, Pause-burst (P-burst) measures were also determined based on this pause threshold. To control for essay lengths which varied across participants, we used indices that are only based on means, proportions or ratios were selected. These included *proportion of* *pause (P) time*, *mean length of pause (in seconds)*, *number of P-burst per minute*, *mean length of P- burst (in characters)*, *average strokes per minute*, *product vs. process ratio*. We then conducted a principal components analysis (PCA) on the six writing fluency indices to examine underlying structure of these indices. The PCA results showed that *Average strokes per minute, mean length of P- burst (in characters)* and *proportion of pause time* cluster on the same component that can be labeled as *general fluency in process* as these indices collectively indicate the speed of production in the writing process. The index of *product vs. process ratio* stands alone as a measure of *revision*.

Cohesion features. Three main types of cohesive devices were calculated in this study: reference, conjunction, and lexical cohesion. With respect to reference, personal reference was measured using SiNLP (Crossley, Allen, Kyle & McNamara, 2014) by calculating the proportion of first-person pronouns, second-person pronouns, third-person pronouns, and all pronouns used in the essays. Demonstrative reference in the essays was analyzed through TAACO (Crossley, Kyle & McNamara, 2016) by selecting three indices related to demonstratives: the percentage of attended demonstratives, unattended demonstratives, and all demonstratives. Comparative reference, operationalized as the proportion of comparative adjectives, comparative adverbs, superative adjectives, and superative adverbs, was analyzed using part-of-speech tags reported by spaCy v2.2 (spaCy core team, 2017), an open-source library for NLP in python. Next, the use of conjunction was assessed by TAACO (Crossley, Kyle & McNamara, 2016) using a list of connective indices that cover the four types of conjunction in Halliday and Hasan (1976): additive, adversative, causal, and temporal. Finally, lexical cohesion was calculated through TAACO (Crossley, Kyle & McNamara, 2016) by analyzing sentence overlap (overlap between words and related semantic content between adjacent sentences) and paragraph overlap (overlap between words and related semantic content between adjacent paragraphs). Semantic content overlap was estimated based on computational models including semantic vector spaces using Latent Semantic Analysis (LSA; Landauer et al., 1998), topic distributions in Latent Dirichlet Allocation (LDA, Blei et al., 2003), and word2vec vector space representations (Mikolov et al., 2013). Cohesion features related to substitution and ellipsis were not included in this study as they are difficult to be quantify. In total, 56 potential indices were selected to assess text cohesion in the essays.

Statistical Analyses

Prior to statistical analysis, the 56 potential indices were pruned to avoid overfitting of the models. First, all indices were checked for normality. Second, correlation analyses were conducted among all the cohesion indices to check for multicollinearity among the indices. Indices that were highly collinear (absolute r > .7) were flagged, and the index with the strongest correlation with the selected fluency measures (*general fluency in process* and *revision*) reported by the PCA was retained. Accordingly, 24 out of the 56 indices were selected for further analyses.

Two separate linear mixed effects models were built using R (R core Team, 2015) with the lme4 (Bates, Maechler, Bolker, & Walker, 2015) packages to examine whether cohesion indices were significant predictors for each writing fluency measures (*general fluency in process* and *revision*). For each model, correlations were calculated between the cohesion indices and the dependent writing fluency variable, and only indices that showed a meaningful relationship (absolute r > .1) with the writing fluency variable were entered into the model as fixed effects (independent variables) to avoid over fitting.

Additionally, time (time1 and time2), college year (1st, 2nd, 3rd, and 4th) and prompt (*Competition* and *Appearance*) were also entered as fixed effects because time, school year, and prompt type have been shown to affect students' writing fluency (see Chenoweth & Hayes, 2003; Kowal, 2014; Way, Joiner, & Seaman, 2002). Participants and their first languages were included into the model as random effects. To find the best model, independent variables that were significant predictors in each of these three full LME models were entered into a new LME model along with the random effects to predict the corresponding writing measure. To obtain a measure of effect size of each LME model, r.squaredGLMM function from the MuMIn package (Nakagawa & Schielzeth, 2013) was used to calculate two separate R^2 values: a marginal R^2 for the variance explained by merely the fixed factors, and a conditional R^2 for the variance explained by both fixed and random factors together.

Results

General Fluency in Process

The results showed that *general fluency in process* was significantly predicted by *prompt* (t = 2.643, p < .01), *unattended demonstratives* (t = 2.92, p < .01), and *LSA cosine similarity (adjacent sentences)* (see Table 1).These results indicate that participants tended to write more fluently on *competition* than *appearance,* and that participants writing more fluently were likely to use more unattended demonstratives but relied less on semantic overlap between adjacent sentences in their essays. This final model reported a marginal R^2 of .092 and a conditional R^2 of .508, suggesting that the three variables collectively explained 9.2% of the variance in participants' general fluency in process. Visual inspection of the residual distribution suggested that this model was not affected by heteroscedasticity.

Table 1

Linear Mixed Effect Model for General Fluency in Process

Fixed Effect	Coefficient	SE	t	p
(Intercept)	0.241	0.277	0.869	0.386
Prompt: competition	0.305	0.116	2.643	< .01
Unattended demonstratives	18.745	6.42	2.92	< .01
LSA cosine similarity (adjacent sentences)	-2.197	0.702	-3.128	< .01

Note: marginal R^2 = .092, conditional R^2 = .508.

Revision

The results showed that *revision* was significantly predicted by *conjunctions*, *attended demonstratives*, and *unattended demonstratives* (see Table 2). These results indicate that participants who used more conjunctions, more unattended demonstratives, and fewer attended demonstratives tended to revise more in their writing process. This final model reported a marginal R^2 of .103 and a conditional R^2 of .671, suggesting that the three variables collectively explained 10.3% of the variance in participants' revision behaviors. Visual inspection of the residual distribution suggested that this model was not affected by heteroscedasticity.

Table 2

Linear Mixed Effect Model for Revision

Fixed Effect	Coefficient	SE	t	p
(Intercept)	0.017	0.089	0.192	0.848
Conjunctions	0.179	0.067	2.665	< .01
Attended demonstratives	-0.211	0.064	-3.328	< .01
Unattended demonstratives	0.27	0.067	4.061	< .001

Note: marginal R^2 = .103, conditional R^2 = .671.

Discussion

This study identifies a set of cohesive features in L2 undergraduate writers' essays as significant predictors for writing fluency. The results from LME models showed that *LSA cosine similarity (adjacent sentences)* negatively predicted *general fluency in production*, suggesting that writers who produced greater semantic overlap among adjacent sentences wrote less fluently. The study also found that writers who produced more conjunctions (e.g., *and*, *but*) had greater revisions.

The use of demonstratives related to writing fluency in a more nuanced manner. While the use of more unattended demonstratives predicted greater general fluency in production, more unattended demonstratives in essays also predicted more revisions. In contrast, students who used more attended demonstratives tended to have fewer revisions. The use of unattended demonstratives (e.g., *This makes her happy*) related to greater fluency compared to attended demonstratives (e.g., *This compliment makes her happy*) because such writing is more economical (Geisler, Kaufer, & Steinberg, 1985) and efficient (Rustipa, 2015). In terms of revision, attended demonstratives present textual reference with more clarity and recontextualization of the previous text than unattended demonstratives (Swales, 2005), potentially leading to fewer revisions.

Apart from cohesive features, this study also found that L2 undergraduate students' revision behaviors varied with different prompts, underscoring the importance of prompt-based effects. Such effects have been shown for linguistic feature production, but not for fluency measures.

Reference

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). lme4: Linear mixed-effects models using Eigen and S4_. R package version 1.1-9. Retrieved from https://CRAN.R-project.org/package=lme4

Chenoweth, N. A., & Hayes, J. R. (2003). The inner voice in writing. *Written Communication*, 20, 99-118.

Crossley, S. A., Allen, L. K., Kyle, K., & McNamara, D.S. (2014). Analyzing discourse processing using a simple natural language processing tool (SiNLP). *Discourse Processes, 51*(5-6), pp. 511-534, DOI: 10.1080/0163853X.2014.910723

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods 48*(4), pp. 1227-1237. doi:10.3758/s13428-015-0651-7

Dustmann, C. (1994). Speaking fluency, writing fluency and earnings of migrants. *Journal of Population Economics*, *7*(2). https://doi.org/10.1007/BF00173616

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

Geisler, C., Kaufer, D. S., & Steinberg, E. R. (1985). The Unattended Anaphoric “This.” *Written Communication*, *2*(2), 129–155. doi: 10.1177/0741088385002002002

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Place of publication not identified: Longman.

Leijten, M., & Van Waes, L. (2015). *Inputlog Manual 8.0.* Retrieved from http://www.inputlog.net/wp-content/uploads/Inputlog_manual.pdf.

Olive, T., Favart, M., Beauvais, C., & Beauvais, L. (2009). Children's cognitive effort and fluency in writing: Effects of genre and of handwriting automatisation. *Learning and Instruction, 19*(4), 299-308.

R Core Team. (2015). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rustipa, K. (2015). The Use of Demonstrative Pronoun and Demonstrative Determiner this in Upper-Level Student Writing: A Case Study. *English Language Teaching*, *8*(5), p158. https://doi.org/10.5539/elt.v8n5p158

spaCy Core Team. (2017). Industrial-strength Natural Language Processing in Python. Retrieved from https://spacy.io/

Swales, J. M. (2005). Attended and Unattended “this” in Academic Writing: A Long and Unfinished Story. *ESP Malaysia*, 11, 1-15.

Way, D. P., Joiner, E. G., & Seaman, M. A. (2000). Writing in the Secondary Foreign Language Classroom: The Effects of Prompts and Tasks on Novice Learners of French. *The Modern Language Journal*, *84*(2), 171–184. https://doi.org/10.1111/0026-7902.00060