University of Groningen

Aligning keystrokes with cognitive processes in writing

Galbraith, David; Baaijen, Veerle M

Published in:
Observing Writing

DOI:
[10.1163/9789004392526_015](https://doi.org/10.1163/9789004392526_015)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2019

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Galbraith, D., & Baaijen, V. M. (2019). Aligning keystrokes with cognitive processes in writing. In E. Lindgren, & K. P. H. Sullivan (Eds.), *Observing Writing: Insights from Keystroke Logging and Handwriting* (Vol. 38, pp. 306-325). (Studies in Writing Series; Vol. 38). Brill.
https://doi.org/10.1163/9789004392526_015

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Aligning Keystrokes with Cognitive Processes in Writing

David Galbraith and Veerle M. Baaijen

Keystroke logging provides a detailed record of the process of writing as it unfolds in time. However, by themselves, the measures it provides—pauses, bursts and revisions—do not provide an unambiguous interpretation of the cognitive processes involved, particularly when the measures are aggregated across texts and writers. We have described this as a problem of alignment (Baaijen, Galbraith & de Glopper, 2012; see also Wengelin, Frid, Johansson & Johansson, this volume, pp. 30–49). How does one map keystroke units of analysis to specific components of the writing process? Pauses may reflect not just various different levels of planning and reflection, but also re-reading and text production during revision. Bursts of language may reflect an initial formulation of thought or an attempt to improve previously formulated text. Revisions may reflect semi-automatic correction of errors or a systematic attempt to modify content. In this chapter we will illustrate this problem of alignment with the help of Hayes' (2009) model of text production. We will then describe procedures that we have used in order to increase the alignment between keystroke log measures and cognitive models of writing. Finally, we will demonstrate that when these measures are combined into more overarching measures they show systematic relationships with text quality and the development of understanding through writing.

1 Problems of Alignment

To illustrate the problem of alignment between keystroke logging measures and cognitive models of writing we will briefly outline Hayes' model of text production (see Hayes, 2009, for a review), and consider how the basic units of keystroke logs can be interpreted in terms of the model. A sketch of the model, taken from Chenoweth and Hayes (2003) is shown in Figure 13.1.

The model distinguishes between four different processes. The *proposer* proposes ideas for expression. This component is assumed to include the higher level processes involved in planning and reflection as characterised in global

FIGURE 13.1 Model of the text production process from Chenoweth and Hayes (2003, p. 113)

models of the writing process, and is responsible for creating an idea package to be formulated in language. Essentially, it is responsible for deciding what to say next. The *translator* is responsible for converting this message into linguistic strings. The *transcriber* then converts this linguistic string into written text. Finally, the *evaluator/reviser* is responsible for monitoring and evaluating word strings and text as they are produced, and for revising them when they are found wanting. Revision may consist either of revision of the language currently being produced (which we call revision at the leading edge) or more global revision of the text as a whole (see Lindgren, Westum, Outakoski & Sullivan, this volume, pp. 346–365, for an extended discussion of revision at the leading edge). The processes involved in text production, therefore, are embedded within the reflective and revision processes characteristic of the writing process as a whole (Hayes, 1996).

In its general form, this model corresponds directly with what has become the standard model of spoken language production (Levelt, 1989, 1999; Vigliocco & Hartsuiker, 2002). In particular, it distinguishes between the conceptual processes involved in creating content and the linguistic processes involved in formulating content in language. (Note, though, that it remains silent about the extent to which this is a strictly serial process or incorporates feedback

from lower levels back to higher levels of conceptual processing.) One might expect therefore that one could draw on the long tradition of pause analysis of spontaneous speech (Allum & Wheeldon, 2007; Goldman-Eisler 1961, 1968; Levelt, 1989) to aid the interpretation of pauses in writing. However, there are important differences between speech and writing which make it even harder to interpret pauses during writing compared to pauses in speech.

The most important of these, in our view, is the extent to which internal and external monitoring can be carried out. Vigliocco and Hartsuiker (2002) have suggested that, although inner speech can in principle be monitored, this cannot be done at the same time as speaking, arguing that externally articulated speech may overwrite the representation of inner speech in short term verbal memory. If Vigliocco and Hartsuiker are right, speaking prevents the monitoring of inner speech. By contrast, writing, partly because of its slower output, but mainly because it is produced manually rather than vocally, allows—indeed encourages—monitoring of inner speech. In extreme cases, this can turn writing into a form of permanent stuttering. More prosaically, this means that, whereas pauses during speech are a relatively direct reflection of planning, pauses during writing may reflect revision of both the content and form of planned language before transcription.

This is clearly illustrated in a study by Kaufer, Hayes and Flower (1986), which compared thinking-aloud protocols produced by expert and novice writers. Kaufer et al. segmented the protocols into language bursts. Bursts that ended in a pause of two or more seconds were classified as P-bursts; bursts that were terminated by an evaluation, revision or some other grammatical discontinuity were classified as R-bursts (Chenoweth & Hayes, 2001; Hayes, 2009). They give as an example a segment of the protocol corresponding to the production of the sentence, “The best thing about it is that it allows me the opportunity to use my mind and my ideas in a productive way.” In a keystroke log, with PAUSES defined as intervals of two seconds or more between keystrokes, this would appear as, **“The best thing about it is that PAUSE it allows me the opportunity to PAUSE [(DELETION OF “the opportunity to”), (replacement with to use)] PAUSE my mind and my ideas in a productive way. PAUSE”**. Using their criteria for the definition of bursts and pauses, this sentence would be analysed as consisting of a P-burst followed by an R-burst, a brief P-burst revising the preceding R-burst, and a final P-burst. Later in the chapter, we will argue that the brief P-burst revising an earlier burst should not be classified as a P-burst but, setting that aside for the moment, a keystroke analysis would therefore indicate that this sentence was produced in three (perhaps two) P-bursts and one R-burst. By contrast, Kaufer et al.’s analysis of the spoken protocol suggests that it consists of six P-bursts and an R-burst (Kaufer et al., 1986, p. 125). The disparity is

because a great many of the bursts that are taken to reflect the output of the translator are not apparent in writing since they are rejected before they are written down. For example, during the first pause in the above example, the think aloud protocol shows that the writer proposes several possible continuations before they actually write the next sentence part.

This simple example of a writer producing a single sentence demonstrates three important features of keystroke analysis. First, even in this simple example, where cognitive processing is focussed on the creation of the content and language for the current sentence, pauses are multiply determined. In particular, they include not just conceptual and linguistic planning but also re-reading and revision of language prior to transcription. The range of possible processes expands even further when one considers the production of extended text. In what follows, we suggest ways in which the range of potential processes can be narrowed down for analysis.

Second, Hayes (2009) has suggested that the length of P bursts may reflect the capacity of the translator and, more generally, individual differences in language skills. This rests on the assumption that the length of P bursts (clean segments of language produced between 2 second pauses) reflects the quantity of material that can be formulated in one planning episode, whereas R-bursts reflect output that has been interrupted prior to completion. Under this hypothesis only P-bursts should be used to compare the language capacities of individuals. The comparison we have just made of keystroke logs with think aloud protocols suggests however that, even if analysis is restricted to P-burst length, written burst size may not simply reflect the capacity of the translator. Thus, while the written burst, "it allows me the opportunity to", is preceded solely by conceptual planning in the think aloud protocol, the burst at the end of the sentence, "my mind and my ideas in a productive way" is preceded in the spoken protocol by both conceptual planning and by a series of attempts at formulation which are evaluated and modified prior to transcription. We will argue in what follows that the key feature of bursts produced during extended text production is how cleanly they are produced rather than their length, and that this depends crucially on the writer's strategy for combining translation and revision rather than on their linguistic capacities alone.

Finally, given that think aloud protocols provide so much more detailed information about the processes taking place between segments of transcription, one might conclude that they provide a more effective method than keystroke logs, and that they therefore should be used instead of, or alongside, keystroke analysis. There are a number of reasons for rejecting this conclusion. First, think aloud protocols only provide information about relatively high level processes accessible to introspection; keystroke analysis provides much

more detailed, albeit less directly interpretable, information about moment by moment variations in processing. Second, think aloud protocols necessarily impose a serial order on reporting of processes, and may therefore underestimate the extent of parallel processing during text production (Alamargot, Dansac, Chesnet, & Fayol, 2007). Third, and most importantly, given the potential difference in capacity for monitoring inner speech, spoken and written text production may involve fundamentally different kinds of processing.

2 Procedures for Increasing Alignment between Keystrokes and Cognitive Processes

A natural first impulse when assessing the characteristics of a keystroke log is to average the various measures across the whole log. The value of doing this depends to a certain extent on the research question. Thus, Chenu, Pellegrino, Jisa, and Fayol (2014) (in a study of handwriting) were concerned with establishing a threshold for distinguishing between cognitive pauses and graphomotor pauses, and hence analysed all the pauses greater than 15ms present in the log. In studies concerned more specifically with questions about the different kinds of processes involved, however, a more discriminating approach is required in order to relate keystroke measures more directly with cognitive processes.

In a study using Inputlog (Leijten & van Waes, 2006; 2013), Baaijen et al. (2012) were concerned more specifically with the characteristics of text production and so took a number of steps to isolate this from other aspects of the writing process. These included the exclusion of titles, episodes of explicit planning, and revision carried out after completion of the initial draft (see Baaijen et al., 2012, for full details). Thus, the analysis of text production was restricted to sections of the log involving the production of full text during the initial drafting of text.

3 Bursts

As we have already mentioned, one approach to analysing the characteristics of text production is to divide the text up into individual bursts of language (Chenoweth & Hayes, 2001; Kaufer et al., 1986). There are a range of approaches to this. Some studies (e.g., Alves, Castro, Sousa & Strömquist, 2007) simply divide up the text according to the location of pauses greater than two seconds. Others (e.g., Lindgren, Spelman Miller & Sullivan, 2008) use both pauses

and revisions to identify burst boundaries but then treat these as equivalent, and calculate a single estimate of burst length. Kaufer et al. (1986), as mentioned above, distinguish between bursts terminated by a pause greater than 2 seconds in length, called P-bursts, and bursts terminated by a revision, called R-bursts. P-bursts are assumed to represent complete units of language production, and hence to reflect the capacity of the translator component of text production. The rationale being that 2 second intervals represent “conceptual planning” episodes, and hence occur when the translator has run out of content and time is needed to create the next item of content. By contrast, R-bursts are assumed to represent episodes where evaluation has led to the termination of a burst before completion, and hence not to reflect the capacity of the translator. (See Alves, Leal, & Limpo, this volume, pp. 50–70; Breuer, this volume, pp. 190–211; Leijten, Van Horenbeeck, & Van Waes, this volume, pp. 71–95, for examples of studies analysing bursts. Note that these studies vary in whether they distinguish between P- and R-bursts, and that it is not always clear whether P-bursts are defined as including or excluding revisions within bursts).

Baaijen et al. (2012) found in their sample of university students writing articles for the student newspaper that P-bursts (with a mean length of approximately six words) were significantly longer than R-bursts (approximately 4.5 words). In addition, bursts consisting of language inserted elsewhere in the text were significantly shorter in length (approximately 2 words) than either P- or R-bursts. We say “approximately” here because their analysis included a range of sub-types of bursts, varying in the nature of the link between P-bursts and preceding content and in the type of revision carried out at the end of an R-burst (see Baaijen et al., 2012, for full details). The most important of these were bursts involving the revision of a previous segment like the example we gave from Kaufer et al. in the introduction. These were typically much shorter than other P-bursts (less than 2 words in length on average), and hence we would argue that they should not be classified as P-bursts.

These data suggest that there are important differences between different types of burst, and that in order to achieve a better alignment with underlying cognitive processes these differences need to be taken into account in the analysis. In particular, to the extent that bursts do give an indication of the capacity of the translator, these estimates should be based specifically on P-bursts rather than bursts in general. That said, we would also question whether bursts in a keystroke log can be interpreted in this way, given that pauses preceding them may include attempts at language formulation (as suggested by the think aloud protocols collected by Kaufer et al., 1986).

4 Distinguishing Linear from Non-Linear Transitions between Units of Text

Having separated out text production during the initial drafting of text from explicit planning episodes before and during writing (often in note form) and from text produced during post-draft revision, Baaijen et al. (2012) then further distinguished within text production between episodes of continuous text production and episodes involving an interruption of text production. This distinguished *linear* transitions between units of text, where the writer maintained the forward progression of text production without interruption, from *non-linear* transitions between units—defined as *events*—where the writer carried out some other action before continuing with the next unit. These interruptions ranged from simple actions (such as mouse movement or scrolling that might indicate re-reading or evaluating previously produced text), through movements across larger units (such as the deletion of a word earlier in the current sentence) up to substantial interruptions, in which the writer moved further back in the text to carry out revisions and then returned to carry on with text production at the leading edge of the text. The aim was to separate transitions primarily concerned with preparation of the next unit of text from transitions concerned with revising or re-reading earlier units of text.

In combination, these procedures meant that analysis of the transitions within text production was restricted to those sections of the text where production was at its most “pure”, uninterrupted by global planning and revision processes, and by more local revision of ongoing text production. Pauses were estimated purely between units produced as part of the forward progression of text production, where it was assumed that processing was most likely to be directly concerned with the production of the next unit in the text.

5 Analysing Pauses

A general feature of the research on keystrokes during the production of full text has been a focus on higher level thinking processes rather than the more local processes involved in formulating thought in language. The standard method for analysing pauses during text production has been to use a threshold to distinguish “cognitive” pauses, assumed to reflect higher level thinking, from briefer pauses, assumed to be associated with lower level transcription and translation processes, and then to examine where these occur in the text. Although this can be informative, and we include such measures as part of our analyses, it has the drawback that it neglects variation in processes below this

threshold. Furthermore, it is difficult to establish what an appropriate threshold is, and this is the source of much debate. We have therefore approached this from the opposite direction (Baaijen et al., 2012). We identify the boundaries at different levels within the text—within words, between words, between sentences, and between paragraphs—and then estimate the extent to which the text is continued linearly at these boundaries, and when it is, how long these linear transitions are. This has the important consequence that it enables us to assess variations between writers in how they manage the transitions between units at different levels of text production. In particular, it enables us to identify variations in how sentence production is carried out.

A characteristic feature of even this sub-sample of pauses associated only with linear text production is that they do not form a homogeneous set. In part, this is due to the fact that they reflect different levels of language planning. Thus, pauses at paragraph boundaries are typically longer than pauses at sentence boundaries which are in turn longer than pauses between words, which are longer than pauses within words (e.g., Schilperoord, 1996; see also Leijten, Van Horenbeeck, & Van Waes, this volume, pp. 71–95; Linnemann, this volume, pp. 326–345). Baaijen et al. (2012) replicated this finding for the pauses associated just with linear transitions. However, they also found that even within these narrower sub-categories at different text boundaries, there was clear evidence that these did not form a homogeneous category.

Pause lengths, in both speech and writing, tend to have highly positively skewed distributions, with the majority of pauses being relatively short, but with a minority being much longer. This makes the mean an overestimate of the length of a typical pause. One approach to this is to use the median to represent the central tendency of scores or to transform the scores—typically using a log transformation—so that they are normally distributed. However, the validity of this depends on whether the pauses have a homogeneous, unimodal distribution. In fact, Baaijen et al. (2012) found that the distribution of pause lengths at between- and within-word boundaries was multimodal. Using mixture models (McLachlan & Peel, 2000) to identify sub-components of these distributions, they found that between-word pauses could be modelled as having 3 separate components. Figure 13.2 shows the distribution of between-word pause lengths for an individual participant, with best-fitting curves overlaid on the plot.

As can be seen in Figure 13.2, even after taking logs, the pauses fall into three distinct clusters. Baaijen et al. (2012) speculated that the left-hand distribution (65 % of the pauses; mean 330 ms) represents word retrieval processes, the middle distribution (26 % of the pauses, mean 735 ms) represents phrase boundary processes, and that the right hand distribution (9 % of the pauses; mean 2697 ms) represents higher level message planning or reflection. Their conclusions

FIGURE 13.2 Frequency distribution of between-word intervals (after $\log_e$ transformation) for a single writer

for this participant were: (i) that there are two normally distributed groups of pauses and that these would be better represented by two lognormal means than by a single median; (ii) that the more extreme pauses should be treated as a miscellaneous group identified by a threshold determined by the tail of the second lognormal distribution. In this case, this is equivalent to a pause duration of 1686 milliseconds. Note, that this case would provide some support for the assumption that a two-second threshold would safely capture only “abnormal” pauses reflecting higher level processes separate from word retrieval, but also highlights the fact that focussing on such pauses neglects important structure at lower levels of processing.

The multimodal nature of the distributions of pause durations was characteristic of both within-word and between-word intervals. There was some variation between individuals in how many clusters there were, with some individuals showing evidence for between-word pauses of three clusters (58 % of the participants), others two clusters (19 %), and others indeterminate between two and three clusters (23 %). Within-word pauses were similarly multimodal in nature, and varied as to whether they consisted of three or four clusters. In general, then, intervals between keystrokes and words are determined by a

number of different processes, and this gives rise to clusters of different pause durations. We would argue that these predominantly involve formulation and transcription processes (with the occasional outlier where the writer has broken off from formulation to think about higher-level content issues). Furthermore, they all fall well below the threshold typically used to define a burst, and hence reflect within-burst processes.

By contrast, pauses at sub-sentence locations (defined by the presence of a within sentence punctuation marker) and sentence locations were best characterised by a single log normal distribution, with the mean for sub-sentence locations being 1211 milliseconds and for sentence locations being 2981 milliseconds. The sub-sentence pauses are something of a mixed bag (they were originally intended as an exclusion category for between-word pauses, designed to restrict the analysis of between word pauses to linear transitions), and vary widely in duration, which in some cases overlap with between-word pause durations. Between-sentence pauses, however, showed much less overlap with the other distributions. We would argue, therefore, that pauses at this location are the ones that are most directly representative of variations in content planning. Other locations may occasionally be associated with longer pauses but these are rare and harder to interpret. Pauses at between-sentence boundaries are consistently longer than others and, in the context of the forward progression of the text, can be interpreted as predominantly concerned with the preparation of the next unit of content. We would therefore suggest that pauses between sentences represent a homogeneous sample to establish variations in content planning.

Before we conclude this section, three caveats are in order about the homogeneous nature of the distribution of between-sentence pauses. First, it should be noted that it is important to carry out a preliminary screening of the between sentence pauses to determine whether they involve linear transitions or events, since in our dataset more than 40% of the pauses happening at between sentence locations were associated with other material than 'empty thinking' time before the next unit of text production. Second, although the texts we have examined are relatively long compared to previous research in the area (about a page and a half of typescript, averaging 650 words), it is possible that distribution of between-sentence pauses may be less homogeneous for longer texts like journal articles or books. Finally, these results may depend on the language proficiency of the writer. L2 writers, in particular, may show less clear-cut distinctions between the different text locations, and would be expected generally to produce a lower proportion of linear transitions between units of text.

6 Relationships between Pauses and Bursts

A natural question to ask is how the amount of time a writer spends preparing content is related to the length of the bursts that they subsequently produce. A straightforward possibility is that the longer time that a writer spends preparing an utterance prior to transcription, the longer the bursts that will be produced. However, Baaijen et al. (2012) found no evidence to support this. Not surprisingly, given that between-word pauses typically occur within bursts, there was no relationship between the length of between-word pauses and burst length. However, for pauses between sub-sentence units and between sentences, the relationship with burst length was actually negative, and this applied for both P-bursts and R-bursts (correlations ranging from $-.41$ to $-.56$).

Although these relationships are consistent, they are hard to interpret. Our suggestion is that they may reflect the extent to which the message is structured prior to transcription. Under this assumption, longer pauses reflect the development of a plan for the content which imposes message structure on bursts. This means that bursts are no longer a direct reflection of linguistic capacity—bursts do not continue until the translator “runs out of steam” (Hayes, 2009)—but rather are constrained by the sub-units in the planned message. This receives some support from the fact that Baaijen et al. (2012) also found that the length of between-sentence pauses was positively related to the proportion of P-bursts produced during text production ($r = .52$). Taken together with the negative relationship with burst length, this suggests that pauses between sentences include content planning and evaluation of formulation prior to transcription. The result being that longer pauses between sentences are associated with cleaner text production—a sequence of shorter, well-formed bursts. An important goal for future research is to establish the generality of this finding, and to investigate the relationship of between-sentence pause time with burst length and quality in a more controlled way.

7 Global Measures of the Writing Process

So far, we have focussed on the characteristics of text production in isolation from the other components of the writing process. A complete description of the writing process requires, however, a characterisation of how text production is combined with planning and revision. In addition, to the sentence production measures that we have already discussed, Baaijen et al. (2012) also measured features of the keystroke logs associated with more global organising processes. These included (see Baaijen et al., 2012 for full details):

- i. A *sentence linearity index*, which compared the order that sentences had in the final product with the order in which they were produced in the keystroke log. Scores ranged from 1, indicating that sentences were produced in the same order in the log as in the text, through to progressively lower scores depending on the proportion of the sentences that were produced in a different order in the log.
- ii. Percentage of time devoted to *non-linear events*, including time spent scrolling through the text (presumably for re-reading) and time spent revising parts of the text away from the current leading edge.
- iii. *Production cycles*, which measured how many continuous episodes of text production were contained within the keystroke log. Lower scores (with text length controlled for) indicated that the text had been produced more continuously, with, at the extreme, a score of 1 indicating that the text had been produced in a perfect, uninterrupted, first draft.
- iv. A range of measures of *linear transition* between units, including the percentage of linear transitions between sentences, which measured the extent to which successive sentences were produced without interruption by events. In this sample, this averaged 67% of the sentence transitions, ranging from 21% to 89%.

Using principal component analysis to identify independent components of the writing process, Baaijen et al. (2012) identified 5 components. Two of these corresponded to sentence production variables, one corresponded to a global organisation variable, and two consisted of only a few measures, corresponding to miscellaneous aspects of the process. More recently, Baaijen and Galbraith (2018) used a more selective set of measures derived from the same data to create two independent scales corresponding to a global linearity and a sentence production variable. These are shown in Table 13.1.

As can be seen in Table 13.1, the two components have good reliability, with Cronbach's alpha around .80 for both components. The *Global Linearity* scale consists of direct measures of how linearly sentences are produced, as well as indicators of what proportion of bursts are inserted away from the leading edge of text production (I-bursts) and of the amount of time spent on non-linear events. High scores represent a very linearly produced text, with successive sentences produced sequentially; low scores represent non-linearly produced text, with sentences being produced recursively. The *Sentence Production* scale consists of measures of pause time between and within sentences combined with measures of within-sentence revision. High scores represent more controlled and clean sentence production, with longer pauses before sentences and relatively little revision during production; low scores represent more spontaneous sentence production, with relatively brief pauses before sentences and more

TABLE 13.1 Summary of principal component analysis with varimax rotation for 2 factor solution

Variables	Rotated factor loadings	
	Component 1 Global linearity	Component 2 Sentence production
1 Sentence linearity index	.849	.200
2 Percentage of I-bursts	-.830	.087
3 Percentage of time spent on events	-.824	-.009
4 Percentage of linear transitions between sentences	.773	.075
5 Number of production cycles	-.709	-.270
6 Percentage of linear transitions between words	.601	.363
7 Percentage of bursts terminated by revision at the leading edge	.113	-.901
8 Percentage words produced in P-bursts	.362	.809
9 Percentage of >2second pauses between words	-.072	.741
10 Text modification index	-.399	-.675
11 Mean pause duration between sentences	.181	.636
Eigenvalues	4.7	2.4
% of variance	42.47	21.40
α	.80	.79

Note that factor loadings over .5 appear in bold

revision as the sentence is produced. The two components are uncorrelated, indicating that texts can be produced by different combinations of these processes: a text may be very linearly produced at the global level, but individual sentences can then be worked out in the course of production; the global structure of a text may be produced very recursively, while at the same time individual sentences may be produced in a clean and controlled manner. (Note that, when the two components were allowed to correlate, there was an indication of a small, positive correlation between the components (.24) but since this fell below the threshold for concluding that there was a relationship between the components (Tabachnick & Fidell, 2001), the orthogonal solution was preferred for the analyses).

An important property of these two measures is that they represent features of the writing process above the level of individual events but also below the global categories used to characterize the writing process as a whole. Thus,

the sentence production variable combines both pause and revision measures to produce a measure of how sentence production varies across texts but at the same time is a more specific category than the text production / translation component of classic models of writing (e.g., Flower & Hayes, 1980; Hayes, 1996). In a think aloud protocol, for example, these variants of sentence production (controlled and spontaneous) would be uniformly classified as translation. Similarly, the global linearity measure is a more specific measure than is embodied in a general claim that writing is a recursive process involving different combinations of planning, translation and revision. Given that these two dimensions are derived from a principal component analysis, we cannot make assumptions about their generality. It may be that different measures may load in different ways on these dimensions when research is carried out on other samples. We would suggest, however, that the general distinction between the processes occurring in sentence production and the processes involved in global organization is likely to be a useful distinction, and that these should be the target for deriving measures from keystrokes.

The final point we would like to make about these two dimensions is that they suggest that sentence boundaries are a crucial “hinge” in the writing process. These boundaries appear to provide information both about the processes involved in planning and formulating sentences and about how the overall organization of the text is developed. This suggests that they should be the focus of analysis in examining keystroke logs. Thus, rather than examining all the pauses above a certain threshold, where it is difficult to establish exactly what the threshold should be, and then to interpret what any given pause may represent, it may be more productive, at least for reasonably long texts produced in a writer's first language, to focus on the pauses at sentence boundaries. An important step here being to distinguish between linear transitions at these boundaries, which we have argued represent sentence preparation, and non-linear events, which we have argued are indicators of global organization.

Focusing on sentence boundaries has the added benefit of enabling links between studies of sentence production in the course of full text production and more controlled experimental studies of written sentence production (e.g., Hayes, 2009; Nottbusch & Torrance, 2011). Such studies may help develop our understanding of the source of the microstructure revealed by mixture modelling of within sentence pauses as well as the interrelationships between pause duration and burst length and quality. In addition, sentence boundaries seem to be particularly useful for examining global organization. By contrast, although paragraph boundaries could potentially be important indicators of global organization insofar as they are often the location of non-linearities associated with within paragraph revision, they do not occur very frequently in

the relatively brief texts we have been discussing here, and furthermore are not very reliable as indicators because we have found that they are often inserted after a draft has been completed.

8 Validity of These Measures

Before we conclude, we want to briefly consider two examples drawn from our research using keystrokes as illustrations of the importance of alignment in the analysis of keystrokes.

The first of these is a study investigating the relationships between writing beliefs, text quality and the development of the writer's understanding (Baaijen, Galbraith & de Glopper, 2014). White and Bruning (2005) distinguished between *transactional* and *transmissional* beliefs about writing. Transactional beliefs are beliefs about the extent to which writing involves the development of the writer's understanding during writing; transmissional beliefs are beliefs about the extent to which writing involves the transmission of information from authoritative sources. White and Bruning (2005) found that writers with a combination of high transactional beliefs and low transmissional beliefs produced higher quality text than writers with the opposite combination (low transactional and high transmissional beliefs).

More recently, Baaijen et al. (2014) replicated the relationship between high transactional beliefs and text quality, but found that this was moderated by the type of planning carried out before writing. In addition, they found that high transactional beliefs were also related to the extent that writers experienced a development of their understanding through writing, though this was again moderated by planning. For present purposes, their key finding was that these differences were related to the amount of revision that the writers carried out during writing. Thus, for writers with high transactional beliefs, amount of revision was correlated with the extent to which the writers experienced a development in their understanding but unrelated to the quality of the text. By contrast, for writers with low transactional beliefs, amount of revision was unrelated to the development of understanding, and negatively correlated with text quality. Baaijen et al. (2014) concluded that writers with high and low transactional beliefs differed in the goals towards which revision was directed. For writers with high transactional beliefs, revision was directed towards clarifying their understanding; for writers with low transactional beliefs, revision was aimed at repairing problems in implementing the writer's plan. Revision was *proactive* for the high transactional writers but *reactive* for the low transactional writers (Galbraith & Torrance, 2004).

Although the keystroke measure used in this analysis provided additional information about process over and above the measures of quality and the development of understanding, this was of relatively limited usefulness. This is partly because the measure of text modification that was used as an indicator of amount of revision is a very global measure. It consists of the ratio of words in the keystroke log to words in the final product, and hence reflects a wide range of possible processes. But, in addition, these do not align theoretically with any particular component of the writing process. Currently, we are analysing the data in terms of specific measures of revision. Preliminary results indicate that there are indeed differences between high and low transactional writers in the amount of meaning-based revision that is carried out. Methodologically speaking, this greater alignment with cognitive models of writing promises to provide greater understanding of these differences.

The second example comes from a recently completed analysis using the two measures described above (see Table 13.1) (Baaijen & Galbraith, 2018). This showed systematic relationships of the sentence production and global linearity measures with text quality and a subjective measure of the development of understanding.

For text quality, there were systematic relationships with the two measures when writing was not outline planned. Higher quality text was associated with a combination of more controlled sentence production (high scores on the sentence production variable) and greater revision of global structure (low scores on the global linearity measure). By contrast, for outline planned writing, neither measure of processes during writing was related to text quality. This is consistent with the assumption that outline planning enables writers to generate and organise prior to writing and hence obviates the need to plan sentence production and re-organise the text during writing. The writer can use the pre-developed outline to organise the text and to enable the spontaneous retrieval of sentence content during writing.

There were also systematic relationships with the development of understanding. However, these were different to those with text quality. Thus, although there was a similar relationship between greater revision of global structure and greater development of understanding, the relationship with sentence production was in the opposite direction. Development of understanding was at a maximum when writing was non-outline planned and sentence production was most spontaneous (low scores on the sentence production dimension). It was reduced to zero when writing was outline planned and sentence production was highly controlled. These results demonstrate that the two dimensions have external validity: they relate to the quality of the final product and to the experience of the writer.

9 Conclusion

We set out in this chapter to address the problem of interpreting keystroke measures in terms of cognitive models of writing. In isolation, it is difficult to relate a particular pause or a particular burst of language production to any specific process. A pause can reflect a multiplicity of processes, from higher level reflection about the global structure of the text, through planning of an individual sentence, to struggling with the spelling of a word. A burst can be produced while making notes during explicit planning, as part of the ongoing development of an argument, or while revising a previously written passage of text.

The strategy we have described involves two complementary sets of procedures. First, the keystroke log is divided up into activities corresponding to the components of global models of writing. In this case, we distinguished between explicit global planning activities, the drafting of text itself, and post-draft revision, and then further distinguished between different types of pauses and bursts using a combination of bottom-up cluster analysis (mixture modelling in this case) and top-down identification of text boundaries. This, in effect, sorts the miscellany of measures available in a keystroke log into sets of measures better aligned with underlying cognitive processes. Second, these sets of measures are then aggregated across the text as a whole. The fundamental assumption here is that, while any individual instance of a measure may be multiply determined—a between sentence pause, for example, may reflect re-reading of previous text and/or planning of the next unit of content, when aggregated across the text as a whole and combined with other measures, the resulting measures will capture general properties of how a specific activity is carried out. In this case, we have identified two general dimensions of the writing process—sentence production and global linearity—and suggested that they have systematic relationships with both the quality of the final product and the subjective experience of the writer.

References

- Alamargot, D., Dansac, C., Chesnet, D., & Fayol, M. (2007). Parallel processing before and after pauses: a combined analysis of graphomotor and eye movements during procedural text production. In M. Torrance, L. van Waes, & D. Galbraith (Eds.), *Writing and cognition: Research and applications* (pp. 13–29). Amsterdam, the Netherlands: Elsevier.
- Allum, P.H., & Wheeldon, L.R. (2007). Planning scope in spoken sentence production:

- The role of grammatical units. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 33, 791–810. <https://doi.org/10.1037/0278-7393.33.4.791>
- Alves, R.A., Castro, S.L., Sousa, L., & Strömqvist, S. (2007). Influence of typing skill on pause-execution cycles in written composition. In M. Torrance, L. van Waes, & D. Galbraith *Writing and cognition: Research and applications* (pp. 55–65). Amsterdam, the Netherlands: Elsevier.
- Alves, R., Leal, J.P., & Limpo, T. (2019). Using HandSpy to study writing in real time: A comparison between low- and high-quality texts in grade 2. In E. Lindgren & K.P.H. Sullivan (2019), *Observing writing: Insights from keystroke logging and handwriting* (pp. 50–70). Leiden, the Netherlands: Brill.
- Baaijen, V.M., & Galbraith, D. (2018). Discovery through writing: Relationships with writing processes and text quality. *Cognition and Instruction*, Advance Online Publication. <https://doi.org/10.1080/07370008.2018.1456431>
- Baaijen, V.M., Galbraith, D., & de Glopper, K. (2012). Keystroke analysis: Reflections on procedures and measures. *Written Communication*, 29(3), 246–277. <https://doi.org/10.1177/0741088312451108>
- Baaijen, V.M., Galbraith, D., & de Glopper, K. (2014). Effects of writing beliefs and planning on writing performance. *Learning and Instruction*, 33, 81–91. <https://doi.org/10.1016/j.learninstruc.2014.04.001>
- Breuer, E.O. (2019). Fluency in L1 and FL writing: An analysis of planning, essay writing and final revision. In E. Lindgren & K.P.H. Sullivan (Eds.), *Observing writing: Insights from keystroke logging and handwriting*. (pp. 190–211). Leiden, the Netherlands: Brill.
- Chenoweth, N.A., & Hayes, J.R. (2001). Fluency in writing: Generating text in L1 and L2. *Written Communication*, 18(1), 80–98. <https://doi.org/10.1177/0741088301018001004>
- Chenoweth, N.A., & Hayes, J.R. (2003). The inner voice in writing. *Written Communication*, 20(1), 99–118. <https://doi.org/10.1177/0741088303253572>
- Chenu, F., Pellegrino, F., Jisa, H., & Fayol, M. (2014). Interword and intraword pause threshold in writing. *Frontiers in Psychology*, 5:182. <https://doi.org/10.3389/fpsyg.2014.00182>
- Flower, L., & Hayes, J.R. (1980). The cognition of discovery: defining a rhetorical problem. *College Composition and Communication*, 31, 21–32. <https://doi.org/10.2307/356630>
- Galbraith, D., & Torrance, M. (2004). Revision in the context of different drafting strategies. In L. Allal, L. Chanquoy, & P. Lamy (Eds.), *Revision: Cognitive and instructional processes* (pp. 63–86). Dordrecht, the Netherlands: Kluwer Academic Publishers.
- Hayes, J.R. (1996). A new framework for understanding cognition and affect in writing. In C.M. Levy & S. Ransdell (Eds.), *The science of writing* (pp. 1–27). Mahwah, NJ: Lawrence Erlbaum Associates.
- Hayes, J.R. (2009). From idea to text. In R. Beard, D. Myhill, M. Nystrand, & J. Riley (Eds.),

- The SAGE handbook of writing development* (pp. 65–79). London, UK: SAGE Publication Ltd.
- Goldman-Eisler, F. (1961). The distribution of pause durations in speech. *Language and Speech*, 4, 232–237. <https://doi.org/10.1177/002383096100400405>
- Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. New York, NY: Academic Press.
- Kaufer, D., Hayes, J.R., & Flower, L.S. (1986). Composing written sentences. *Research in the Teaching of English*, 20(2), 121–140.
- Leijten, M., & Van Waes, L. (2006). Inputlog: New perspectives on the logging of online writing. In K.P.H. Sullivan & E. Lindgren (Eds.), *Computer keystroke logging and writing: Methods and applications* (pp. 73–93). Amsterdam, the Netherlands: Elsevier.
- Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication* 30(3), 358–392. <https://doi.org/10.1177/0741088313491692>
- Leijten, M., Van Horenbeeck, E., & Van Waes, L. (2019). Analysing keystroke logging data from a linguistic perspective. In E. Lindgren & K.P.H. Sullivan (Eds.), *Observing writing: Insights from keystroke logging and handwriting* (pp. 71–95). Leiden, the Netherlands: Brill.
- Levelt, W.J.M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
- Levelt, W.J.M. (1999). Producing spoken language: A blueprint of the speaker. In C. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 83–122). New York, NY: Oxford University Press.
- Lindgren, E., Spelman Miller, K., & Sullivan, K.P.H. (2008). Development of fluency and revision in L1 and L2 writing in Swedish high school years eight and nine. *International Journal of Applied Linguistics*, 156, 133–151. <https://doi.org/10.2143/ITL.156.0.2034428>
- Lindgren, E., Westum, A., Outakoski, H., & Sullivan, K.P.H. (2019). Revising at the leading edge: Shaping ideas or clearing up noise. In E. Lindgren & K.P.H. Sullivan (Eds.), *Observing writing: Insights from keystroke logging and handwriting*. (pp. 345–364). Leiden, the Netherlands: Brill.
- Linnemann, M. (2019). Anticipation of audience during writing. In E. Lindgren & K.P.H. Sullivan (Eds.), *Observing writing: Insights from keystroke logging and handwriting*. (pp. 326–345). Leiden, the Netherlands: Brill.
- McLachlan, G.J., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley.
- Nottbusch, G., & Torrance, M. (2011). Written production of single words and simple sentences. In V.W. Berninger (Ed.), *Past, present, and future contributions of cognitive writing research to cognitive psychology*. New York, NY: Psychology Press.
- Schilperoord, J. (1996). *It's about time. Temporal aspects of cognitive processes in text production*. Amsterdam, the Netherlands: Rodopi.

- Sullivan, K.P.H., & Lindgren, E. (2006). *Computer keystroke logging and writing: methods and applications*. Amsterdam, the Netherlands: Elsevier.
- Tabachnick, B.G., & Fidell, L.S. (2001). *Using multivariate statistics*. Boston, MA: Allyn and Bacon.
- Vigliocco, G., & Hartsuiker, R.J. (2002). The interplay of meaning, sound, and syntax in sentence production. *Psychological Bulletin* 128 (3), 442–477. <https://doi.org/10.1037/0033-2909.128.3.442>
- Wengelin, Å., Frid, J., Johansson, R., & Johansson, V. (2019). Combining keystroke logging with other methods: Towards an experimental environment for writing process research. In E. Lindgren & K.P.H. Sullivan (Eds.), *Observing writing: Insights from keystroke logging and handwriting* (pp. 30–50). Leiden, the Netherlands: Brill.
- White, M.J., & Bruning, R. (2005). Implicit writing beliefs and their relation to writing quality. *Contemporary Educational Psychology*, 30, 166–189. <https://doi.org/10.1016/j.cedpsych.2004.07.002>