Developing pause thresholds for keystroke logging analysis

Simon Rosenqvist
DEVELOPING PAUSE THRESHOLDS FOR KEYSTROKE LOGGING ANALYSIS.

Simon Rosenqvist

Research on the process of writing uses bursts and pauses as key artifacts of underlying cognitive processes. However, the definition of a pause in writing is primarily based on tradition and ease of comparison between studies. This study explores keystroke logging data collected from middle school pupils (N=46) in northern Sweden, Norway and Finland and questions the traditionally defined pause’s usefulness, comparability and validity for investigating the underlying cognitive processes during writing. By examining the raw computer keystroke log data it was revealed that the group had a large variance in typing speed between participants and that different textual contexts had big variances compared to each other. Through exploration of different pause definitions’ effects on the text it was concluded that the twice the median length of pause (median x 2) was a good measurement for investigating pauses in sentences. Further, the 1.5 times the median (median x 1.5) for pauses between keystrokes within words proved useful for investigating the production of individual words.

Forskning på skrivprocesser har länge använt kaskader och pauser som nyckelartefakter av underliggande kognitiva processer. Definitionen av en paus i skrivande är dock främst baserade på tradition och direkt jämförbarhet mellan studier. Denna studie utforskar loggade tangenttrycknings data insamlade från elever (N=46) i mellanstadiet i skolor i norra Sverige, Norge och Finland och ifrågasätter den traditionellt definerade pausens användbarhet, jämförbarhet och validitet för att utforska underliggande kognitiva processer under skrivande. Genom att granska rå-data visade det sig att gruppen som helhet hade stora skillnader i skrivhastighet mellan deltagarna och att olika textuella kontext hade stora skillnader jämfört med varandra. Genom en undersökning av olika paus definitioners effekter på text kom det fram att dubbla längden på medianen för pauser (medianen x 2) var ett bra mått för att undersöka pauser i meningar. Dessutom var 1.5 gånger medianen (medianen x 1.5) i inom-ords kontexter ett användbart mått för att studera produktionen av individuella ord.

The purpose of this study is to develop a method for defining an individualized lower-threshold for what constitutes the pause during writing when data is collected using computer keystroke logging technology. Current computer keystroke logging research (e.g. Alves, Castro, de Sousa, & Stromqvist, 2007; Wengelin, 2006; Wengelin et al., 2009) applies an arbitrary two-second lower threshold when analyzing pauses in writing. This study explores writing data with the aim of developing a nuanced and individualized threshold that will provide the writer researcher access to useful traces of cognitive processes during a writing session that are lost when the two-second lower threshold is applied.

Writing is ubiquitous, and researchers investigate the processes of writing, with increasingly more complex methods. Apart from investigating different writing styles and strategies of writing some researchers have looked at the underlying processes in writing (for example, Lindgren & Sullivan, 2002; Wenglin, 2006; Stevenson, Sullivan, & Lindgren, 2008). There researchers, have, for example, considered how planning and transcribing of text works, and how does it relate to other cognitive processes such as attention and working memory? To answer these questions methods to record and analyze writing were developed. One of these methods is computer keystroke logging. This study focuses on one of the most
common assumptions in keystroke logging methodologies, namely a common lower threshold for pauses for all writers, and this study asks what we might be missing by keeps a universal lower threshold for all writers.

As the keyboard skills of early 21st century youth are higher than those of 20 years ago, inter-keystroke intervals tend to be highly automated and faster than for earlier generations. This may allow actions that slightly increase working memory load to be seen in text production, for example at morphological boundaries or complex bigram production.

Such increases in working memory load may be more frequent in second language writing as the writer may need to consider inflectional and derivation morphological processes, irregular grammar rules, and unusual spelling. The improved typing skills among the under 20s may mean that these traces are less hidden by non-automated typing skills.

This paper begins by giving an introduction into speech production and how it differs from reading and producing text. Then, before discussing definitions of the pause and the burst, a frequently used model of writing, Hayes (2012), and computer keystroke logging are introduced. Next the data, methods and analyses use in this exploratory study and the findings are presented. Finally I discuss the implications of the findings, and suggest lines for future research.

**Background**

To understand the implications of pauses in writing it is necessary to have an understanding of how spoken language is produced and its similarities with writing.

**Speech**

Compared to spoken language much less research has been done on written language, especially in psycholinguistics (Wengelin, 2002). There is a large difference between writing and speech. In speech an auditory signal formed by the representations in the sender’s mind causes a similar representation in the receiver’s mind (Garman, 1990). Understanding spoken language is a multimodal process, as demonstrated by the McGurk effect (McGurk & MacDonald, 1976), in text the signal is monomodal, based on vision.

The visual signal does not degrade like an auditive signal. Once the auditive signal is sent it disappears within a few tens of a second. The listener then has to keep the content of the signal in their short term memory. This temporal nature of spoken language makes the pragmatics of speech very different than that of text. A reader can keep track of a longer discourse than a listener can. Written language more diversity of words (Johanson, 2008) and longer clauses (Chafe & Tannen, 1987). While many would argue that spoken language is instinctual if not innate, written language is a learned, cultural skill. To become a competent writer requires a very special expertise that needs to be learned. Wengelin (2002) suggests that learning to write involves learning grammatical rules that cannot be found in spoken
language. A phoneme can be spelt in different ways and the same spelling can represent different phonemes (Garman, 1990).

**Early research in writing processes**

Early research on writing was primarily concerned with, among other things, the finished text, its composition, argumentative structure (Wengelin, 2002). Earlier, as Wengelin points out the writing researcher’s area of interest was mainly whole texts, often from acclaimed literary works or religious texts. In the second half of the 20th century researchers became interested in how texts were written. A common method for researching writing has been to ask the participants to verbalize their thoughts while writing, a think aloud protocol (TAP). Under a TAP the participants are instructed to talk and explain what they are doing without self-censorship while they are writing. While useful since you can hear the participants’ rationalizations for their writing behavior as it is performed, it forces the participants to shift their attention from writing to explaining their writing. To get around this problem a retrospective account can be used. Here the participant is recorded while writing and afterwards explains their own behavior while looking at the recording. The recording aids the recall of the participants. Neither a TAP nor a retrospective protocol uses pauses as a way to investigate the writing processes.

Off-line writing research looks at a text after the act of writing has been completed, whereas On-line writing research looks at the act of writing while it is being performed (Spelman Miller, 2006). On-line recordings of a writing session is sometimes combined with retrospective protocols or TAPs. To help the writer remember a recording of their writing process can be played back to them. The first recordings were made with a camera pointed at the page where the writer was writing by hand (Matsuhashi, 1980 in Spelman Miller, 2006). As the use of personal computers became more common researchers started using them as tools to investigate writing. Logging the keystrokes lets us see how the text is being constructed and makes it possible to study the process of writing rather than the finished product.
A cognitive model of writing

Theories about how the mind processes information in the working memory are central for explaining the process of writing; these emphasizes the complexity of the writing process. Writing draws not only from attentional resources and processing power of the working memory but also long term memory, motivations and motor skills (Berninger, Cartwright, Yates, Swanson & Abbott, 1994). Kellog (2008) argues that advanced writers have active representations of the current text, the goal text and the reader active at any given time. To allocate attentional resources for all these functions is crucial for creating advanced texts. Competent writers manage to automate much of what beginners have to deal with manually.

A widely used model of writing is the one proposed by Hayes and Flowers (1981). The model has been revised several times (Flowers & Hayes, 1981; Hayes, 2004; Hayes, 2012), and what started out as three main processes, planning, translating and revising, has now grown to three levels with a host of processes and resources. The main objective has been to emphasize writing as a complex task involving several cognitive, motivational and environmental aspects (Hayes, 2012).

The model in Hayes (2012) consists of 3 levels:
1. Control level with motivation for writing and representations of the goal.
2. Process level that consists of the Task Environment and Writing processes.
3. The resource level that contains cognitive resources that the other levels draws from such as attention, working memory, long-term memory and reading ability.

The control level contains the motivation for writing the text which influences the goal of the text, along with a representation of the current plan and a set of writing schemas.

The process level is divided into the task environment and the writing processes. The task environment is everything in the environment in which the text is being produced. This includes the writing technology (pen and paper, or keyboard), the text that has been written so far, the collaborators and critics that will read the text, and other material for writing, such as reference material. Studies have shown that the level of familiarity with the technology used for writing influences how the final text looks (Berninger, Cartwright, Yates, Swanson & Abbott, 1994; Connelly, Gee & Walsh, 2007; Christensen, 2004).

The writing processes take the plans for the text and translates them into language that can be transcribed. It consists of four interdependent processes: Translator, Evaluator, Proposer and Transcriber. It is important to note that a writer will move through the types of processes recursively (Lindgren, 2005): the proposer will propose a text based on the plans formed on the control level, the translator will translate it into language which the evaluator will monitor to see if it corresponds with the proposed plans. If it passes the evaluation the transcriber will transcribe it into the task environment.
The third level consists of the cognitive resources. These are finite in nature and the other levels of processes need to share the resources. When the resources are stretched thin the production of text has to be limited in speed or halted. This causes pauses or disfluencies which have been studied in the same way that speech disfluencies has been studied (Spelman Miller, 2006). She describes pauses as a way to give insight into the cognitive processes behind writing. Since the cognitive resources are finite, a pause could indicate that those resources has shifted from transcribing texts to elsewhere, such as to planning the next sentence or to review what has already been written.

Keystroke logging
Examples of keystroke loggers used for writing research include Translog and Inputlog. Translog was developed to research translation processes. It features two windows with the source text in one window (which is waiting to be translated) and the target text in another window (in which the translator writes and edits) (Jakobsen, 2006). Inputlog was designed to look like an ordinary writing session by letting the writer use MSWord as their word-processor (Leijten & Van Waes 2006). It logs all keystrokes and mouse movements and it is also equipped to record audio and the screen.

Revisions and Pauses
Revisions are changes to already written text. Revision-taxonomies make a distinction between changes to the plan of the text, and changes to already produced text. Lindgren (2005) called them internal and external revisions, and Lindgren and Sullivan (2006) pre-contextual and contextual revisions. An external or contextual revision is a revision to already produced text and occurs when already produced text has new text inserted to it, gets deleted or rearranged. The revisions can take place within previously written text or at the tail end of the text written so far. The contextual revision can be classified as a revision of the form or a conceptual revision. A form revision can be: correcting spelling mistakes, capitalizing or replacing a word with more appropriate one without changing the semantic meaning of the sentence. A conceptual revision changes the semantic content of a text.

An internal or pre-contextual revision revises the plans for a sentence. Pre-contextual revision is distinct from planning in that it is a revision of a proposed plan before the plan has been actualized. These revisions happen at the point of inscription, the tail-end of the text. Pre-contextual revisions can also be divided into revisions of the form or of the concept. A revision of the form is less demanding of the working memory than a revision of the concept (Lindgren & Sullivan, 2006) as long as you are fluent in the language. If the writer writes in a language she is less proficient in grammatical rules and spelling can be demanding. The pre-contextual revisions show up in keylogging data as moments of inactivity.

If a writer performs a pre-contextual revision the text produced afterwards needs to fit within the already produced text and whatever knowledge of the topic the writer has in their long term memory. A critique of the Hayes (2012) model of writing that Galbraith (1999) has made is that the model can only account for
knowledge that is already present in the writer’s long term memory. There can be no conceptual discoveries during the writing process. Galbraith (1999) criticizes those who see writing as basically a problem solving process, fitting together the right concepts that are stored in the long-term memory in the right way by using the translator process in Hayes (2012) model. Galbraith claims that writing is a process of discovery where one acquires new knowledge through activation in one’s semantic network. This semantic network does not consist of discrete ideas but rather sub-concepts from which the ideas are later formed according to Galbraith (1999). During a pause a writer might be wrestling with problems of form, concepts or trying to fit in the new knowledge they just discovered into the old text.

**Pauses and bursts**

With keystroke logging you get a dataset of keystrokes and their time of occurrence. A logger will usually compile this into a readable text and display different data from it. When writing a text words and sentence fragments are written in bursts with comparatively longer pauses between them. Studies have shown that the length of these bursts correlates with the length of the pause just prior to it (Spelman Miller, 2006). The pauses also correlate with clause and sentence boundaries (Spelman Miller, 2006; Wengelin 2002). In Hayes (2012) model of writing a pause is what happens when a writer needs to reallocate his attentional and working memory resources to formulate future writing plans or translate plans into language.

It is important to note that conclusions about the underlying cognitive processes cannot be inferred by pause durations in keystroke logger data (Spelman Miller, 2006). A moment of inactivity in the logs gives no clue as to what the writer was doing during that time, whether it was planning a sentence, searching through a mental lexicon or rereading text. While rereading can be controlled for with, for example with eye-tracking, keystroke logging remains an indirect measurement of the processes used. As such it is primarily interested in patterns of pauses and how they relate to the grammatical forms and working memory questions like how long bursts of texts can be produced.

Since a burst is production of text between two pauses the definition of a pause determines the definition of a burst. The definition of pauses in this context has in the majority of the literature been defined as 2 seconds of inactivity. The reasons for this are many but include that it was twice the mean typing rate and, the most common one, that it was easy to compare results between studies (Wengelin, 2006).

The justifications for the 2 second pause can be considered to be weak. According to Wengelin (2006) the mean median transition time between lower-case letters within words in the Swedish Spencer corpus varied between 181ms for university students and 568ms for fourth graders. In the English R&W corpus the differences were similar with the fastest writer had a median typing speed of 247ms and the slowest 488ms. Even within groups there can be a big variance in typing speed. The university students had a SD of 0.032ms while the 4th graders had a SD of 0.232. The doubling of these transition times are not close to 2 seconds and there is a difference between the corpuses and the groups within them. To investigate writing between these different groups may not be comparable.
By defining a pause to be 2 seconds long interesting phenomena may be missed. Sahel, Nottbusch, Grimm and Weingarten (2008) found that the time between keystrokes at the conjunction boundary in compound words were longer than mean/median in non-transparent words for German writers writing in German. Although they were trying to answer a different question than the majority of other research it suggests that the data already available might hold interesting answers.

**Examples of pause criteria**
The following is a text with a 2 second pause criterion from a participant in the study. According to this criterion the text consists of 5 bursts. The length and location of the pause is shown inside curly brackets and the time is measured in milliseconds. The “<“-sign within square brackets denotes a deletion and the number following it denotes how many deletions were made.

\{44943\}When we are \{2418\}with our firen[^4]riends do we a lot of things \{15740\}. But when we livees in a small city its not so myuck[^4]\{5429\}activi{1014}tes we can do \{2512\}on our freee timme[^3].

The same text with a 1 second pause criteria reveals more nuances. It shows that there was some hesitation when starting the second sentence. It also shows that it took some time to write the last syllable in “activites”.

\{44943\}When we are \{2418\}with our firen[^4]riends do we a lot of things\{1778\} \{15740\}. \{1139\}But \{1295\}when we livees in a small city its not so myuck[^4]\{5429\}activi{1014}tes we can do \{2512\}on our freee timme[^3]\{1077\}.

When further lowering the criteria, this time to twice the writer’s median typing rate (187ms x 2), we start to see difficulties in spelling and hesitation at clause boundaries. While spelling “friends” the writer has difficulties with the consonants that closes the syllable and also with the morphological infliction. At “...in a small city \{592\}its not \{702\}so...” we can see hesitation at one of the clause boundaries.

\{44943\}W\{546\}hen w\{593\}e are \{2418\}with our firen[^4]rie\{484\}nd\{468\}s \{452\}do we a\{468\} lo\{530\}t of things\{1778\} \{812\} \{15740\} \{577\} \{1139\}But \{1295\}when we live\{515\}d\{640\}es in a small city \{592\}its not \{702\}so my\{436\}uck \{562\}[^4]\{5429\}activ\{577\}i\{1014\}tes \{530\}w\{578\}e c\{484\}an do \{2512\}o\{452\}n o\{405\}ur \{578\} free\{405\} tim\{718\}e[^3]\{1077\}\{811\}

The pause definition brings out different aspects what happened during the production of the text.
Wengelin (2006) developed a taxonomy of pauses depending on the context in which the pauses are located. Her taxonomy distinguishes inactivity between, for example, letters within a word, between a letter and punctuation, between a letter and space followed by a letter. In her notation she uses a caret (^) to denote the location of the pause in relation to the surrounding context. The pause between a letter and another letter is shown as “a^a” and a pause between a space and a letter is shown as “a_\^a”. Her taxonomy describes what she calls micro-contexts. A summary of the notation is shown in Table 1. Wengelin also suggests that the median is a better measurement for the centrality of the values than the mean as the mean is much more receptive to influences of outliers.

Some researchers have recognized the validity problem of using the traditional 2 second pause. Baaijen, Galbraith and de Glopper (2012) tried to develop an alternative pause criteria. They wanted to investigate the length of bursts in texts written by 80 participants. They calculated the time between words, measured from the last letter in the first word to the first letter in the second word, from the raw interval data. Since the distribution of intervals was highly skewed towards the mean burst time, they plotted the intervals on a logarithmic scale. They found that a graphic visualization produced three separate peaks and the triple lognormal distribution was confirmed with a best fit model. They compared the triple-distribution model to all the participants in their data. They also compared the intervals for within-word keystrokes, between-sub-sentences and between sentences. What they found was that within-word intervals best fitted to a three lognormal distribution model (for 54% of participants). For between-word intervals a three lognormal distribution model had the best fit (on 58% of the participants). For sentences and sub-sentences a one lognormal distribution was the best fit.

Because of this they questioned if the median, suggested by Wengelin (2006) to

<table>
<thead>
<tr>
<th>Notation</th>
<th>Micro-context</th>
</tr>
</thead>
<tbody>
<tr>
<td>a^a</td>
<td>Inactivity between two letters</td>
</tr>
<tr>
<td>a^_a</td>
<td>Inactivity after a letter before a space followed by a letter</td>
</tr>
<tr>
<td>a_^a</td>
<td>Inactivity after a space preceded by a letter and followed by a letter</td>
</tr>
<tr>
<td>_^a</td>
<td>Inactivity after a space preceded by a major delimiter followed by a letter (as in a new sentence)</td>
</tr>
<tr>
<td>a^a</td>
<td>Inactivity after a letter followed by a major delimiter</td>
</tr>
<tr>
<td>a^_a</td>
<td>Inactivity after a letter followed by a minor delimiter</td>
</tr>
<tr>
<td>_^a</td>
<td>Inactivity after a minor delimiter followed by a space and a letter</td>
</tr>
<tr>
<td>D^D</td>
<td>Inactivity after a letter followed by a deletion</td>
</tr>
<tr>
<td>D^__D</td>
<td>Inactivity between two deletions</td>
</tr>
<tr>
<td>D^_a</td>
<td>Inactivity after a deletion followed by a letter</td>
</tr>
</tbody>
</table>

Figure 2. Example of how a file from Inputlog is structured
be a good measure of centrality, could be used as a base for pause criteria. They defined the lower threshold of a pause as above three standard deviations of the middle distribution of the trimodal lognormal distributions. They found that with this pause definition the length of the pause correlated negatively with the burst length, something that goes against previous research (Wengelin, 2006).

**Aims of the analysis**
The aim of this analysis is to investigate the validity of the 2 second pause definition in relation to alternative definitions and the context they are in. The study explores what phenomena can be seen at the word and sub-sentence level, and discuss them in the context of underlying cognitive processes. In what follows visualization and statistical tools have been used to explore the data from 46 writers.

**Method**
The data discussed in this study was collected as part of the project Literacy in Sápmi (Outakoski, 2015). The data consists of keystroke logs from pupils between 9 and 15 years old who live in northern Sweden, Norway and Finland. The project concern speakers of North Sámi and since there was a likelihood of a strong correlation between speakers of North Sámi and minority groups in Sweden, Norway and Finland the project sought and was approved permission by the Ethical Review Board in Umeå (2012-119-318). Because of the low number of students in some of the classes no gender was encoded in the data to help preserve the participants’ anonymity. The participants were instructed to write two texts, one argumentative and one descriptive, in English, north Sámi and the national language. The texts were written on laptops in the students’ classrooms using Microsoft Windows 7 as their operating system and Microsoft Word as their word processor. Inputlog was used for keylogging and Camstudio was used for screen capture while they wrote. The screen capture recording has not been included in this study. Only the English texts are included in this study. English is a second or third language for all the subjects.

Of the 126 pupils who participated in Outakoski’s (2015) study, 46 have been included in this study that looks at pauses in fluent writing in English. Of the texts that were excluded, some were too short other consisted of nothing but lists of the few words the writer knew in English. Texts that have some words or shorter sentence fragments in the writer’s language have been included.

**Data preparation**
Although the logs from Inputlog can be presented with two second pauses displayed, this exploratory study required the
raw-format of the files. The logs are encoded into an idfx-format. Within these files
the keystrokes are structured in an xml-format. Each keystroke event was extracted
and formatted with a purpose built analytic script. The events contain information
such as the value of the keystroke (in both unicode and character), the start time of
the stroke and the endtime of the stroke. An example of how the file is structured is
shown in Figure 2. Before the files could be used a script swapped all occurrences of
"&#x8;" and "&#xc;" to "BACKSPACE" and "CTRL" respectively since those codes are
not allowed in the xml-format. When reading the file the script extracted the
relevant information for further calculations. Each log was stored in a list of which
an example can be seen in Figure 3. To calculate the time of inactivity between the
keystrokes the start time of a keystroke until the start time for the next keystroke
were used. Events concerning mouse movements and actions were disregarded
since the study was concerned with production of text, not necessarily revision
types.

The data displayed some artifacts from the logger such as times tend to be clustered
in 15ms steps with a couple of milliseconds variation. Although the cause of these
clusters have not been confirmed it is likely to have to do with the speed of the
processor and how keystrokes were recorded in Inputlog. Madison and Wallace
(2012) found differences between operative systems when they tested how
responsive the keyboard and sleep function were, two common ways to measure
time-to-input in psychological tests. The clusters should not be seen as a feature of
the writing process but rather as an artifact from the logging, since this data deals
with comparatively long times (see Table 2) these artifacts should not present a
problem to the validity of this exploratory study.

The script was written in python 2.7 with the xml-library included to read the xml
file and the scipy-library (ver. 0.15.0) to create the graphs and perform statistical
analyses (Jones, Oliphant, Peterson, et al. 2001-).

Results
Differences between writers and micro-contexts
The sample mean and median typing rate for all participants showed that there was a big
variation between the writers. The 46 participants had a mean
typing rate of 1155ms with a standard deviation of
758.88ms. As discussed in
Wengelin (2006), the median
is a better measurement of
centrality and the mean
median for the participants in
the sample was 304.78ms with
a standard deviation of

<table>
<thead>
<tr>
<th>Micro-context</th>
<th>Average</th>
<th>Median</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>^a</em></td>
<td>10434</td>
<td>5716</td>
<td>12175.88</td>
</tr>
<tr>
<td><em>^a</em></td>
<td>1865</td>
<td>889</td>
<td>2432.13</td>
</tr>
<tr>
<td>^=a</td>
<td>5642</td>
<td>2409</td>
<td>8749.80</td>
</tr>
<tr>
<td>^=a</td>
<td>629</td>
<td>276</td>
<td>1016.24</td>
</tr>
<tr>
<td>^=a</td>
<td>1631</td>
<td>1532</td>
<td>1340.70</td>
</tr>
<tr>
<td>^=a</td>
<td>2623</td>
<td>1556.5</td>
<td>3540.99</td>
</tr>
<tr>
<td>^=a</td>
<td>631</td>
<td>342</td>
<td>999.55</td>
</tr>
<tr>
<td>^=a</td>
<td>1601</td>
<td>1437</td>
<td>1176.31</td>
</tr>
<tr>
<td>^=a</td>
<td>747</td>
<td>424</td>
<td>705.17</td>
</tr>
<tr>
<td>^=a</td>
<td>1532</td>
<td>1277.5</td>
<td>1066.08</td>
</tr>
<tr>
<td>^=a</td>
<td>345</td>
<td>256.5</td>
<td>302.30</td>
</tr>
<tr>
<td>^=a</td>
<td>2135</td>
<td>1709.5</td>
<td>1729.64</td>
</tr>
</tbody>
</table>
207.66ms. Such a large standard deviation shows that applying a single pause criterion for every participant hides differences in writing technique and style between the participants.

Table 2 shows the mean, mean median and standard deviation for all observed micro-contexts in the sample. There is a large difference in the values between different micro-contexts. Basing a single pause criterion for all micro-contexts would fail to take into account the differences between contexts. Based on these observations a pause criterion should take both the individual writer and the micro-context into account.

All intervals between keystrokes for a participant are plotted in chronological order in Figure 4. The lower horizontal line indicates the median inactivity time, the middle horizontal line indicates the mean inactivity time and the upper horizontal line indicates 2 seconds on the logarithmic y-scale. The graph shows that there is a lot of variation in the interval time and that only a fraction reaches above 2 seconds. The keystrokes are clustered around the median time rather than the mean.

In Figure 5 only the between-word intervals defined in Baaijen et al. (2012), as in a^_a+a^a in Wengelin (2006) are plotted. The median for all the participants’
inactivities, not just the between-word intervals, is at the bottom of the plot, unseen. The lower line shows the mean duration for all the participants' inactivities. The upper line shows the traditional 2 second pause. Figure 5 shows the variance within the micro-context which can also be seen in Table 2.

When focusing on the most frequent micro-contexts: within word (a^a) and between word (a^a + a^a) intervals the same definition as Baaijen et al. (2012) of between-word pauses is used; as they claim, it has better correspondence with spoken language. The between-word context is relevant since in much of the literature the assumption has been that higher level planning (sentences, concepts and similar) should take place at the same point as it does in speech: at sentence and clause boundaries. While sentences are marked with a period ("."), and are common in the available data sub-sentences and clauses are harder to find. Commas are uncommon in the data and therefore it is be better to treat all spaces that occur between words as a potential clause-boundary.

The a^a context, the within-word context, is the most common of contexts and is investigated since it is the level where artifacts from spelling processes can be seen. Since these are texts that are written in the participants' second or third language spelling patterns should be less automated than in their first language and the correct morphological inflection could be more difficult to access. Much of the research on writing is concerned about the higher levels of planning. The planning
that has to do with argument structure, sentence and sub-sentence planning. Effects of cognitive processes should be present at not just the sentence and clause level but also at the word level. Sorting and plotting the micro-context reveals a common feature for all participants: the plotted inactivities start with a steep climb in value to a plateau that is centered on the median and the last 10‒15% shows a steep climb again. The second curve (after the median plateau) corresponds roughly with double the value of the median (median x 2). For longer texts the mean and the median x 2 are roughly equal in value and coincides with the second curve. Since the mean is more influenced by outliers than the median the median x 2 would more often coincide with the second curve even for shorter texts. In figure 6 an example can be seen of the plateaus and curves. The lowest horizontal line is the median for this context and the second to lowest horizontal line is the mean for the context. The third lowest horizontal line is the median x 2 and the highest horizontal line is the traditional 2 second pause.

The same visual analysis was undertaken for the between-word micro-context. The between-word micro-contexts show a similar steep climb, plateau, steep climb pattern yet with a steeper slope of the curve. There is a larger variability in what can be seen as a normal between-word interval. In Figure 7 one participant’s between-word intervals have been plotted. The median for the between-word intervals corresponds with the lowest line which is the mean for all intervals. The second lowest horizontal line is the median for between-word intervals. The third
horizontal line from the bottom is the median x 2 and once again it corresponds well with the end of the plateau. The highest horizontal line is the traditional 2 second pause. From this analysis it is apparent that an abnormally long inactivity is likely to coincide with the micro-context’s median x 2.

The triple distribution that Baaijen et al., (2012) found for between-word pauses could not be found in the data. The shape of the curves varied wildly between different writers. Histograms for all participants can be found in the appendix.

Exploring within-word pauses
Using the script, a participant’s text was printed with pauses above the median x 2 within-words. Participant 331202 wrote with a median of 141ms in the within-word context. Pauses occur in compounds (“in{406}side”, “out{359}side”, “some{390}thing”). Pauses also occurred once in the morphological inflection (“rain{468}y”). From the third sentence onward (“{2231}When {437}i’ts...”) would be classified as a single burst in with the classical pause definition but we can see that it is riddled with disfluencies and short pauses.

{21716}W{639}hen {468}it’s ran{983}[<3]rain{468}y
{406}outside{811} an{593}[<2]{390}I sa{452}[<1]ta{811}y
in{406}side. {296}When {4103}it’s sunny out{359}side{6256}
{655}I{530} go {343}out. {2231}When {437}i’ts {327}[<4]t{764}’s
Participant 331201 had a median typing speed of 202ms in the within-word context. Pauses occur again in compounds (“im{624}portand”, “be{406}cause”, “other{514}wise”, “peace{530}ful”, “every{406}one”) and at morphological inflections (“farmer{921}’s”).

A lower pause criteria produces more pauses at syllable boundaries and where there is spelling that is unusual for Nordic writers. By lowering the pause criteria to the median*1.8 these pauses become visible. Participant 220717 has a median typing speed of 219ms in the within word context. With this criteria pauses occur frequently at syllable boundaries (“im{483}por{624}t”, “mo{577}bles”, “be{515}c{405}a{624}use”, “star{421}ts”, “ho{733}use”, “p{827}hones”) and morphological inflections (“ha{405}ve”, “star{421}t”, “you{655}r”)
doi{312}ng {656}with
{577}re{375}ind{281}ii{639}rs{951}{1467}so
much{1155}{1466}Reindii{530}[<1]eers{343}{281}need {1326}clean
air{1716}{422}and {359}good {6256}places {452}to
live{1014}{765}But {344}some peoples {562}dont
un{281}d{718}ersta{483}b{640}[<1]{437}n{546}{327}or {375}care
it{765}.

Participant 220712 has a median typing speed of 234ms in the within word context.
In order to see what we could find at an even lower threshold we printed the text with pauses over the median*1.6. Pauses occur at nearly every syllable at this rate.

{10764}the {1981}e{421}n{562}vi{2995}r{780}o{4056}nt{608}ment
{609}i{842}[<1]{9594}i{890}thin{1045}k{2652}{4758}we
{1358}s{577}take {1358}care {2075}o{733}f{1341}
{6724}[<1]{390}[<1]{468}[<1]{780}be{437}a{1528}[<1]{1217}ca{484}us
es{733}[<1]{780}[<1]{1342}i{577}ts
{437}po{624}rt{765}e{764}nt{1701},{2121}an{2028}d
{1295}be{655}a{405}use{6006}pla{1357}nts
{3276}[<1]{1186}{982}ani{453}m{858}als {1450}an{390}d
die{1202}{18392}a{562}nd
{1653}t{375}o{561}b{827}take {592}care {811}of{2995}t{671}he
{842}plan{1466}

One more analysis of texts were performed with pauses set to the median*1.4. The pauses that appeared then were similar to the pauses of the median*1.6. Participant 331202 types with median of 141ms.

{11949}W{281}e{483}an{219}v{203}e{343}s my{686}w{234}e{312}t{296}
ne{312}[<1]{842}[<1]{218}e{546}d{234}to{265}
{2324}ca{281}re{234}a{312}bout the
{250}n{1170}t{203}r{967}o{234}o{218}n{484}mnt.[...] {484}we
{343}try{203}{562}{296}{687}hel{234}pt{514}{3089}the
ani{234}m{234}als{234}b{390}e{297}fo{234}r{3806}a{297}bout the
{421}cli{250}ma{250}t.
A pause definition of 1.5 times the median for the within-word context creates pauses at syllable and morphological boundaries. Lowering the level further will result in a lower validity of the co-occurrences with these boundaries.

Exploring between-word pauses
Since the pause definition determines the length of the bursts a way to compare them can be to look at how the lengths of bursts change.

The bursts length between a 2 second pause is 4 words and a median x 2 is 3 words per burst on average and the difference is significant (p<.000) in an independent samples t-test. In Figures 8a and 8b the word-per-burst frequency has been plotted for the two pause criteria. With the median x 2 pause criteria there is a single normal distribution but with the 2 second pause criteria there are three normal distributions.

Participant 331203 had a median typing speed of 530ms in the between-word context. When applying the pause criteria of the median x 2 the text contains 178 pauses versus 87 when it had the 2 second pause definition. The new pauses show up at clause boundaries (“movie or {1186}play games”, “need {1326}to rem{2824}ee[<1]ber “). Difficulties with spelling can even be found on this level (“ch{1467}ao{1123}[<4]h[<1]{1295}annoying” = “chore”, “are {1388}alo{1528}ne”). in the example below the pauses are written in milliseconds. All pauses below 2000ms would not have shown up with the traditional pause criteria.

What you do when you are {1388}alo{1528}ne {1388}at home, {1435}you dont do too< much {41948}i usally watch a {5023}movie or {1186}play games{2028} {1592}and then \[<6]i m[<1]might invid[<1]te my friends over t my[<4] to my place {1732}t{1669}o do something{1311}. {6630}[<1], {2262}when its [<2]’s winter [<6]{1185} winter[<7]winter i{2106}[<1]it becomes a bit ch{1467}ao{1123}[<4]h[<1]{1295}annoying [1124] because its cold and yu[<1]ou ne[<6]i need {1326}to rem{2824}ee[<1]ber too[<1]
Participant 151205 has a median typing speed of 577ms in the between-word context. The number of bursts changes from 27 to 51 and the mean length change from 7 words to 4 words per burst. The pauses coincide with morphological inflections (“computer’s”) and clause boundaries (“who is abroad”, “Imagine a situation”).

Participant 211002 writes with a median speed of 422ms in the between-word context. With a median x 2 pause the participant has 191 pauses and with a 2 second pause the participant has 93 pauses. The mean words per burst is 4 for the median x 2 pause and 7 with the 2 second pause. The pauses tend to occur at clause boundaries (“my friends, we talk very much”, “an talk how i like to talk”)

When i am with my friends, we talk very much t[2]gether. {6162}We talk about almost everything there is to talk about. {3027}With my m[1] friends i can[2]an talk how {1342}i like to talk{5866} and then i dont have [6]t have to worry {2371}that my p[1451]arents{1731}{5}ents are going to say something a[1]bad about what i just sair.[<2]d. {11575}

If i am with my [4]<4967>h my friends at {1560}home, we use[8] and we are out sid[<4]side [1], we{1108} {889}usually just walk {2028}a little[<19] just take a little walk and talk about things, {1108}[2]{2652}, but when we are[<17] or {6880}[<4], bacy(905)kle[<2]x[<2]x[<1]ck[<1]le {3791}at t[<1]summer time or have snow ball wars at winter.{1840}[<9]when it's winter.
To investigate if the median x 2 pause may works, two writers with high between-word medians were investigated.

Participant 331204 has a between-word median typing speed of 1107ms. A median x 2 pause for this participant is therefore 2214ms, higher than the traditional 2 second pause. With the median x 2 pause the participant’s text contains 63 pauses with a mean burst length of 4 words. With the traditional 2 second pause the participant’s text contains 64 pauses with a mean burst length of 4 words. The writer’s median typing rate for within-word contexts is 297ms. Below is an example text with pauses over 1500ms printed in order to see where the different pauses occur. As can be seen the participant can even with a pause set 700ms below their median x 2 still produce quite long bursts. The pauses occur where we expect them to occur for the median x 2 pauses.

Discussion

This study explored pause definitions for investigating pauses at the word level and at the sentence level. The traditionally used 2 second pause definition is problematic since it does not take into account differences in writing ability. If the goal is to research underlying process of writing the methods used have to be sensitive to phenomena caused by these processes. The results show that to increase sensitivity, whether on word or sentence level, there is a need to take the individual participants writing ability into account. The median typing speed for the within-word context x 1.5 was a good pause definition for investigating spelling difficulties and other phenomena at the word level. The median x 2 for the between-word context raised the amount of pauses and lowered the mean burst length. The median x 2 pause changed the distributions of means for all participants from three distributions to one. This change could be an indicator that the median x 2 is a better measurement for pauses since with three distinct distributions it is likely that what is studied is writing style not the underlying structures.
What makes these pause definitions more valid than the traditional 2 second pause is that they take the writer's own typing speed into account. To be able to compare different writers there needs to be some sort of normalization for the data. Especially, since researchers call for studies with more participants (Baaijen et al., 2012) in order to be able to draw more valid conclusions.

The texts used in this study with within-word median x 1.5 pause definitions show that pauses crop up at syllable boundaries for writers who write in English as a foreign language. At the word level motor skills and how used you are to the keyboard become influencing factors (Christensen, 2004) and should be controlled for if what you want to look at is, for example, word retrieval and how phonemes match against graphemes.

The example of Participant 331204’s between-word context shows that the median, although more robust than the mean, is still a measure that is strongly influenced by outlier values. The participant seems to prefer to plan longer and then write a long burst. The pauses for between-word contexts might need to be normalized by the median for within-word contexts for example. This participant has little trouble finding the keys nor the words to express themself but the writing style, with longer pauses and rather well formed bursts that follows them, does not let the same structure be seen with the between-word pause definition as is seen among other writers.

Since this has been an exploratory study the conclusions are tentative. In order to determine the validity of these findings more research is needed. One way to move forward would be to use similar pause definitions to Baaijen et al. (2012). What was missing from their analysis were more robust findings: only 58% of the participants could be fitted into their between-word three distribution model.

The definitions proposed in this paper and the definitions used in Baaijen et al. (2012) could both be evaluated using a Signal Detection framework (Green & Swets, 1966) for the within-word context. A sample text could be used where the researcher marks, for example, syllable and morphological boundaries for a within-word condition. The different pause criteria would then be evaluated for how many hits, misses and false alarms they found. Since this context deals with quite fast times the hardware and software needs to be sensitive, and both the layout of the keyboard and the individual’s writing habit (Karnan, Akila, & Krishnaraj, 2011) needs to be accounted for. This type of evaluation works in the within-word context since it is studying automated behavior. It is not possible to use this method on between-word contexts since the pauses are not predictable. Although there are grammatical features that correlate with longer interval times (e.g. clause-boundaries, sentences and paragraphs) the characteristics of a pause is that it is unusually long for this context. A researcher could not identify where these pauses would occur and therefore we cannot make a hit, miss and false alarm table.
An ocular examination of the data suggests that pauses are more common and longer at sub-sentence and sentence boundaries such as at “,” or “.” characters. This could be systematically tested and if this suggestion were shown to be the case, this would agree with earlier research.

Since pauses are often seen right before a word, and the mean times for the a^a micro-context in this study confirms that the inactivity in that context is longer than mean, words are likely accessed serially. This means that there has to be a string generator that generates word series up to the length of sub-sentences.

According to the results in this study these strings tend to be around 4 words long. With the flat 2 second pause there were some writers who managed a mean of up to 10 words per pause. Even if they are very competent writers keeping many items active in working memory should be extremely hard. Especially if they also have to spare room for the other processes that occur such as the text goal and representations for the current text and imagined readers. The reason for the limited bursts length could be the capacity of the phonological loop.

The question is if these results are not simply as arbitrary as the 2 second pause. In the within-word context lowered the pause from median x 2 to median*1.8 because they provided the results that I was looking for, most of the time. What makes these pause definitions more valid than the traditional one is that for the between-word pause burst have a single distribution of words rather than three. For the within-word pause the validity is less strong and both the usefulness and correctness of that type of pause is more speculative.

**Conclusion**

From these results it is clear that not all pauses are created equal. Keystroke loggers for research purposes should adjust pause definitions for the individual participants and the micro-contexts that are present in the text. To get the best results our tools have to be as good as possible. As can be seen from the methods in this study analyses of the raw data can be done and can yield results.

One way to move forward could be to use Natural Language Processors (NLPs) to identify clause boundaries for example and then see what kind of interval times are present and work backward from that. The NLP would, at least in part have to be purpose built for this assignment since a text under production is very different from a finished text which most commercially available NLPs are designed to be used on.

The study needs to be replicated with another and perhaps larger group of writers. Further, it would be useful, but difficult, to develop an automated process to separate keystroke logs into sentences and clauses that is reliable.
References


Appendix A

This appendix contains between-word pauses for each participants in a histogram with a logarithmic x-scale.