Revising in a non-alphabetic language: The multi-dimensional and dynamic nature of online revisions in Chinese as a second language

Xiaojun Lu[a]

Andrea Révész[b]

[a] Southeast University, College of International Students, Nanjing, People's Republic of China

[b] University College London, Institute of Education, 20 Bedford Way, WC1H 0AL, London, UK

Abstract

While second language (L2) writing processes have received increased attention in recent years, few studies have considered writing processes in non-alphabetic languages. To help fill this gap, this study examined the online revision behaviours and associated cognitive processes of L2 writers of Chinese from a multi-dimensional and temporal perspective. Thirty-two L2 Chinese writers performed four writing tasks while their keystrokes were logged. Based on their last writing performance, participants engaged in a stimulated recall session, during which they were asked to describe their thoughts during revisions. Baseline data were collected from 32 first language (L1) writers of Chinese following the same procedure. Revisions were coded for linguistic domain, context, and level of transcription. Stimulated recall comments were categorised according to the orientation of revision. We found that L2 writers more frequently revised language than content, smaller linguistic units than larger ones, and text at the inscription point than previously produced text. They also revised Pinyin more than characters. Revisions occurred more frequently in the middle stages of writing, except for most contextual revisions made in the final stage. Similar trends were observed in L1 writing, apart from character revisions outnumbering Pinyin revisions and proportionately more revisions focusing on content.

Keywords: online revision; second language writing; Chinese as a second language; temporal dimension; keystroke-logging; stimulated recall

Introduction

Revision refers to any reflection and/or transformation made at any point of writing (Barkaoui, 2016; Piolat, 1997). It is an effortful process (Chenoweth & Hayes, 2001), which can be triggered by various activities, from noticing a mismatch between task requirements and the evolving text to correcting a mistake in a lexical item. Although revision has been extensively studied in first and second language (L1, L2) writing (e.g., Bereiter & Scardamalia, 1987; Raimes, 1985), little is known about revision processes in L2 Chinese writing. Unlike previous L2s studied, Chinese has a logographic writing system, which, due to the non-transparent sound-form mappings (Zhang, 2017), may generate differential demands on writing activities such as planning, translating, monitoring, and execution (Ellis & Yuan, 2004). The goal of this study was to examine whether and how the differential demands posed by using a logographic writing system may influence the amount and nature of online revisions. To achieve this, we adopted a multi-dimensional and temporal perspective building on previous research. Rather than following the common practice of focusing on a single revision type, we investigated revision behaviours by taking into account various dimensions (e.g., linguistic domain, context, and orientation of revision) (e.g., Stevenson et al., 2006). Also, instead of adopting a static view, we analysed online revisions for both the whole and different stages (beginning, middle, end) of the L2 Chinese writing process, given increasing evidence that writing is a dynamic process (e.g., Tillema, 2012). In addition, we explored L2 Chinese writers' online revision by combining quantitative (keystroke-logging) and qualitative (stimulated recall) data sources. We expected that

a mixed-methods approach would allow for a more valid and reliable interpretation of the data obtained (Révész & Michel, 2019).

Literature review

2.1 Revision: a multi-dimensional construct

Producing an essay entails a set of complex activities, including planning goals and ideas, translating ideas into a written message, transcribing the text using either a pen or a keyboard, and monitoring the developing text (e.g., Flower & Hayes, 1981). These activities are assumed to occur in parallel, manifest in observable behaviours such as revision (Lindgren & Sullivan, 2006b).

Revision in writing varies in quality. Among the frameworks developed to categorise revisions (e.g., Matsuhashi, 1987), two similar taxonomies, proposed by Lindgren and Sullivan (2006a) and Stevenson et al. (2006), are the most comprehensive, modelling revision from multiple perspectives. They first distinguish revisions as external and internal based on whether or not, respectively, the revision entails a visible change to the text. Next, external revisions are classified according to context; pre-contextual revisions are made at the point of inscription (i.e., end of the current text), and contextual revisions involve changes to previously written text. Both pre-contextual and contextual revisions may have different orientations; changes can relate to content, language, or typography. External revisions can also involve different levels of linguistic units (i.e., linguistic domain), such as revisions within a word or to a word, clause, or paragraph. In addition, external revisions can be classified in terms of action,

for example, whether they include additions, deletions, or substitutions. Due to working memory limitations (Kellogg, 1996), different types of revision are expected to compete for attentional resources. If more attention is allocated to a certain type of revision, the amount of attention available to other types of revision is likely to be constrained.

Many studies have investigated revision in paper-based writing using verbal protocols (e.g., think-alouds) and found that L2 writers tended to pay more attention to linguistic aspects of the text and revise more frequently within lower linguistic domains (e.g., Chenoweth & Hayes, 2001; Roca de Larios et al., 2008; Zimmermann, 2000). However, these findings may be questionable due to several methodological limitations. Think-alouds may be reactive, therefore failing to provide a valid reflection of writers' thoughts during revision (e.g., Janssen et al., 1996). Another problem is that verbal protocols may not capture all the cognitive activities of writers during revision given that some may be unconscious and thus not verbalisable (e.g., Ericsson & Simon, 1993). Also, most aforementioned research included a relatively small sample size; the results need further confirmation by larger-scale studies.

To address these issues, a few scholars have employed keystroke-logging to obtain a more fine-grained picture of the multi-dimensional nature of revision (e.g., Barkaoui, 2016; Choi, 2007; Khuder & Harwood, 2015; Stevenson et al., 2006; Xu, 2018). Similar to findings obtained from verbal protocols, these researchers observed more language- than content-oriented revision in L2 writing (e.g., Barkaoui, 2016; Stevenson et al., 2006). However, findings were mixed regarding other dimensions. For example, some studies revealed a decreasing number of revisions as the level of linguistic units

increased (e.g., Choi, 2007; Stevenson et al., 2006), while others found sentence-level revision to be the most frequent revision type (e.g., Khuder & Harwood, 2015). As for context, in Barkaoui (2016), Stevenson et al. (2006) and Xu (2018), L2 writers revised more often at the inscription point, whereas those in Choi (2007) and Khuder and Harwood (2015) carried out more contextual revisions.

It also remains to be explored how types of revisions may vary when L2 writers compose in a non-alphabetic language. In doing so, at least one additional dimension, level of transcription, needs to be included in multi-dimensional revision taxonomies. The indirect connection between Chinese characters and their pronunciation (Zhang, 2017) makes it important to differentiate between two levels of transcription. First, writers can make changes to Chinese characters, demonstrated as logographic symbols, during writing. Second, writers may also revise before characters are formed. For example, they may delete and re-enter Pinyin (i.e., the phonetic reading of a character) when typing using the Pinyin input method (see 3.2.3). Due to their partial acquisition the sound-form mapping of Chinese, L2 writers are expected to revise phonetic readings more frequently, given that processing texts written in letters poses lower cognitive demands than processing texts with logographic symbols. As previous L2 investigations on revision have predominately focused on alphabetic language writers, this assumption is yet to be tested.

2.2 Temporal dimension of revision

In previous research, the analysis of L2 writing processes has largely been limited

to counting frequencies of writing activities for a whole composition session. According to Van Den Bergh and Rijlaarsdam (1996), however, writing activities are expected to differ in terms of how probable they are to occur at different points of the writing process. In other words, writers may engage in specific writing activities (i.e., planning, formulation, revising) to different extents across various writing stages.

Following Van Den Bergh and Rijlaarsdam's (1996) proposal, researchers have begun to explore the role of time in the writing process. In investigating revision as a function of time, Roca de Larios et al. (2008) divided think-aloud protocols into three equal stages based on participants' total writing time. The researchers found that most revisions were made in the last stage. However, Tillema (2012), also utilising the think-aloud procedure, reported that the occurrence of revision was equally frequent throughout text production. Similar, Gánem-Gutiérrez and Gilmore (2018) observed no change in the amount and duration of revision using eye-tracking. A limitation of all these studies, however, was that they considered revision as a homogeneous entity.

Only a single study, Barkaoui (2016), investigated revision behaviours by taking into account both the multi-dimensional and temporal nature of writing. The writing processes of 54 L2 writers of English were recorded by the keystroke-logging software *Inputlog* and segmented into three phases based on total writing time. The researcher found that most revisions occurred in phase two. When revisions were sub-divided according to context, the majority of pre-contextual revision was found in phase two whereas contextual revision occurred most frequently in phase three.

Based on earlier studies in alphabetic writing, we anticipated that L2 Chinese

writers would also make more pre-contextual revisions in the middle phase of writing. However, unlike in alphabetic language writing, we did not expect many contextual revisions towards the end of writing, given the relatively high cognitive demands posed by re-reading drafts written in Chinese characters.

2.3 The present study

While previous studies have provided some useful information about L2 writers' revision behaviours, several issues need further investigation. First, the results are inconclusive for the incidence of various types of revision behaviours (e.g., Barkaoui, 2016; Khuder & Harwood, 2015; Xu, 2018). Also, little is known about the temporal dimension of revision; the few studies that have considered the temporal nature of writing did not take into account how this might interact with various revision types (except Barkaoui, 2016). In addition, revision studies to date have mainly utilised a single data elicitation method (verbal protocols or keystroke-logging), not allowing for the triangulation of various data sources (Révész & Michel, 2019). Lastly, so far, little research has looked into non-alphabetic writers' revision behaviours. To address these gaps, we formulated three research questions:

1. To what extent do L2 Chinese writers engage in different types of revision involving various dimensions, as reflected in

    a. revision behaviours captured by keystroke-logging?

    b. stimulated recall comments associated with revision behaviours?

2. To what extent do the stages of writing affect types of revision in L2 Chinese

writing, as reflected in

a. revision behaviours captured by keystroke-logging?

b. stimulated recall comments associated with revision behaviours?

3. To what extent do these patterns differ in L1 Chinese writing?

Notably, we used data collected from L1 writers of Chinese as a baseline to determine if the observed trends were L2-specific given the limited research into non-alphabetic language writing. In other words, we were interested in whether revision patterns in L1 and L2 Chinese writing were similar, rather than making direct quantitative comparisons between the two groups.

Methodology

3.1 Participants

The dataset for this study was part of a larger project examining the separate and joint effects of writing stage, genre, and proficiency on L2 writing behaviours and associated cognitive activities in Chinese writing. The participants were 32 L2 writers of Chinese (19 males and 13 females) studying or living in London. They were all L1 users of an alphabetic language, and their ages ranged from 19 to 41 years (M = 26.62, SD = 6.52). They varied in their L2 Chinese proficiency, from pre-intermediate to advanced (see 3.2.1), with an average of 58.84 months (SD = 37.85) of previous Chinese study.

Baseline data were provided by 32 L1 writers of Chinese (5 males and 27 females). They were students studying at universities in London. Their ages ranged from 18 to

32 years (M = 24.72, SD = 3.54). They received primary and secondary education in China and had on average been in the UK for 6.37 months (SD = 2.32). They were using Chinese daily (39.16 h per week), with an average of 9.33 hours per week spent writing.

3.2 Instruments

3.2.1 Proficiency test

L2 writers' proficiency in Chinese was estimated by a cloze test adapted from the Test of Chinese as a Foreign Language. The internal consistency reliability of the test was found to be good (Cronbach's alpha = .89). The cloze test scores indicated that L2 writers' proficiency levels in Chinese ranged from pre-intermediate to advanced (range 13 to 41 out of 45 points, M = 28.25, SD = 7.57). Our rationale for using a cloze test was that it took less time to complete than available standardised assessments. Also, close test scores have been shown to positively correlate with those of standardised proficiency tests (e.g., Bachman, 1986), thus arguably generating a reliable 'proxy' for L2 proficiency (Gellert & Elbro, 2013).

3.2.2 Writing tasks

Two argumentative and two narrative writing tasks were used to elicit participants' writing performances to avoid a potential prompt effect. When developing the prompts, we consulted five experienced instructors of L2 Chinese to ensure that the topics were suitable for both high- and low-proficiency L2 Chinese writers. See Table S1 in Supplementary material for the full prompts.

3.2.3 Pinyin input method

The Pinyin input method is the most popular phonetic-based input method among writers of Chinese (Xiao et al., 2007). Pinyin refers to the official Romanisation system of Chinese characters based on their pronunciation. For example, /wǒ/ is how the Chinese character *我* (I) spells in Pinyin. When using the Pinyin input method, writers first type the Pinyin of a character (without the tone mark) and then select the character from a list of homophones. Writers can also enter the Pinyin of multiple characters when typing multi-character phrases. Second, writers select the desired character/multi-character phrase from among options offered by the program. Figure 1 provides an example of typing the Pinyin of a multi-character phrase (*绝妙创意*, excellent idea). The character options appear once writers have pressed the space bar after the character at the inscription point (*个* in Figure 1). Given that using the Pinyin input method involves two steps, writers can revise both Pinyin and characters during composition. This is a feature different from typing in alphabetic languages, where transcription typically involves a single step.

Figure 1 ABOUT HERE

3.2.4 Keystroke-logging

The keystroke-logging software, *Translog 2.0* (Carl, 2012), was used to record writers' revision behaviours. Keystroke-logging is a relatively new technique for studying L2 writing behaviours. This technique captures every keystroke during text

production and, unlike think-alouds, it does not interrupt the writing process (Van Waes et al., 2009). Keystroke-logging, however, has the disadvantage of providing no direct information about writers' cognitive activities. To address this, we combined stimulated recall (see 3.2.5) with keystroke-logging.

3.2.5 Stimulated recall

To tap into participants' cognitive activities associated with revision behaviours, we conducted stimulated recalls. To mitigate issues with veridicality (Gass & Mackey, 2016), all participants took part in a stimulated recall interview immediately after they finished writing. To avoid reactivity, the interviews were based only on the last writing task they had performed, no stimulated recall session was scheduled between writing sessions. The interviews were conducted in English for L2 writers and in Mandarin Chinese for L1 writers by the first author. To prompt recall, the participants watched a recording of their writing performance, presented as keystroke logs at the character level (*Translog 2.0* did not allow for playing back of the Pinyin transcription process). The researcher paused the recording whenever a revision occurred and asked the participant to recall the thoughts they had at the time. The sessions were video recorded.

3.3 Data collection

Each participant attended two individual sessions. In each session, they wrote two essays using the Pinyin input method. The order of the prompts was counterbalanced across participants. Thirty minutes were given for completing each task. While

composing, the participants' keystroke behaviours were recorded by *Translog* 2.0. The automatic spelling and grammar checker functions were turned off. The first author also monitored the participants to ensure that they did not refer to additional materials. In session 2, immediately after the participants had finished writing, the stimulated recall session followed. L2 Chinese writers completed the proficiency test after the stimulated recall session. Figure 2 demonstrates the data collection procedures.

3.4 Data analysis

3.4.1 Analysis of the *Translog* files

This study focused on external revisions only. After all revisions had been identified from the *Translog* files, they were manually coded using an adapted version of Stevenson et al.'s (2006) multi-dimensional taxonomies (see Table 1). First, each revision was categorised in terms of linguistic domain, whether it involved a change below the word level, at the word level, below the clause level, or at the clause level and above. Next, revisions were classified according to context, whether they were pre-contextual or contextual. To accommodate the nature of typing in Pinyin, we added level of transcription to Stevenson et al.'s (2006) taxonomies, distinguishing revisions in terms of whether they entailed a change to Pinyin or Chinese characters. We did not include 'orientation' in the coding scheme for the *Translog* files, as the coding, due to ambiguity, was likely to lack reliability without complimentary data from verbal reports (cf., Stevenson et al., 2006). However, orientation was considered when analysing the stimulated recall protocols for cognitive activities associated with revision behaviours

(see 3.4.5).

Table 1 ABOUT HERE

The data for three L1 and three L2 writers were double coded, and Cohen's kappa was .93. Finally, the revisions were added up to obtain a frequency count for each participant by category.

3.4.2 Analysis of stimulated recall comments

The full stimulated recall session was transcribed for each participant. All comments were reviewed, and those associated with typos (34% for both groups) were excluded from the analysis. The remaining comments were coded for emergent categories. The resulting categories were grouped into macro-categories (i.e., content and language) following Stevenson et al. (2006). To determine the linguistic domain and context of revision relative to its orientation, we triangulated the stimulated recall data with the *Translog* files. Table 2 provides a typical example for each category and how data triangulation was carried out. Finally, comments that fell into a specific category were added up, and a percentage for each category was calculated. The data for six L1 and six L2 writers were coded by another coder. Inter-coder reliability was high (Cohen's kappa = .94).

Table 2 ABOUT HERE

3.4.3 Analysis of the temporal dimension of revision

The temporal dimension of revision was investigated by dividing the writing

session into five stages for each task (Gánem-Gutiérrez & Gilmore, 2018; Tillema, 2012), based on the total time the participant had spent on task performance. Next, each *Translog* file and verbal protocol was segmented into five parts, each corresponding to one of the five stages in the writing session. In the last step, all revision indices were calculated for each stage.

The raw frequency count by revision category was corrected for time (in minutes) by dividing the frequency counts by the duration of the whole writing session or stage. The rationale for using time rather than text length (e.g. Barkaoui, 2016; Stevenson et al., 2006) for standardising was that the actual time that each participant took to complete the tasks varied (some participants finished writing in less than 30 minutes). Therefore, a time-based standardisation procedure was likely to yield more valid results than a word-based one by controlling for any effects arising from differences in task completion times.

3.5 Statistical Analyses

Statistical analyses were performed using the *RStudio* package. The data for revision behaviours were first inspected for outliers. These were trimmed to values of three standard deviations from the mean for each index per participant. To examine the extent to which L2 Chinese writers engaged in different types of revision, we constructed three linear mixed-effects regression models, one for each dimension (linguistic domain, context, level of transcription). For all models, the dependent variable was the log frequency count for revision (log transformation was employed to

improve model fit). The fixed effect was a dimension of revision, and subcategories of the dimension served as levels of the fixed effect (e.g., the dimension *context* had two levels: contextual and pre-contextual). Two random effects, participant and prompt, were initially included in all models. However, we needed to remove prompt, the random effect explaining less variance, to achieve model convergence. A similar set of analyses was run for the L1 group, except that the L1 models for linguistic domain and level of transcription included both participant and prompt as random effects. As there were four levels under linguistic domain, Bonferroni post-hoc tests were conducted to compare each of the two levels.

To investigate the extent to which stages of writing affect revision in Chinese writing, we conducted linear mixed-effects regression analyses for the L1 and L2 data separately. The dependent variable in the models was, again, the log-transformed frequency count of revision (however, for total, pre-contextual, and character revision in L1 writing, the data was not transformed as the transformation did not result in better model fit). For all models, the fixed effect was stage of writing, and participant and prompt were included as random effects. Prompt was removed from models that failed to converge. When a significant effect for stage emerged, Bonferroni post-hoc tests were conducted to identify pairwise differences.

Notably, for all L2 models, we initially included L2 proficiency, operationalised as the cloze test score, as a moderator, given the difference in participants' L2 Chinese proficiency. However, adding the moderator increased the models' *BIC* (Bayesian Information Criteria, see Table S2 in Supplementary material for details), indicating

better-fit models without L2 proficiency.

The threshold for *p* was set at .05 for all mixed-effects models but lowered to .01 for post-hoc tests. To obtain effect-size estimates for the mixed-effects regressions, we calculated marginal R^2 (R^2_m) values which indicated the amount of variance explained by the fixed effects in the models. Cohen's *d* was computed as an effect size for the post-hoc tests; d-values of .60, 1.00, and 1.40 were considered as small, medium, and large respectively (Plonsky & Oswald, 2014).

Results

4.1 Types of revisions in Chinese writing

Descriptive statistics for revision behaviours are summarised in Table 3. Mixed-effects analyses and Bonferroni post-hoc tests yielded significant effects of linguistic domain, context, and level of transcription on revision frequency for both writer groups (see Tables 4 and 5).

The results indicate that L2 Chinese writers revised most frequently below the word level, followed by word-level and below-clause revisions. The fewest revisions were made at the clause level and above. The differences between below-word and word-level revisions were small, while the rest of the differences had effects sizes in the medium to large range. A similar pattern was observed for L1 Chinese writing, the only difference was that L1 writers made equal amounts of word-level and below-clause revisions. Turning to context, more pre-contextual than contextual revisions were found in both L1 and L2 writing, with context explaining 69% and 86% of the variance in

each group, respectively. Finally, L2 writers revised Pinyin more often than characters; level of transcription explained 15% of the variance. A reverse trend was seen in L1 writing, with L1 writers revising characters more frequently than Pinyin, level of transcription accounting for 4% of the variance.

Tables 4 and 5 ABOUT HERE

Tables 6 and 7 summarise the stimulated recall comments for linguistic domain and context (no stimulated recall data was available for level of transcription, see 4.2.3). In L2 Chinese writing, considerably more comments referred to language- (67%) than content-oriented revisions (27%). The majority of content-oriented revisions (93%) were associated with changes to ideas, whereas most language-oriented revisions concerned lexis (40%) and grammar (36%). In terms of linguistic domain, most revisions below the word level (92%), at the word level (66%) and below the clause level (67%) were language-oriented. On the other hand, clause-and-above revision involved similar number of references to content (46%) and language (45%). For context, language was recalled as the primary focus regardless of whether participants made pre-contextual (68%) or contextual revisions (66%).

The stimulated recall data yielded largely similar patterns for L1 writers, but we also observed some small differences. While the majority of revisions by L1 writers concerned language (56%), there was a higher proportion of comments related to content (37%) than in the L2 group. Similar to the L2 group, most content-oriented comments referred to changing ideas (89%), and the majority of language-related comments focused on lexis (37%). Like L2 writers, L1 writers made most reference to

language when revising smaller linguistic units (below word: 65%, word level: 63%, below clause: 58%). However, unlike L2 writers, they attended to content (58%) more frequently than language (35%) when revising at the clause level and above. In terms of context, similar to L2 writing, most pre-contextual (56%) and contextual (57%) revisions were language-oriented.

Tables 6 and 7 ABOUT HERE

4.2 Effect of stage on revision in Chinese writing

The descriptive statistics for stage of writing are presented in Table S3 in Supplementary material. The mixed-effects regressions found significant effects for stage for all revision indices (see Table S4 in Supplementary material for model functions and results). As shown in Table 8, the post-hoc tests revealed that, in L2 Chinese writing, stages 2, 3 and 4 featured more revision in total than stages 1 and 5. In terms of linguistic domain, more below-word, word-level and below-clause revisions occurred in the middle stages. Moving onto context, L2 writers made more pre-contextual revisions in stages 2, 3 and 4, while the amount of contextual revision increased from beginning to end. For level of transcription, we found less Pinyin revision in stages 1 and 5 and fewer changes to characters in stage 1 only. Similar stage effects were identified in L1 Chinese writing. In general, L1 writers revised more often in the middle stages. The only exception to this trend was more frequent contextual revision observed in stage 5 as compared to stage 1. Most effect sizes were found to be in the small range.

Table 8 ABOUT HERE

The descriptive statistics for stimulated recall comments by stage are summarised in Table 9 (see Tables S5 and S6 in Supplemental material for details). Similar to the pattern observed for the whole session, both L1 and L2 writers referred to language more often than content when recalling their thoughts during each stage. In L2 writing, however, we found a gradual decrease in content-oriented comments from stage 1 to 5, whereas stage 1 in L1 writing featured proportionally more language-oriented comments than subsequent stages.

Table 9 ABOUT HERE

When considering linguistic domain, the majority of below-word, word-level and below-clause revisions described language-related changes for both groups across all stages. Also, pre-contextual and contextual revisions were primarily language-oriented for both groups regardless of the stage. Notably, however, the percentage of contextual revision related to language increased considerably in stage 5 in L2 writing, whereas it stayed relatively stable across stages in L1 writing.

Discussion

5.1 Revision in Chinese writing: from a multi-dimensional perspective

Our first research question asked the extent to which L2 Chinese writers engage in different types of revision, as reflected in keystroke-logging and stimulated recall data. Our results for L2 Chinese writers yielded similar patterns to those observed for L2 English writers, confirming the multidimensional nature of revision for logographic

writing. In particular, keystroke-logging analyses revealed that L2 Chinese writers revised smaller linguistic units more frequently than larger units, echoing the results of Stevenson et al. (2006) and Choi (2007). More pre-contextual than contextual revisions were found, which also mirror the findings of earlier studies (e.g., Barkaoui, 2016; Stevenson et al., 2006). The stimulated recall data revealed that L2 writers attended to language issues more frequently than content during revision, which is, again, consistent with the findings of Barkaoui (2016) and Stevenson et al. (2006). Combining the two data sources, our data suggest that L2 Chinese writers, similar to L2 writers of English, tend to focus mainly on language issues during revision, particularly when revising smaller linguistic units and text at the inscription point.

The increased focus on language during revision may be attributed to L2 writers' limited L2 proficiency (Murphy & Roca de Larios, 2010). As compared to L1 writers, L2 writers likely devoted more conscious effort to language use than other aspects of writing, as they encountered more difficulty with linguistic encoding processes, leaving less attention to allocate to content revision. This explanation also receives support from the L1 data. Although L1 Chinese writers revised language more often than content, the difference in the distribution of language- versus content-oriented revisions (approximately 1.5:1) was much smaller than that in L2 writing (approximately 3:1). Less likely to encounter linguistic barriers, L1 writers were probably more able to direct more cognitive resources to content-oriented revision.

It is also worth highlighting that L2 Chinese writers, as expected, revised Pinyin more frequently than characters, whereas the opposite trend was observed for L1 writers.

One reason for the more extensive Pinyin revision by L2 writers might be that some Pinyin letters do not exist or correspond to sound in the same way in the writers' L1 script (i.e., English, German, French, Italian, and Polish). For instance, *x*, pronounced as /ɕ/ in Pinyin, does not exist in the Polish or the Italian alphabet, while it is pronounced as /ks/ in English, German, and French. This probably led to frequent Pinyin mistakes and subsequent revision by L2 writers. In addition, the link between Chinese characters and their pronunciation is often non-transparent (Kang, 2011), making it harder to proceduralise sound-form mappings in Chinese. As a result, selecting target characters after typing Pinyin likely caused extra effort for L2 writers, maybe making them less inclined to change characters once they have succeeded in producing them. On the other hand, L1 writers had more automatised knowledge of the sound-form mappings associated with Chinese characters, enabling them to commit fewer mistakes and thus revise Pinyin less often.

5.2 The role of writing stage in revision in Chinese

Our second research question was concerned with the effects of stage of writing on different types of revision by L2 Chinese writers, as reflected in keystroke-logging and stimulated recall data. We found that, similar to alphabetic language users, L2 Chinese writers displayed differential revision patterns depending on writing stage, extending evidence for the dynamic nature of the writing process (Van Den Bergh & Rijlaarsdam, 1996) to the context of non-alphabetic language writing.

Turning to specific trends, we found most differences between stages 1 and/or 5

and the middle stages. Stage 1 in L2 Chinese writing featured fewer total revisions than subsequent stages, replicating the results of Barkaoui (2016) and those of Roca de Larios et al. (2008). This pattern may be attributed to the fact that writers primarily focused on planning at the beginning of writing (Ong, 2014), which, in turn, resulted in less text production and thus fewer revisions. This explanation received support from the verbal data, with L2 writers making proportionally more references to content in stage 1 than in sequential stages. L1 writers in this study also revised less in stage 1, indicating that this trend is not unique to L2 writing.

Interestingly, total revision amount was also found to drop in the final stage in our study, countering Roca de Larios et al.'s (2008) results who found an increase in revision after the initial stage. Our observation of decreased revision towards the end of writing may be associated with the logographic Chinese writing system. During stage 5, our participants seemed to engage in systematic reviewing, indicated by the drastic reduction in pre-contextual revision signalling the end of the initial drafting stage. Systematic reviewing involves reading one's previously written text, which might be more cognitively demanding for L2 writers of logographic systems given the difficulty posed by character recognition (Gunderson et al., 2011). This, in turn, is likely to make systematic reviewing less productive due to working memory limitations (Kellogg, 1996), leading L2 writers to prioritise correcting small language mistakes over larger chunks of text involving more characters. This account is aligned with the stimulated recall data, which saw an increase in language-oriented contextual revision in stage 5. Another explanation for fewer revisions in stage 5 may be related to writers' inability

to self-correct due to their insufficient linguistic knowledge of L2 Chinese.

Surprisingly, unlike L2 writers who revised language more frequently in stage 5, L1 writers made proportionally more language-oriented revisions in stage 1. This difference might be explained by L1 writers' beliefs about what makes a good Chinese text. Traditionally, an appealing beginning is considered to be a key feature of a good Chinese text (Tao, 2012). One way to achieve this is through using idioms or parallel structures in the opening paragraph. It is possible that L1 writers, being aware of this belief, revised language more often in stage 1 to make the beginning more impressive.

Conclusions and implications

This study examined revision patterns in Chinese writing from a multidimensional and dynamic perspective. The results revealed that L2 Chinese writers revised language more often than content, with most revisions involving smaller textual units and occurring at the point of inscription. L2 Chinese writers also made more frequent changes to Pinyin than characters. Differences in revision patterns mainly set apart the initial and/or final stage from the middle writing stages. L1 writers largely demonstrated similar patterns, the only notable differences being that they made more changes to characters and engaged in a larger proportion of content-oriented revisions. Our findings for L2 Chinese writers are largely in line with those obtained for L2 English writers, extending the observation that revision possesses are multi-dimensional and dynamic to non-alphabetic language writing.

In our analyses, a revision dimension unique to writing in a logographic language

has also emerged. We added level of transcription, Pinyin versus character, to existing revision taxonomies. The differences we observed between L1 and L2 Chinese writers in terms of this dimension suggest that transcribing in an orthographic system different from one's L1 may pose extra demands during writing. That is, writers' L2 orthographic knowledge could be a crucial factor in determining L2 Chinese writing difficulty.

Some tentative pedagogical implications can also be drawn based on this study. The results suggest that, apart from the difficulty of encoding language, a major obstacle faced by L2 Chinese writers is associated with transcribing and reading logographic characters. This indicates the importance of developing learners' knowledge of sound-form connections in L2 Chinese (writing) instruction. Other than the difficulties linked to producing and processing Chinese characters, the findings demonstrated largely similar revision patterns between L2 Chinese and L2 English writers, suggesting that pedagogical implications derived from L2 English revision/writing studies can potentially be applied to L2 Chinese writing. For example, L2 Chinese writers, similar to their L2 English counterparts, could be advised to balance language and textual concerns when revising their text, rather than focusing primarily on linguistic problems (e.g., Stevenson et al., 2006). Probably, L2 Chinese writers, like L2 English users, would also benefit from instruction about how to allocate time and attentional resources during the writing process to achieve more efficient revisions (Barkaoui, 2016). However, it should be noted that this study was conducted in a laboratory setting. Thus, follow-up studies in real-world writing situations are warranted to investigate how these pedagogical recommendations could be effectively

implemented in L2 writing classrooms.

It is also important to acknowledge the limitations of the study. One methodological flaw concerns the partial triangulation of the data, as stimulated recall interviews were conducted based on recordings of writing performance at the character level. It is possible that more language-oriented revisions would have been recalled if the participants had had access to stimuli at the Pinyin level. Future studies could use a screen recorder to capture writing processes at both Pinyin and character levels to create a more effective stimulus for recall. Another shortcoming is that the findings provide limited information about writers' viewing behaviours. Given that the process of writing is shaped by the text previously produced (Galbraith, 1999), it is impossible to understand revision fully without exploring the interaction between re-reading and revision. Future researchers could incorporate eye-tracking to explore viewing behaviours before and after revision.

CRediT author statement

Xiaojun Lu: Conceptualization, Methodology, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Visualisation, Project administration, Andrea Révész: Conceptualization, Methodology, Writing – review & editing, Supervision.

Acknowledgement

This study was funded by a joint doctoral scholarship awarded to the first author by

University College London, Institute of Education and the China Scholarship Council, No. 201606530021.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.system.2021.102544.

References

Bachman, L. F. (1986). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly*, *19*(3), 535–556.

Barkaoui, K. (2016). What and when second-language learners revise when responding to timed writing tasks on the computer: The roles of task type, second language proficiency, and keyboarding skills. *Modern Language Journal*, *100*(1), 320–340.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Lawrence Erlbaum Associates.

Carl, M. (2012). *Translog-II*. 13th International Conference on Intelligent Text Processing and Computational Linguistics.

Chenoweth, N. A., & Hayes, J. R. (2001). Fluency in writing: Generating text in L1 and L2. *Written Communication*, *18*(1), 80–98.

Choi, Y. H. (2007). On-line revision behaviors in EFL writing process. *English Teaching*, *62*(4), 69–93.

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition*, *26*(1), 59–84.

Ericsson, A. K., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data (revised edition)*. The MIT Press.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*(4), 365–387.

Galbraith, D. (1999). Writing as a knowledge-constituting process. In M. Torrance & D. Galbraith (Eds.), *Knowing what to write* (pp. 139–160). Amsterdam University Press.

Gánem-Gutiérrez, G. A., & Gilmore, A. (2018). Tracking the real-time evolution of a writing event: Second language writers at different proficiency levels. *Language Learning*, *68*(2), 469–506.

Gass, S. M., & Mackey, A. (2016). *Stimulated recall methodology in applied linguistics and L2 research*. Routledge.

Gellert, A. S., & Elbro, C. (2013). Cloze tests may be quick, but are they dirty? Development and preliminary validation of a cloze test of reading comprehension. *Journal of Psychoeducational Assessment*, *31*(1), 16–28.

Gunderson, L., Odo, D. M., & D'Sliva, R. (2011). Second language literacy. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 472–487). Routledge.

Janssen, D., Van Waes, L., & Van Den Bergh, H. (1996). Effects of thinking aloud on writing processes. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 233–250). Lawrence Erlbaum Associates.

Kang, H. (2011). *Computer-based writing and paper-based writing: A study of beginning-level and Intermediate-level Chinese learners' writing* [PhD Dissertation]. The Ohio State University.

Kellogg, R. T. (1996). A model of working memory in writing. In C. M. Levy & S.

Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 57–72). Lawrence Erlbaum Associates.

Khuder, B., & Harwood, N. (2015). L2 writing in test and non-test situations: Process and product. *Journal of Writing Research*, *6*(3), 233–278. Scopus.

Lindgren, E., & Sullivan, K. P. H. (2006a). Analysing online revision. In K. P. H. Sullivan & E. Lindgren (Eds.), *Computer key-stroke logging and writing* (pp. 157–188). Elsevier.

Lindgren, E., & Sullivan, K. P. H. (2006b). Writing and the analysis of revision: An overview. In K. P. H. Sullivan & E. Lindgren (Eds.), *Computer key-stroke logging and writing* (pp. 31–44). Elsevier.

Matsuhashi, A. (1987). Revising the plan and altering the text. In A. Matsuhashi (Ed.), *Writing in real time: Modelling production processes* (pp. 197–223). Alex Publishing Corporation.

Murphy, L., & Roca de Larios, J. (2010). Searching for words: One strategic use of the mother tongue by advanced Spanish EFL writers. *Journal of Second Language Writing*, *19*(2), 61–81.

Ong, J. (2014). How do planning time and task conditions affect metacognitive processes of L2 writers? *Journal of Second Language Writing*, *23*(Mar), 17–30.

Piolat, A. (1997). Writer's assessment and evaluation of their texts. In *Encyclopaedia of language and education* (Vol. 7, pp. 189–198). Longman.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912.

Raimes, A. (1985). What unskilled ESL students do as they write: A classroom study of composing. *TESOL Quarterly*, *19*(2), 229–258.

Révész, A., & Michel, M. (2019). Introduction to the special issue. *Studies in Second Language Acquisition*, *41*(3), 491–501.

Roca de Larios, J., Manchón, R. M., Murphy, L., & Marín, J. (2008). The foreign language writer's strategic behaviour in the allocation of time to writing processes. *Journal of Second Language Writing*, *17*(1), 30–47.

Stevenson, M., Schoonen, R., & De Glopper, K. (2006). Revising in two languages: A multi-dimensional comparison of online writing revisions in L1 and FL. *Journal of Second Language Writing*, *15*(3), 201–233.

Tao, Z. (2012). *Nancun chuo geng lu (Records of discontinuing farming)*. Shanghai Guji Chubanshe (Shanghai Ancient Works Publishing House).

Tillema, M. (2012). *Writing in first and second language: Empirical studies on text quality and writing processes*. LOT.

Van Den Bergh, H., & Rijlaarsdam, G. (1996). The dynamic of composing: Modeling writing process data. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 207–233). Lawrence Erlbaum Associates.

Van Waes, L., Leijten, M., & Van Weijen, D. (2009). Keystroke logging in writing research: Observing writing processes with Inputlog. *GFL - German as a Foreign Language*, *2*, 41–64.

Xiao, J., Liu, B., & Wang, X. (2007). Exploiting pinyin constraints in pinyin-to-

character conversion task: A class-based maximum entropy markov model approach. *International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 3, September 2007: Special Issue on Invited Papers from ISCSLP 2006*, *12*(3), 325–348.

Xu, C. (2018). Understanding online revisions in L2 writing: A computer keystroke-log perspective. *System*, *78*, 104–114.

Zhang, D. (2017). Word reading in L1 and L2 learners of Chinese: Similarities and differences in the functioning of component processes. *The Modern Language Journal*, *101*(2), 391–411.

Zimmermann, R. (2000). L2 writing: Subprocesses, a model of formulating and empirical findings. *Learning and Instruction*, *10*(1), 73–99.

Table 1: *Coding scheme for keystroke logs*

Dimension	Operationalisation
Linguistic domain	
Below word	Changes made to Pinyin before converting into a character (e.g. [K][A][I][S]◀[X][I][N]*开心* (happy))
	Changes made to character(s) within a word (e.g. [Y][I][N][W][E][I]*因为* (because)◀[C][I]*此* (thus))
	Substitution of a character with another with the same spelling in Pinyin (e.g. [Q][I][E]*切* (cut)◀[Q][I][E]*且* (in addition))
Word	Changes made to a word in Pinyin or characters
Below clause	Changes made to part of a clause but more than a word in Pinyin or characters
Clause and above	Changes made to more than clauses in Pinyin or characters
Context	
Pre-contextual	Changes made at the point of inscription in Pinyin or characters
Context	Changes made to already-written text in characters
Level of transcription	
Pinyin	Addition, deletion, or substitution of Pinyin
Character	Addition, deletion, or substitution of character(s)

Table 2: *Coding scheme for stimulated recall comments*

Comments	Orientation - subcategory	Linguistic domain	Context
After I wrote *同意* (agree), I was still thinking whether it was how it happened. Then I realised it was not, so I deleted that.	Content - Idea	Below clause	Pre-contextual
Because then I decided to put the last bit at the beginning.	Content - Organisation	Clause and above	Contextual
I looked at this bit, and it looked wrong. I didn't think it was right, so I tried to rephrase.	Language - Translation in general	Below clause	Contextual
I didn't think *自作主张* (take it upon oneself) was a suitable word to describe what the dean of my department did.	Language - Lexis	Word	Pre-contextual
I thought this part was redundant in terms of the sentence structure.	Language - Grammar	Below clause	Pre-contextual
I thought *每当* (whenever) was not appropriate, so I had to replace it with another cohesive device.	Language - Cohesion	Word	Pre-contextual

Table 3: *Revision behaviours by dimension (standardised by total writing time in minutes)*

	L2 Chinese ($N = 32$)				L1 Chinese ($N = 32$)			
Dimension	*M*	*SD*	95% *CI* low	95% *CI* up	*M*	*SD*	95% *CI* low	95% *CI* up
Linguistic domain								
Below word	1.85	1.51	1.59	2.11	2.51	.95	2.34	2.67
Word	1.06	.65	.95	1.17	1.62	.63	1.51	1.73
Below clause	.65	.40	.58	.72	1.56	.66	1.45	1.67
Clause and above	.15	.13	.13	.17	.27	.21	.23	.31
Context								
Pre-contextual	3.30	2.51	2.86	3.73	5.55	1.93	5.22	5.88
Contextual	.57	.42	.50	.64	.63	.48	.55	.71
Level of transcription								
Pinyin	2.28	1.62	2.00	2.56	2.70	1.08	2.51	2.89
Character	1.42	1.04	1.24	1.60	3.25	1.36	3.01	3.49

Note. Full descriptive data (including data by stages) are available in Table S3 in Supplemental material.

Table 4: *Results from linear mixed-effects regressions examining the effect of dimension on different types of revision*

	L2 Chinese ($N = 32$)				L1 Chinese ($N = 32$)			
Fixed effect: dimension	*E*	*SE*	*t*	*p*	*E*	*SE*	*t*	*p*
Linguistic domain (below clause as reference)								
Below word	.49	.02	22.17	< .01	.31	.02	13.95	< .01
Word	.21	.02	9.49	< .01	.03	.02	1.24	.22
Clause and above	-.34	.02	-15.50	< .01	-.68	.02	-30.82	< .01
L2 model = lmer(log(frequency) ~ linguistic.domain + (1|participant), L2data, REML=F), $R^2_m = .60$								
L1 model = lmer(log(frequency) ~ linguistic.domain + (1|participant) + (1|prompt), L1data, REML=F), $R^2_m = .71$								
Context (contextual as reference)								
Pre-contextual	.94	.03	30.98	< .01	1.39	.03	47.10	< .01
L2 model = lmer(log(frequency) ~ context + (1|participant), L2data, REML=F), $R^2_m = .69$								
L1 model = lmer(log(frequency) ~ context + (1|participant), L1data, REML=F), $R^2_m = .86$								
Level of transcription (character as reference)								
Pinyin	.29	.03	10.62	< .01	-.13	.03	-4.52	< .01
L2 model = lmer(log(frequency) ~ level.of.transcription + (1|participant), L2data, REML=F), $R^2_m = .15$								
L1 model = lmer(log(frequency) ~ level.of.transcription + (1|participant) + (1|prompt), L1data, REML=F), $R^2_m = .04$								

Table 5: *Significant results from post hoc Bonferroni test comparing the amount of revision at four levels of linguistic domain*

	L2 writers ($N = 32$)			L1 writers ($N = 32$)		
Linguistic domain	*SE*	*p*	*d*	*SE*	*p*	*d*
Below word - Word	.02	< .01	.88	.02	< .01	1.10
Below word - Below clause	.02	< .01	1.71	.02	< .01	1.10
Below word - Clause and above	.02	< .01	2.39	.02	< .01	3.20
Word - Below clause	.02	< .01	1.15			
Word - Clause and above	.02	< .01	2.54	.02	< .01	3.23
Below clause - Clause and above	.02	< .01	1.92	.02	< .01	2.83

Table 6: *Stimulated recall comments by linguistic domain*

	Content			Language					Others	No recall	Total
	Idea	Organisation	Total (%)	Translation in general	Lexis	Syntax	Cohesion	Total (%)	Total (%)	Total (%)	Total (%)
L2 Chinese (N =32)											
Below word	2	0	2 (0%)	0	49	5	3	57 (6%)	0 (0%)	3 (0%)	62 (7%)
Word	50	4	54 (6%)	34	101	101	24	260 (28%)	0 (0%)	35 (4%)	349 (37%)
Below clause	78	8	86 (9%)	46	69	97	13	225 (24%)	2 (0%)	24 (3%)	337 (36%)
Clause and above	66	25	91 (10%)	14	11	50	14	89 (9%)	5 (1%)	12 (1%)	197 (21%)
Total	196	37	233 (25%)	94	230	253	54	631 (67%)	7 (1%)	74 (8%)	945 (100%)
L1 Chinese (N =32)											
Below word	11	0	11 (1%)	2	19	3	0	24 (2%)	0 (0%)	2 (0%)	37 (2%)
Word	166	13	179 (12%)	62	181	92	50	385 (26%)	5 (0%)	41 (3%)	610 (41%)
Below clause	176	11	187 (13%)	128	94	71	27	320 (22%)	0 (0%)	45 (3%)	552 (37%)
Clause and above	131	33	164 (11%)	28	11	43	17	99 (7%)	2 (0%)	17 (1%)	282 (19%)
Total	484	57	541 (37%)	220	305	209	94	828 (56%)	7 (0%)	105 (7%)	1481 (100%)

Table 7: *Stimulated recall comments by context*

	Content			Language					Others	No recall	Total
	Idea	Organisation	Total (%)	Translation in general	Lexis	Grammar	Cohesion	Total (%)	Total (%)	Total (%)	Total (%)
L2 Chinese (N =32)											
Pre-contextual	109	12	121 (13%)	55	183	115	26	379 (40%)	6 (1%)	55 (6%)	561 (59%)
Contextual	87	25	112 (12%)	39	47	138	28	252 (27%)	1 (0%)	19 (2%)	384 (41%)
Total	196	37	233 (25%)	94	230	253	54	631 (67%)	7 (1%)	74 (8%)	945 (100%)
L1 Chinese (N =32)											
Pre-contextual	382	31	413 (28%)	183	251	156	53	643 (43%)	5 (0%)	95 (6%)	1156 (78%)
Contextual	102	26	128 (9%)	37	54	53	41	185 (12%)	2 (0%)	10 (1%)	325 (22%)
Total	484	57	541 (37%)	220	305	209	94	828 (56%)	7 (0%)	105 (7%)	1481 (100%)

Table 8: *Significant results from post hoc Bonferroni tests comparing the amount of revision between stages*

	L2 Chinese ($N = 32$)				L1 Chinese ($N = 32$)			
	Stage	*SE*	*p*	*d*	Stage	*SE*	*p*	*d*
Total	S1 < S2	.03	< .01	.56	S1 < S2	.21	< .01	.51
	S1 < S3	.03	< .01	.68	S1 < S3	.21	< .01	.54
	S1 < S4	.03	< .01	.66	S1 < S4	.21	< .01	.54
	S1 < S5	.03	.01	.26	S5 < S3	.21	.01	.37
	S5 < S3	.03	< .01	.32	S5 < S4	.21	.01	.38
	S5 < S4	.03	.01	.35				
Linguistic domain								
Below word	S1 < S2	.03	.01	.32	S1 < S2	.04	< .01	.35
	S1 < S3	.03	< .01	.39	S1 < S3	.04	< .01	.42
	S1 < S4	.03	< .01	.43	S1 < S4	.04	< .01	.41
	S5 < S2	.03	< .01	.30				
	S5 < S3	.03	< .01	.34				
	S5 < S4	.03	< .01	.41				
Word	S1 < S2	.03	.01	.33	S1 < S3	.03	< .01	.29
	S1 < S3	.03	< .01	.43	S1 < S4	.03	< .01	.36
	S1 < S4	.03	< .01	.32				
Below clause	S1 < S2	.03	< .01	.40	S1 < S2	.03	< .01	.39
	S1 < S3	.03	< .01	.39	S1 < S3	.03	< .01	.33
	S1 < S4	.03	< .01	.35	S1 < S4	.03	< .01	.36
Context								
Pre-contextual	S1 < S2	.04	< .01	.44	S1 < S2	.21	< .01	.43
	S1 < S3	.04	< .01	.55	S1 < S3	.21	< .01	.48

	S1 < S4	.04	< .01	.37	S1 < S4	.21	< .01	.50
	S5 < S2	.04	< .01	.47	S5 < S3	.21	< .01	.41
	S5 < S3	.04	< .01	.49	S5 < S4	.21	< .01	.46
	S5 < S4	.04	< .01	.43				
Contextual	S1 < S2	.03	.01	.37	S1 < S5	.03	< .01	.39
	S1 < S3	.03	< .01	.34				
	S1 < S4	.03	< .01	.59				
	S1 < S5	.03	< .01	.79				
	S2 < S5	.03	< .01	.52				
	S3 < S5	.03	< .01	.47				
	S4 < S5	.03	.01	.29				
Level of transcription								
Pinyin	S1 < S2	.04	< .01	.37	S1 < S2	.04	< .01	.37
	S1 < S3	.04	< .01	.45	S1 < S3	.04	< .01	.37
	S1 < S4	.04	< .01	.34	S1 < S4	.04	< .01	.40
	S5 < S2	.04	< .01	.45	S5 < S4	.04	< .01	.36
	S5 < S3	.04	< .01	.49				
	S5 < S4	.04	< .01	.46				
Character	S1 < S2	.03	< .01	.48	S1 < S2	.14	< .01	.42
	S1 < S3	.03	< .01	.49	S1 < S3	.14	< .01	.47
	S1 < S4	.03	< .01	.67	S1 < S4	.14	< .01	.45
	S1 < S5	.03	< .01	.67				

Table 9: *Stimulated recall comments by stage*

		Content			Language					Others	No recall	Total
Stage		Idea	Organisation	Total (%)	Translation in general	Lexical retrieval	Syntactic encoding	Cohesion	Total (%)	Total (%)	Total (%)	Total (%)
L2 Chinese (N = 32)												
1	Total	38	9	47 (28%)	13	43	44	11	111 (65%)	1 (0%)	11 (6%)	170 (100%)
2	Total	37	9	46 (22%)	24	56	50	11	141 (68%)	1 (0%)	18 (9%)	206 (100%)
3	Total	31	9	40 (22%)	21	52	43	7	123 (68%)	2 (1%)	17 (9%)	182 (100%)
4	Total	46	4	50 (28%)	18	43	43	9	113 (63%)	0 (0%)	17 (9%)	180 (100%)
5	Total	44	6	50 (24%)	18	39	72	16	145 (70%)	0 (0%)	11 (5%)	207 (100%)
L1 Chinese (N =32)												
1	Total	63	12	75 (31%)	40	55	37	17	149 (62%)	0 (0%)	16 (7%)	240 (100%)
2	Total	110	12	122 (37%)	36	67	57	16	176 (53%)	2 (1%)	29 (9%)	329 (100%)
3	Total	114	12	126 (39%)	40	77	42	17	176 (55%)	1 (0%)	17 (5%)	320 (100%)
4	Total	103	8	111 (39%)	58	42	38	16	154 (54%)	2 (1%)	18 (6%)	285 (100%)
5	Total	94	13	107 (35%)	46	64	35	28	173 (56%)	2 (1%)	25 (8%)	307 (100%)

Note. Stage-wise data by linguistic domain and context are available in Tables S5 and S6 in Supplemental material, respectively.

Captions Figure 1

Figure 1: *Typing the Pinyin for multiple characters*

Figure 2: *Data collection procedures*

Session 1	Session 2
Writing Task 1 (30 minutes)	Writing Task 1 (30 minutes)
Break (5 minutes)	Break (5 minutes)
Writing Task 2 (30 minutes)	Writing Task 2 (30 minutes)
Break (5 minutes)	Break (5 minutes)
Background questionnaire (10 minutes)	Stimulated recall (30 to 60 minutes)
	Break (5 minutes)
	L2 proficiency test (L2 writers only, 30 minutes)