Writing and rewriting:
Keystroke logging's colored numerical visualization

Hélène-Sarah Bécotte, Polytechnique Montréal, Canada & TERS, ITEM, ENS-CNRS, Paris, France,
Caporossi, Gilles, HEC Montréal, Canada & TERs, ITEM, ENS-CNRS, Paris, France
Hertz, Alain, Polytechnique Montréal, Canada
Leblay, Christophe, University of Turku, Finland & TERs, ITEM, ENS-CNRS, Paris, France

There are currently several systems for collecting online writing data via keystroke logging. Each of these systems provides reliable and very precise data. By applying a genetic criticism approach to the text, together with software for recording the written data, we seek to show how various visualizations are a very fecond way to understand re-writing.

The goal is to show that the time dimension (time pattern) is an essential dimension of writing, as well as the space dimension. This study of temporality is closely related to the differents methods of representation, specifically of colored numerical visualization. A new type of visualization based upon a transposition of mathematical graph theory is then described. These graph-related visualizations, also built from source documents that are the exhaustive recording (logs) of activities from the keyboard and cursor movements, clearly shows temporal (and local) phases in the writing and rewriting modes (revision activity).

Keywords: Classification, Keystroke logging, Numerical visualization, Textgenetic, Time-oriented production

1 Introduction

Filled with a lot of data, the keystroke logging files are difficult to read and to analyze (Wengelin, et al., 2009). This is due to many reasons which include their chronological format and the high amount of details (Kollberg, 1996). The writing process is complex and not linear. Many subprocesses can be embedded in other processes at any time (Flower & Hayes, 1981) and representations of it are so far one of the main analysis tools. The reason why analyzing the writing process is that important derives from the genetic methodology where the more the text is changed, modified, the better it becomes (Leblay, 2011, p. 16). The ultimate goal is then to understand how modifications are made in order that the text is always better.

We seek to position, in a complementary way, the genetic methodology among other disciplines of writing. The latter offers a real opportunity of understanding of digital writing (Leblay & Caporossi, 2015).

Two main points are crucial : the first is the refusal to give a value judgement on the written and rewritten text(s). Unlike the philological approach, genetic methodology is not looking for the better text, or the better recorded version on screen. We believe that philology is placed in a systematic perspective of degradation of textual quality ; thus, successive versions, away from the very first version (called *original*) suffer losses as time goes by. Instead of such a nostalgic approach, elapsed time, recorded time allows to highlight how writing is mostly rewriting, without quality losses. To simplify: *Best does not echo with older, but with newer*. The second point is that the closure imposed by the completed text,

depending on the structural approaches, does not take into account what happens before the so-called final text.
The first section (§ 2) presents what we mean by data representation *versus* data presentation and how the two are embedded in a larger concept called visualization. This major distinction allows us to undertake a classification of the main visualizations of the writing process, to better bring out the features of the graph representation (§ 3).

2 Data representations as a data analysis tool

The data and context complexity of the writing process and the keylog files need a specific contextual framefork and methodology. Already suggested by some authors to analyse the writing process, the use of data mining and big data can help processing and analyzing the data (Leijten & Van Waes, Keystroke Logging in Writing Research : Using Inputlog to Analyze and Visualize Writing Processes, 2013).

2.1 Big Data and the data science

In data analysis, the *data mining* techniques are included in big data (Manyika, et al., 2011). Big data is a set of techniques and tools used to explore and analyse computer databases (Karimi, 2014, p. vii). Statistics, A/B tests, data fusion and integration, regression, classification, association rules, visualization, text mining, sentiment analysis, neural networks, network analysis, graph theory, natural language processing, simulation and chronological series analysis (Manyika, et al., 2011) are examples of techniques and tools used.
However, there is not a strict definition of big data. For some authors, big data is the use of various data analysis techniques to process a huge amount of data (Manyika, et al., 2011, p. 27). Other authors have a broader definition of this discipline. For Lynch (2008), the data can be 'big' because the descriptive complexity of their analysis requires a specific experimental framework. Boyd and Crawford (2012) point out that "it is increasingly important to recognize the value of 'small data' [and that] research insights can be found at any level, including at very modest scale". In this case, big data is truly defined by its methodology then by the real quantity of data processed.
The simplified process by which the data is acquired, processed and analyzed is resumed by this series of actions: capture, sharing, sorting, analysis and visualization of the data (Manyika, et al., 2011). Some of these actions have already been studied individually and have been subject of software development aimed to record the writing process and to create representations of it in one of its dimensions (Caporossi & Leblay, 2011). Each of those software are different and are intended for a specific goal (Sullivan & Lindgren, 2014).
This chapter will address mainly the analysis through visualization and the importance of colors.
The goal of data representations is to help researchers with the analysis, to understand the data and find patterns over it. More than just drawings of data, visualization is an analysis tool (Manyika, et al., 2011, p. 33). By seeing how data interacts, it is possible to discover and to understand patterns and changes over time in a database (Yau, 2011, p. xvi) (Minelli, Chambers, & Dhiraj, 2013, p. 110). Allowing the researcher to using representations in a way

that does more then just describe a dataset requires visualization techniques. Those techniques are multi-disciplinary and include statistics, cognitive science, graphic design, computer science and cartography (Kirk, 2012), and also, as it was sait befor, textgenetic.
On the same visual surface, it is important to consider two complementary concepts when creating *data visualizations*: *data representation* and *data presentation*. Data representation is how the data is transformed by visual variables in order to create graph or charts. It's the physical form of the data and unfortunately it's often the only one considered (Kirk, 2012). Data presentation concerns the appearance and the delivery format of the entire data visualization design. It includes the choice of colour, the interactive features and the annotations (Aigner, Miksch, Schumann, & Tominski, 2011).

2. 2 Data representation

The choice of the representation format whether it's a line, bar, circle or a graph, should be in function of data as a raw material. The key points of this aspect is what are the variables used, how they are represented, what are their physical properties and what is the degree of precision. The goal is to create a representation that best portray the data's attributes (Kirk, 2012).
To choose the right method to represent the data, it's important to know the data's specifications. Detailing its characteristics is the first step of the representation. It includes setting the independent and dependant variables (Aigner, Miksch, Schumann, & Tominski, 2011, p. 4) . The independent variables consists of the *location* and *time*. The time variable has multiple characteristics. Its scale can be ordinal, discrete or continuous, the scope can be either point based or interval based, the arrangement is linear or cyclic and finally the viewpoint is either ordered, branching or with multiple perspectives (Aigner, Miksch, Schumann, & Tominski, 2011, pp. 71-72). The dependant variables includes what has been measured or observed (Aigner, Miksch, Schumann, & Tominski, 2011, p. 4).
The second aspect to consider is the specification of the task, the reason why the data are visualized. Three different goals can motivate data representation: explorative analysis, confirmative analysis or presentation of analysis results (Unwin, Chen, & Hardle, 2008).
Another important aspect is the degree of accuracy of the representation. It's how the variables are considered and illustrated, on a micro or on a macro level. Some variable can be as precise as the exact value on an axis position. Other variables can be shown less accurately with color ranges or shapes for example (Kirk, 2012). It can be the case of time when it's interpreted relatively rather than absolutely. The duration between two occurences of the same data item can be illustrated with transparency or color (Aigner, Miksch, Schumann, & Tominski, 2011, p. 78). Color therefore is a way to add another dimension of data to the visualization (Blanchard, 2005, p. 2).

2.3 Data presentation and design options

The presentation of data involves the overall design of the visualization tool. While it is more important when the goal of the visualization is to present some results, some features are worth considering in an explorative-orientated model (Unwin, Chen, & Hardle, 2008).

Data visualization relies traditionally on static representations. The most common problem is how to illustrate well the relationships within the data as well as the change over time (Bartram, 1997). Static representations have to be in two dimensions and incidentally can't show accurately data problems that have multiple dimensions (Minelli, Chambers, & Dhiraj, 2013, pp. 110-113). One of the best ways to visualize spatiotemporal changes is to animate the data (Yau, 2011, p. 309). However, another and more simple way to add dimension is the use of colors (Blanchard, 2005, p. 65). Having the right combination of visual codes and dimensions increase the analyst's efficiency and speed while looking at the representation because that kind of visual information doesn't require cognitive effort and is rather processed by the preattentive visual system (Bartram, 1997).

While colors can help illustrate variables, an other consideration when choosing those colors would be to make the visualization tool accessible to the greater number of researcher by choosing colors while paying attention to how color blind persons see (Aigner, Miksch, Schumann, & Tominski, 2011, p. 89).

2. 4 Finding patterns

There is no specific method that exists to guide researchers in finding patterns. Visualizations are meant to be an exploratory way of finding patterns (Manyika, et al., 2011, p. 18). They essentially are a tool that allow the human eye to observe underlying structures (Blanchard, 2005, p. 40).

The process of finding patterns starts by giving a visual version of the data to a researcher or an analyst that would let them understand it (Dzemyda, Kurasova, & Zilinskas, 2013, p. v). The « graphical displays should induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production, or something else » (Tufte, 2001). After understanding the data, the researcher can therefors gain insight into it and draw conclusion from the patterns discovered (Dzemyda, Kurasova, & Zilinskas, 2013, p. v). As humans are very effective at discovering certain types of patterns, their ability to process a lot of data is limited (Manyika, et al., 2011, p. 33). The quality and ergonomy of the visualization is extremly important to ensure the maximization of the possible outputs (Tory & Moller, 2004).

While finding patterns, it's important to be careful with the interpretation. Data analysts can become apopheniacs (Boyd & Crawford, 2012) which means they detect correlations between unrelated and random items (Merriam-Webster, 2014). Interpretation is at the center of data analysis and understanding the methodological processes allows reducing data limitation and bias (Boyd & Crawford, 2012). While the goal of creating visualizations of data is to integreate the human in the analyzing process, this step is often made alongside with algorithms or with the help of a software to reveal patterns or clusters (Dzemyda, Kurasova, & Zilinskas, 2013, p. v; Moller, Hamann, & Russell, 2009).

3 Visualizations of the writing process

Without pre-processing, the writing process data is big, complex and not appropriate for human analysis (Caporossi & Leblay, Outils de visualisation de données enregistrées, 2014). The writing process data is multi-dimensional. Those dimensions, which are temporality, chronology and spatiality, concern the genetic operations at the most basic level. Each operation of the writing process could be considered as a *substitution* operation (Van Waes & Schellens, Writing Profiles : The Effect of the Writing Mode on Pausing and Revision Patterns of Experienced Writers, 2003). The *insertion* would be the replacement of an empty space by a keystroke and the *deletion*, the *replacement* of a keystroke by an empty space. These operations are characterized by the fact they are made in one operation with the mouse or the keyboard. The more complex operations, such as the substitution and the replacement, which are made in two steps (Caporossi & Leblay, Online Writing Data Representation : A Graph Theory Approach, 2011), are considered as a combination of the simple operations.

Each basic operation owns its specific position for the combination of the dimension. The temporality is defined as the order of the operations in the text as they have been made in time. Therefore, two operations can't have the same temporality position. The chronology is the order of the operations in the text as it is read. As for the temporality, each operation has a distinct chronology position. Finally, the spatiality concerns where the operations have been made geographically in the text. More than one spatiality dimension exists. The position of an operation as it's recorded in the log file concerns where it has been made considering the text's state at that specific moment. The absolute spatiality is the operation's position compared to all the operations that have been made in the text.

Another type of data is the cursor's position. It has the same characteristics as the operations: at a specific time T, the cursor is at the recorded position R and the absolute position P.

In some cases, researchers also include variables that are excluded from the writing actions process itself (actions made by the hands) but contribute to the writing process. Amongst them, an eyetracking feature allows to know where the writer look while typing or to know at what they look exactly when they look back over the text they wrote (Wengelin, et al., 2009). While it concerns manual writing, the software Eye and Pen can record where the eye look and, at what time it does (Chesnet & Alamargot, 2011). The variables are even more complex.

To facilitate the analysis, researchers are using representations of the writing process to find patterns. The specification of the task of the data representation is therefore an *explorative analysis*. Writing process representations are predominantly static. The only exception is the animated replay of the process that allows to visualize chronologically the operations made by the writer that were recorded in the log file (ITEM, 2014). Even if it's animated, this representation is only a film of what happened, without any data preprocessing.

The actual visualizations of the writing process are bidimensional and because of that, they are oriented either on the revision, the temporal aspect or the writer's retrospection for example (Latif, 2008). Even if the spatiotemporal dimension of the process is an important aspect to analyze and understand (Stromqvist, Holmqvist, Johansson, Karlsson, & Wengelin, 2006), none of the actual visualizations represents completely the problem.

Another aspect of the writing process is the micro and macro aspects of the text i.e. the detailed operations made and the global structure of the process. Because those two aspects

can't be together visualized on the same representation unless using interactivty and the view adjustment feature (Aigner, Miksch, Schumann, & Tominski, 2011, p. 107), researchers usually use many representations to understand more completely the process (Alamargot, et al., 2011, Breetvelt, et al., 1994, Caporossi & Leblay, 2011, Cox, et al., 2009, Doquet-Lacoste, 2003, Haas, 1989, Latif, 2008, Leijten & Van Waes, 2013, New, 1999, Southavilay, et al., 2013, Van Waes & Schellens, 2003).
When dealing with the representation of the writing process, two dimensions necessarily arise : time and space. The *diplomatic* transcription deliberately emphasis on space, and could barely be considered as a representation of the writing process. On the opposite, the *linear* representation focuses on the temporal aspect and represents the various actions of the writer on the text. Between these two extremal representations, many variants may exists.

3. 1 Linear representation

The linear representation allows to see closely the operations made while writing the text. Many different linear representations exists but they have similar properties. The spatial dimension is explicit while the temporal dimension is present but not well detailed. While it is called a representation, it is not graphical.

3.1.1 S-notation

The **S-notation** is a linear representation that is used mainly in the *Inputlog*, *Trace-it* and *JEdit* softwares to indicate the text's evolution in demonstrating the usual text production, the insertions, revisions and deletions (Lindgren & Sullivan, The LS Graph : A Methodology for Visualizing Writing Revision, 2002). It has been created in order to better understand the revisions and "was inspired by an informal notation for revisions in handwriting developed by Matsuahi (1987) (Kollberg, Rules for the S-notation: a compuyter-based method for representing revisions, 1996). It focuses on the editing actions and makes it neutral by representing it as independent, unrelated operations (Kollberg, S-notation as a tool for analysing the episodic nature of revisions, 1996).

The following symbols have been used to illustrate the revision operations:

Table 1. S-notation's main symbols

$\|_i$	The interruption (break) with sequential number i
{inserted text}^{i}	An insertion occurring after interruption number i
[deleted text]^i	A deletion occurring after interruption number i

The example comes from Kollberg (S-notation as a tool for analysing the episodic nature of revisions, 1996). *I am writing a {short}[1] text. |[1] It will [probably][2] |[3] be revised [somewhat][3] later. |[2] Now [I am |[4]][4] it is finished.* The final version of that text would be: *I am writing a short text. It will be revised later. Now it is finished.*

This representation allows the researcher to detect where the interruptions and revisions are made during the writing session. However, due to the low level of information available in this representation, an interpretation of the writer's actions is needed to complete the

analysis (Kollberg, Rules for the S-notation: a compuyter-based method for representing revisions, 1996).

Here are the data representation and data presentation characteristics of the S-notation :

Data representation	
What are the variables used?	Chronology of the operations.
How are they represented?	It is simply the written text with symbols that indicate where in the text are made the revisions and in which order (Kollberg, 1996).
What is the degree of precision?	Precise, the text and its modifications are clearly readable.
Data presentation	
What are the characterstics of the data presentation?	The S-notation is integrated in an interactive program for revision analysis. It is possible for the user to navigate between different versions of the revisions, play the writing session and generate statistics of the session.

The main flaw of this way of representing the writing process is the low number of dimensions used.

There exists a more comprehensive version of the S-notation including references to sources used when the text was written. The copied text is underlined and different font types are used to illustrate the same sources (Perrin, 2003).

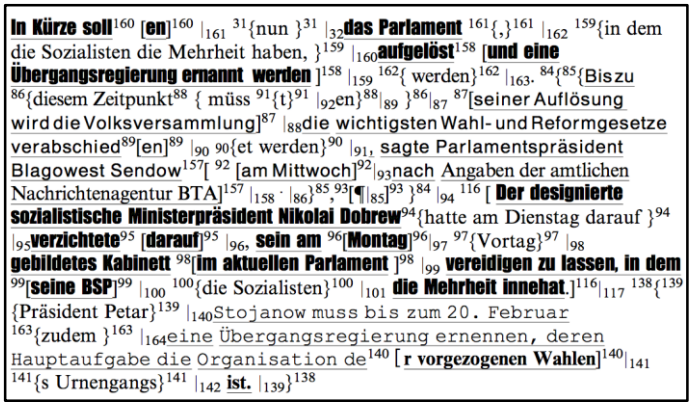

Figure 1. S-notation example

Here are the data representation and data presentation characteristics of this other version of S-notation :

Data representation	
What are the variables used?	Chronology of the operations.
How are they represented?	It is simply the written text with symbols that indicate where in the text are made the revisions and in which order. In the « sources » version, the use of different fonts help the user noticing the origin of the text displayed (Perrin, 2003).
What is the degree of precision?	Precise, the text and its modifications are clearly readable.
Data presentation	

What are the characterstics of the data presentation?	The different font used in the « sources » version are sufficiently different for the reader to see the difference but when the amount of sources use dis greater then 3-4 it became hard for the analyst.

There is also a third variation of the same visualization that include the use of the mouse, the touch pad or the arrows by the writer and is called the genetic linear representation (Leblay, 2009 ; Doquet & Leblay, 2014). The revisions symbols are not the same then those used in the S-notation.

The following symbols have been used to illustrate the genetic linear representation:

Table 2. Genetic linear representation's main symbols

Transcription conventions	Type 1 operations, following what has already been written	Type 2 operations, by returning in what has already been written
Insertion	Insertion 1 Simple	Insertion 2 (Simple +) **Bold**
Deletion	Deletion 1 Immediate deletion ~~*Crossed italics*~~ - Delayed deletion ~~Crossed without italics~~	Deletion 2 Immediate deletion ~~***Crossed bold italics***~~ - Delayed deletion ~~**Crossed bold without italics**~~
Substitution	Substitution 1 ~~Substituted~~substitute	Substitution 2 ~~**Substituted**~~**substitute**
Displacement	Displacement 1 ~~Displaced~~displacer	Displacement 2 ~~**Displaced**~~**displacer**
Text [i]	A revision occurring after interruption number i	
🖱 □ ←→ ↑↓	Movements of the mouse, the touchpad or the arrows	

TRANSCRIPTION LINEAIRE (.DOC)

Que~~ *me vionet*~~'est-ce qui me vient en tête spontanément quand on me donne la tâche d'~~*e*~~e décrire un environnement idéal?[18] Be**u**[1]au[2]coup de verdure←[1]→[2], peu de pollution, suffisamment d'espace pour que t~~*uot*~~out un chacun se trouve ~~`~~*l'i*à l'aise **a**[19]~~*sv*~~**a**[4]vec soi-m^~~*me*~~ême et**v**[3] avec les autres🖱[3]🖱[4]🖱. Le climat? Je dois dire que les saisons disctinctes et variées typiques du Nord m**ae**[5] plaisent🖱[5]🖱, etv que c*e*'est vquelque chose que ~~*nje*~~ je voudrais avoir dans un **ue**[10]nvironnement idéal. ~~*Do*~~Autrement **n**[11]dit, un été chaud et ensoileillé ~~ave~~esans oublier évidemment les journées lumineuses et e**x**[6]trêmement longues🖱[6]🖱, ~~`la~~comme en Finlande en été, un hiver blanc, avec du **ds**[12]oleil et de la n~~*aige*~~ei~~*d*~~ge, et encore d**ueux**[13] sais~~*io*~~ons de "transmisson", de passage on peut dire, pour passer du froid,[14] au chaud ou ~~*läinverse*~~inversemen~~*rt*~~t: le pri~~*np*~~ntemps et l'a~~*i*~~utomne.

Figure 2. Genetic Linear representation

Here are the data representation and data presentation characteristics of the genetic linear representation :

Data representation

What are the variables used?	Chronology of the operations.

How are they represented?	It is simply the written text with symbols that indicate where in the text are made the revisions and in which order. The use of the mouse, the touch pad or the arrows is also indicated by symbols (Leblay 2009). The revisions symbols are not the same then those used in the S-notation.
What is the degree of precision?	Precise, the text and its modifications are clearly readable.
Data presentation	
What are the characterstics of the data presentation?	The use of icons to represent the cursor's position is quite user-friendly. It is easy to read the final text if there's not a lot of revisions but it can become overwhelming if the same portion of the text is modified more then once.

3.1.2 Scriptlog software linear representation

The Scriptlog software has its own linear representation that is somehow a simplification of its log file and can "be used to follow the writing process step by step, as an alternative to watching the ScriptLog replay" (Wengelin, et al., 2009). This representation's goal is to highlight the pauses.
The following example was written in Swedish and come from Wengelin et al. (2009).

<5.762>Filmen jag såg nyss handla<BACKSPACE6>bistro<BACKSPACE2>od av korta klipp<BACKSPACE4>s<BACKSPACE2>sekvenser ur några ung<3.639>domars vardag<9.930> <4.027>.Nästan<3.825><BACKSPACE9>. I princip alla sekvenser ha<BACKSPACE2>visar<BACKSPACE>de hur ungdomar utsätts för jobbiga situationer u<BACKSPACE>avsiha<BACKSPACE2>n a<BACKSPACE2>a kamrater.

Figure 3. Scriplog linear representation

The final version would be:

Filmen jag såg nyss bestod av korta sekvenser ur några ungdomars vardag. I princip alla sekvenser visade hur ungdomar utsätts för jobbiga situationer av sina kamrater.

Here are the data representation and data presentation characteristics of the Scriptlog software linear representation :

Data representation	
What are the variables used?	Temporality.
How are they represented?	The numbers between the brackets < > indicate a pause and it's time length. <BACKSPACE> refers to the deletion of a keystroke while <BACKSPACE6> would mean that the writer deleted 6 keystrokes. This representation of the writing process doesn't allow to see well returns in the text (Wengelin, et al., 2009).
What is the degree of precision?	Precise, the keystrokes are clearly displayed.
Data presentation	
What are the characterstics of the data presentation?	The presentation is barely limited to some symbols used and the use of uppercase for certain aspects.

The main flaw of this representation is that it doesn't allow an intuitive reading (Wengelin, et al., 2009).

3.2 *TimeLine* representation

Timeline is another representation that aims to show what the writer is doing while writing a text. This representation uses Scriptlog and eyetracking data in order to give an overview of the allocation of the writer's attention (Wengelin, et al., 2009).

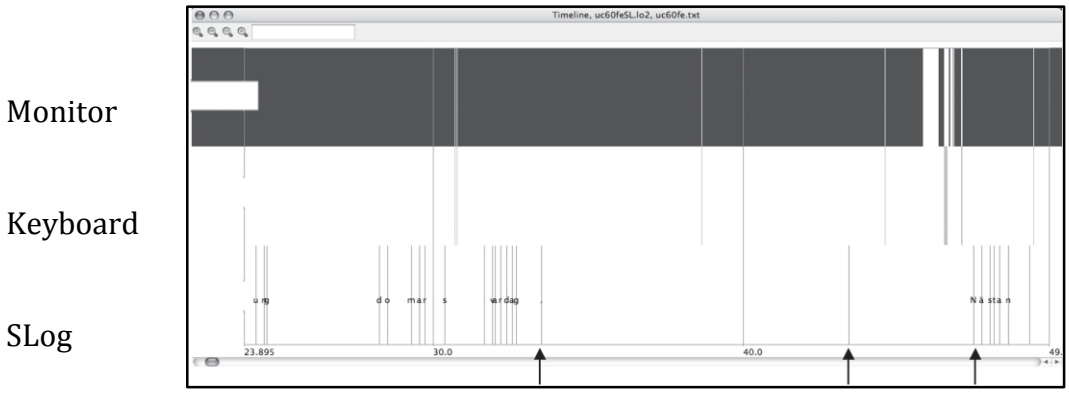

Figure 4. Timeline representation (Wengelin, et al., 2009)

This visualizing tool has been created to overcome the linear representation's lack of information about what the writer was doing during the pauses (Wengelin, et al., 2009).

Here are the data representation and data presentation characteristics of *Timeline* :

Data representation	
What are the variables used?	Temporality of each operation (keystroke).
How are they represented?	The x-axis shows the time. The y-axis is separated in 3 sections. The Monitor and Keyboard sections represent where the writer looks at. The dark sections indicate where the writer is looking at, at this precise moment. The SLog section is allowed to the operations (keystrokes). The operations are represented by a thin line and the exact keystroke is indicated on the line (Wengelin, et al., 2009).
What is the degree of precision?	Extremely precise. It is possible to view exactly the operation made but it is impossible to read or understand what has been written.
Data presentation	
What are the characterstics of the data presentation?	The presentation is quite minimalistic. The use of a dark hue to where the writer is looking at is efficient. The size and placement of the keystrokes (letters) however may not be optimal as it is difficult for the user to really see the information.

While it is impossible to read what key have been used in the Slog part, this tool is never used alone. It is a complement to the S-notation (Wengelin, et al., 2009).

3.3 Overview of the writing process visualizations

The overview visualizations are those that illustrate the global flow of the process. It is almost impossible to know what the text and the operations actually look like from that perspective (Caporossi & Leblay, Online Writing Data Representation : A Graph Theory Approach, 2011). Those visualizations are mainly inspired by Geographical Information Systems (GIS) because of the similarities between those fields such as the geographical and temporal identification of items (Lindgren & Sullivan, 2002). In their original field, GIS are a tool for both visualization and data mining (Lindgren E. , Sullivan, Lindgren, & Spelman Miller, 2007).

3.3.1 Progression diagram

The goal of this visualization is to emphasize the revisions made within the process (Lindgren & Sullivan, The LS Graph : A Methodology for Visualizing Writing Revision, 2002). The progression diagram has been created to be used along with the S-Notation for a better analysis of the writing process (Perrin, Progression Analysis (PA): Investigating Writing Strategies at the Workplace, 2003).

The following visualization corresponds to the « sources » S-notation presented in figure 5.

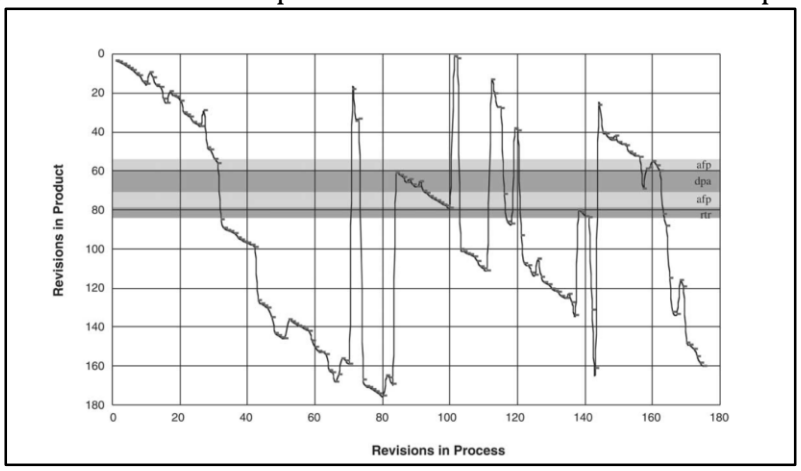

Figure 5. Progression diagram

Here are the data representation and data presentation characteristics of the Progression diagram :

Data representation

What are the variables used?	Temporality and spatiality. However, the variables are not the operations (keystrokes) but the revisions that occured in the process. The sources used are also included.
How are they represented?	On two axis, the number of revisions in process (x axis) and the number of revisions in product (y axis): the 'revisions in product' axis help the researcher knowing how much revisions has been made on a specific section of the text. For the sources, for example « AFP » refers to copy-paste from Agence Science Presse.
What is the degree of precision?	Not precise, it's an overview of the « revision » aspect of the process. It is not a representation of the whole process.

Data presentation

What are the characterstics of the data presentation?	The axis's names are well identified. The light and dark grey sections refer to a specific section of the text (spatiality) that has been selected on the computer and is seen on the screen as the S-notation.

The progression diagram's strength is to emphasize on the revision's order compared to the final text. In this representation, the temporal dimension is put forward (Latif, 2008).
The main limitation of the progression diagram is that it's "seriously limited in that it is based solely upon the relative position of the revision with no relation to time, text size, or distance of movement" (Lindgren & Sullivan, The LS Graph : A Methodology for Visualizing Writing Revision, 2002).

3.3.2 LS graph

This graph shows the same information as the *S-notation*, but using it differently from the progression diagram. This representation was largely inspired by the Perrin's work (2003) and the progression diagram (Leijten & Van Waes, Inputlog : New Perspectives on the Logging of On-Line Writing Processes, 2006).

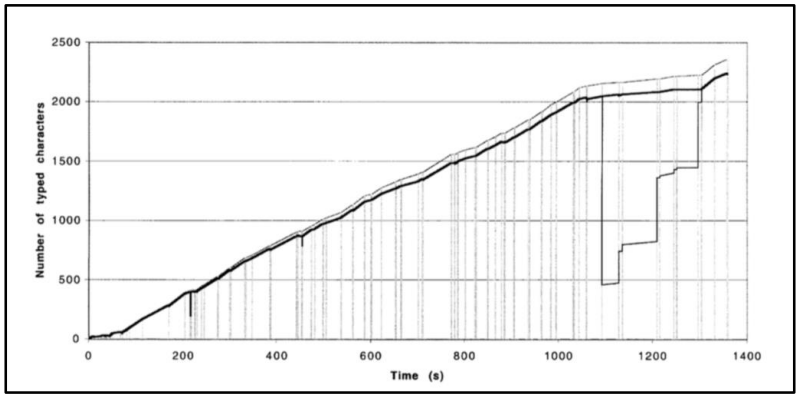

Figure 6. LS graph

Here are the data representation and data presentation characteristics of the LS graph :

Data representation	
What are the variables used?	Temporality and spatiality.
How are they represented?	The *x-axis* represents time while the vertical axis represents the total number of characters typed. The top line shows the progression of the total typed characters. The centered line (the darker) is the length of the text as a function of time. Finally, the lower line is the cursor position in the text (Lindgren & Sullivan, The LS Graph : A Methodology for Visualizing Writing Revision, 2002)
What is the degree of precision?	Not precise, it's an overview of the process.
Data presentation	
What are the characterstics of the data presentation?	The axis's names are well identified. It would be useful to identify clearly on the side which line color refers to what aspect of the text. This representation is purely static, the only way to know exactly what is happening in a certain point is to

manually search it in the data and examine this event in details (Lindgren E. , Sullivan, Lindgren, & Spelman Miller, 2007).

The main differences with the Progression diagram are the nature of the operations identified. The Progression diagram focuses on revisions while the LS Graph is more general and represents the number of operations made in time. The Progression diagram's y-axis's scale is from top to bottom, similarly to the way we read and write texts while the LS Graph's y-axis's scale is from bottom to top.

3.3.3 Representation based on a GIS Software

Considering all the strenghts of the GIS, this software visualization has also been used. Compared to the LS graph that is purely static and which it is impossible to automatically have insights on a precise graph point, the GIS allows the researcher to gain insight instantly (Lindgren E. , Sullivan, Lindgren, & Spelman Miller, 2007).

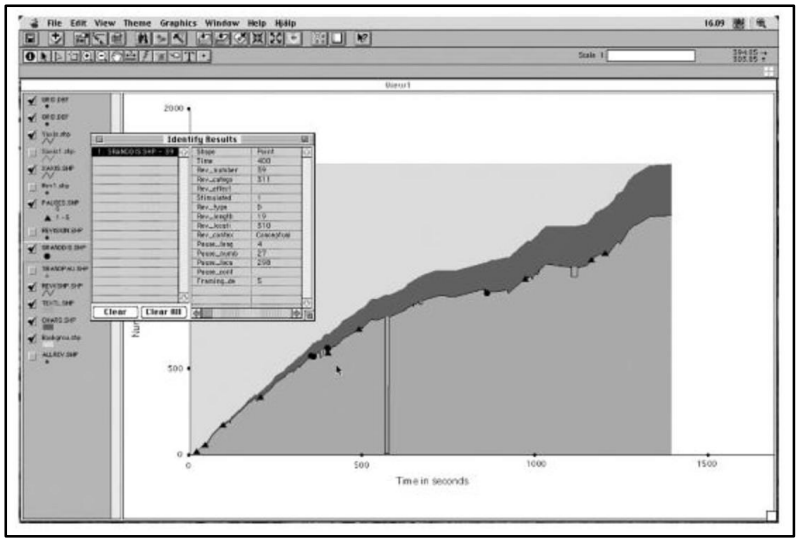

Figure 7. GIS sofware representation

Here are the data representation and data presentation characteristics of the representation based on a GIS software :

Data representation

What are the variables used?	Temporality and spatiality.
How are they represented?	The *x-axis* represents time while the vertical axis represents the total number of characters typed. The top line shows the progression of the total typed characters. The centered line (the darker) is the length of the text as a function of time. Finally, the lower line is the cursor position in the text (Lindgren E. , Sullivan, Lindgren, & Spelman Miller, 2007)**.**
What is the degree of precision?	Not precise, it's an overview of the process.

Data presentation

What are the characterstics of the data presentation?	The axis's names are well identified. This representation is dynamic in the sense that the secondary window display information corresponding to a specific coordinate (x,y) of data which helps the user to analyze the writing process (Lindgren E. , Sullivan, Lindgren, & Spelman Miller, 2007).

3.3.4 *AFP* representation

The *Genèse du texte* software provides this representation called *Au fil de la plume* (AFP), which is part of the same category as *GIS* and *LS Graph* (Caporossi & Leblay, 2011). The focus here is on the cursor position in the text and tracking back and forth in it (Doquet-Lacoste, 2003).

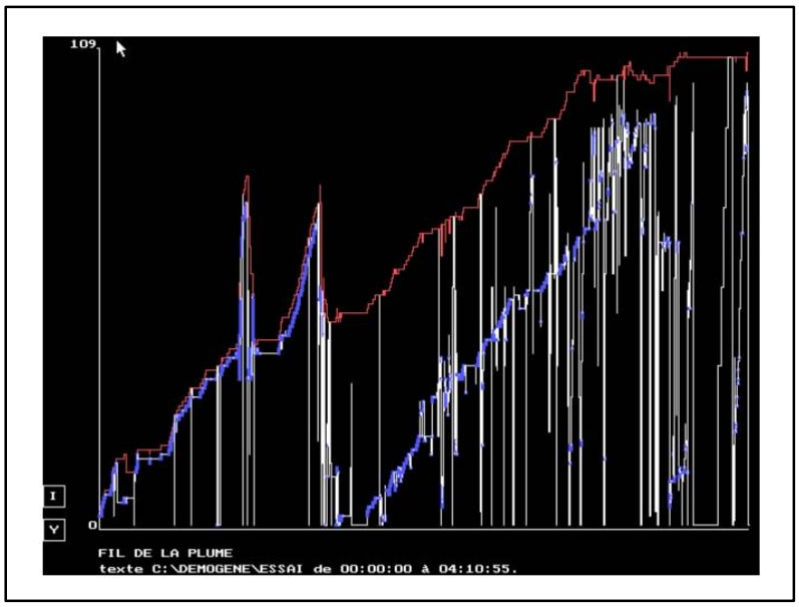

Figure 8. AFP sofware

Here are the data representation and data presentation characteristics of the AFP representation :

Data representation	
What are the variables used?	Temporality and spatiality.
How are they represented?	The *x-axis* represents time while y-axis represents the the number of lines of text (Doquet-Lacoste, 2003).
What is the degree of precision?	Not precise, it's an overview of the process.
Data presentation	
What are the characterstics of the data presentation?	The axis's names are not well identified. The overall presentation has not been optimized for the viewer. The dark background forces the eye to deal only with dominant colors and therefore is not optimal. It is better to use the strongest colors for the data the user needs to draw attention to (Kirk, 2012).

3.3.5 Temporal representation combined with the use of sources

Graph representation *LS / GIS* was echoed by Leijten and Van Waes (2013) by adding the source elements, which the writer was referring to while writing, which provides a better overview of the activities the writer at a given time t.
A strength of this representation is the combination of more then one representation. It allows the researcher to have a more comprehensive overview of the writing process.

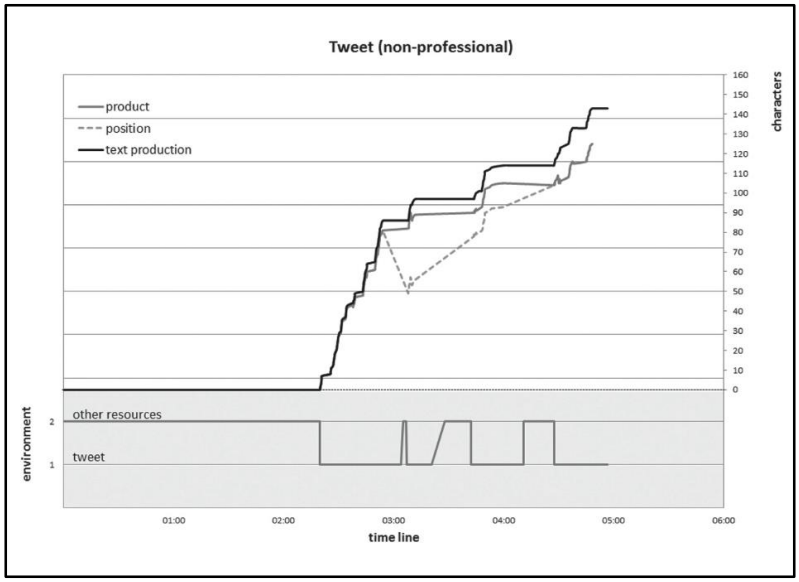

Figure 9. *InputLog* temporal representation

Here are the data representation and data presentation characteristics of the *Inputlog* temporal representation :

Data representation	
What are the variables used?	Temporality and spatiality.
How are they represented?	The *x-axis* represents time. The *y-axis* represents the total number of characters typed. The top line shows the text production. The centered line is the actual lenght of the text at this specific time. The lower line is the cursor position in the text. (Leijten & Van Waes, 2013) The writer's attention is represented by a line showing that either the person is looking at sources or to the writing window (the tweet in this case).
What is the degree of precision?	Not precise, it's an overview of the process.
Data presentation	
What are the characterstics of the data presentation?	The axis's names are well identified. This representation is usine a combination of colors to separate many things : environment section and the writing process evolution. This helps visually the user to analyze.

3.3.6 The graph representation

Graphs are mathematical objects that consists in nodes (points) and edges (line eventually joining them). As such, graphs are based upon relations between nodes and may be used for modeling purpose.

The position of the nodes is not important in the definition of the graph and is left to the user.

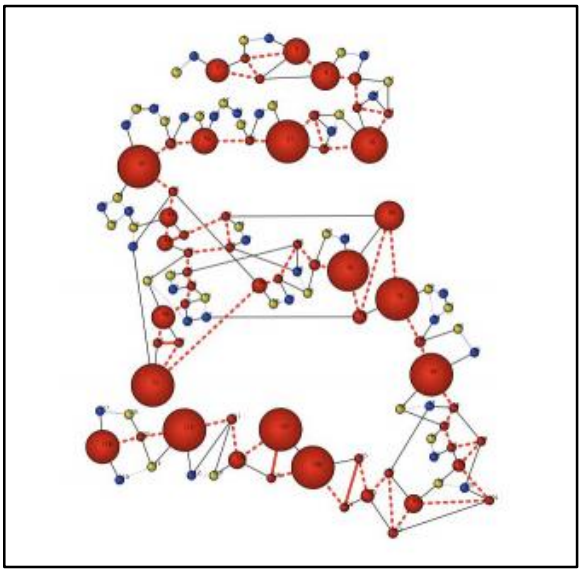

Figure 10. The graph representation

Here are the data representation and data presentation characteristics of the graph representation :

Data representation	
What are the variables used?	Chronology and spatiality.
How are they represented?	Nodes represent a a sequence of consecutive (both in time and space) operations of the same nature (addition/suppression). Edges represent a spatial or temporal relation between nodes. The position of the nodes is not important in the definition of the graph and is left to the user (Caporossi and Leblay 2011,2014 and 2015).
What is the degree of precision?	Intermediate : it is possible to gain precise insight by seeing the content of the nodes that contain keystrokes while the whole process is also globally viewable with the general graph structure.
Data presentation	
What are the characterstics of the data presentation?	The use of colors help the user knowing the nature of the grouped operations. Red is used for the text that is present in the final text. Yellow is for the additions that have been suppressed and blue stands for the suppression operations. The use of different line rendering for the edges according to the fact the operations are made consecutively (black line) or that they are chronologically successives in the final text (red dotted line). The fact that it is possible to move the nodes and to know the content of it by clicking on it makes it interactive. Allowing the user to interact with a visualization is a way to increase the probabilities to find patterns (Kirk, 2012).

This representation is halfway between the detailed representations and the overviews. The dynamic aspect of the writing process is highlighted (Caporossi & Leblay, Online Writing Data Representation : A Graph Theory Approach, 2011) and the linear progression represents the

chronology (Leblay, Le Temps de l'Écriture. Genèse, durée, représentations, 2011). One of its strength is to show well the temporal and chronologic relation between the operations making it possible to identify it in a structured way. Another advantage of this visualization of the writing process is that it « *can handle moving text position* » (Southavilay, Yacef, Reimann, & Calvo, 2013).

3.4 The progressive visualization

This visualization is a GIS version of the graph representation. It has been created to merge the advantages of other visualizations and to illustrate more variables in order to represent the writing process as a whole (Becotte-Boutin, Caporossi & Hertz, 2015).

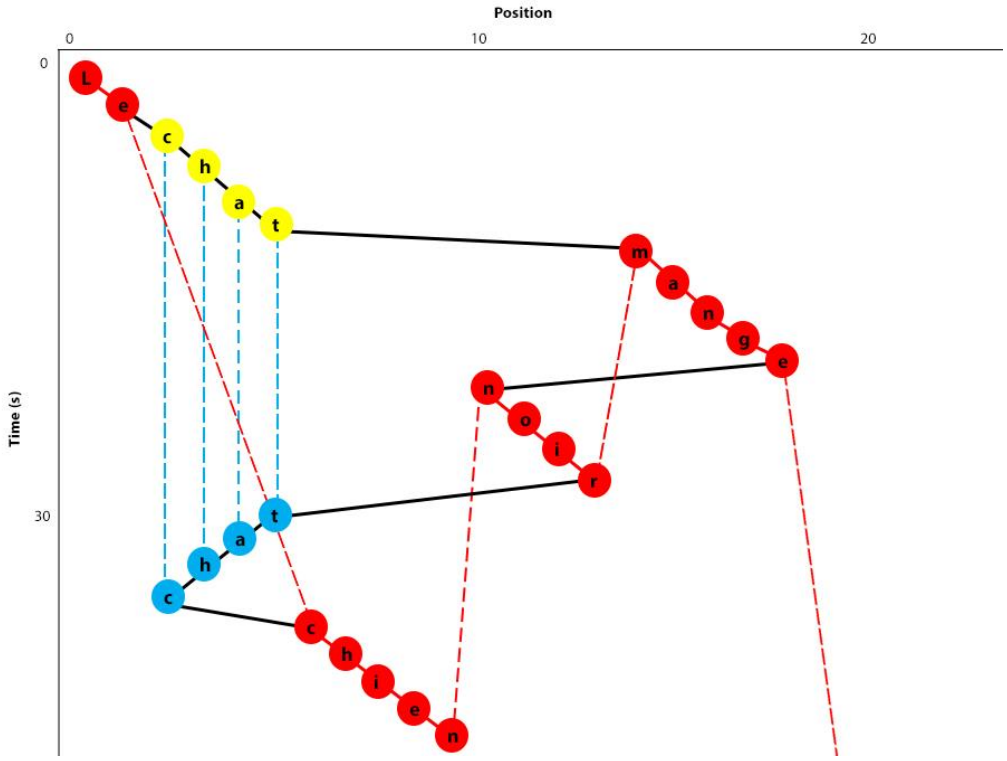

Figure 11. The progressive visualization

Here are the data representation and data presentation characteristics of the graph representation :

Data representation

What are the variables used?	Temporality, chronology and spatiality.
How are they represented?	Similar to the GIS, each operations (keystroke) owns its coordinate (x,y). The *x-axis* represents the time and the *y-axis* represent the spatiality of the operations. Each keystroke is also a node and is linked with an edge to the two nodes that are temporally neighbourd to it (previous and next). It can also be linked with an edge to another operation, if they are chronologically neighbours in the final text.

What is the degree of precision?	Precise and large depending of the level of « zoom » of the process.
Data presentation	
What are the characterstics of the data presentation?	The use of colors help the user knowing the nature of the sole operation. The same code as for the graph representation is used. Red is used for the text that is present in the final text. Yellow is for the additions that have been suppressed and blue stands for the suppression operation. The use of different line rendering for the edges according to the fact the operations are made consecutively (black line) or that they are chronologically successives in the final text (red dotted line). It is also possible to zoom in or out, to have a really precise view of the process or to view it as an overview, similarly to the GIS.

4. Conclusion

In the previous chapters, visualization and interaction has been described to help users to visually analyze time-oriented data. Analysts can look at the data, explore how they are interconnected sot hey can understand it. This is possible thanks to human visual perception and the fact that humans are quite good at recognizing patterns, finding interesting and unexpected solutions, classifying data, combining knowledge from different (colored) sources, and being creative in general.

This holds true unless the problem to be solved exceeds a certain size. Very large time-series or data that consist of many thousands of time-dependent variables can usually not be grasped by human observers. In such cases, we need the proficiency of computing systems to assist the knowledge crystallization from time-oriented data. If the problem size is sufficiently large, computers are better (i.e., faster and more accurate) than humans at numeric and symbolic calculations, logical reasoning, and searching (Aigner, Miksch, Schumann, & Tominski, 2011, p. 127).

Within the rewriting study field (i.e. revision), digital assessment tools are becoming a real object of research. Among these tools, we think that visualization occupies a prominent place. Lying in the direct lineage of transcripts made by text genetics, time-oriented data renews the paradigm by being complemented with mathematical colored models. One of the principal strengths of graph-related visualizations is the data structure that lies behind it and the post-process analysis possibilities (Vathy-Fogarassy & Abonyi, 2013, p. vi).

It is not only question of assessing the image of the product which has been written, but assess the image of the process of what is being written. In a digital world, the exploitation of the image and what it represents (representation & presentation) is key to the analyze of contemporary written productions.

BIBLIOGRAPHY

Abukhodair, F. A., Riecke, B. E., Erhan, H. I., & Shaw, C. D. (2013). Does interactive animation control improve exploratory data analysis of animated trend visualization? Burlingame, California, USA: Visualization and Data Analysis 2013.

Aggarwal, C. C., & Wang, H. (2010). Graph data management and mining: a survey of algorithms and applications. In C. C. Aggarwal, & H. Wang (Eds.), *Managing and mining graph data* (pp. 13-68). New York: Springer.

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. New York: SIGMOD '93 Proceedings of the 1993 ACM SIGMOD international conference on Management of data.

Ahuja, R. K., Magnanti, T. L., & Orlin, J. B. (1993). *Network flows: Algorithms and applications.* Upper Saddle River: Prentice Hall.

Aigner, W., Miksch, S., Schumann, H., & Tominski, C. (2011). Visualization of Time-Oriented Data. *Human-Computer Interaction Series* . London: Springer.

Alamargot, D., & Lebrave, J.-L. (2009). The study of professional writing: A joint contribution from cognitive psychology and genetic criticism. *European Psychologist* (doi:10.1027/1016-9040/a000001).

Alamargot, D., Caporossi, G., Chesnet, D., & Ros, C. (2011). What makes a skilled writer? Working memory and audience awareness during text composition. *Learning and Individual Differences , 21* (5), 505-516.

Alamargot, D., Caporossi, G., Chesnet, D., & Ros, C. (2011). What Makes a Skilled Writer? Working Memory and Audience Awareness During Text Composition. *Learning and Individual Differences , 21*, 505-516.

Baaijen, V. M., Galbraith, D., & Glopper, K. d. (2012). Keystroke analysis: reflections on procedures and measures. *Written Communication , 29* (3), 246-277.

Barabasi, A.-L., Albert, R., & Jeong, H. (2000). Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A , 281*, 69-77.

Bartram, L. (1997). Perceptual and interpretative properties of motion for information visualization. *Proceedings of the 1997 workshop on new paradigms in information visualization and manipulation* , 3-7.

Bath, P. A., Craigs, C., Maheswaran, R., Raymond, J., & Willett, P. (2005, Nov-Dec). Use of graph theory to identify patterns of deprivaton and high morbidity and mortality in public health data sets. *Journal of the American Medical Informatics Association , 12* (6), pp. 630-641.

Becotte-Boutin, H.-S., Caporossi, G., Hertz, A. (2015). The progressive visualization, a new tool for analyzing the writing process. Cahiers du Gerad.

Berthold, M. R. (2011). Bisociative knowledge discovery. In J. Gama, E. Bradley, & J. Hollmen (Eds.), *Advances in intelligent data analysis X* (pp. 1-7). Porto: Springer.

Bethel, E., Prabhat, P., Byna, S., Rubel, O., Wu, K., & Wehner, M. (2013). Why high performance visual data analytics is both relevant and difficult. Burlingame, California, USA: Visualization and Data Analysis.

Bondy, A. J., & Murty, U. (1982). *Graph theory with applications.* New York: Elsevier Science Publishing Co. Inc.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society , 15* (5), pp. 662-679.

Bramer, M. (2013). *Principles of data mining* (2nd ed.). Portsmouth: Springer.

Breetvelt, I., Bergh, H. v., & Rijlaarsdam, G. (1994). Relations between writing processes and text quality: When and how? *Cognition and Instruction , 12* (2), 103-123.

Breetvelt, I., Van Den Bergh, H., & Rijlaarsdam, G. (1994). Relations between Writing Processes and Text Quality : When and How. *Cognition and Instruction , 12* (2), 103-123.

Caporossi, G., & Leblay, C. (2011). Online Writing Data Representation : A Graph Theory Approach. In *Lecture Notes in Computer Sciences 7014* (pp. 80-89).

Caporossi, G., & Leblay, C. (2014). Outils de visualisation de données enregistrées. In C. Leblay, & G. Caporossi (Eds.), *Temps de l'écriture: enregistrements et représentations* (pp. 147-166). Louvain-la-Neuve: Academia.

Caporossi, G. & Leblay, C. (2015). A graph theory approach to online writing data visualizaion. In G. Cislaru (Ed.) Writing(s) at the Crossroads: The Process-Product Interface. Amsterdam: John Benjamins. 171-181.

Carbone, A., & Gromov, M. (2001). Mathematical sclices of molecular biology. *Gazette des mathématiciens, édition spéciale , 88*, 11-80.

Chakrabarti, D., & Faloutsos, C. (2012). *Graph mining: laws, tools, and case studies.* Morgan & Claypool Publishers series.

Chesnet, D., & Alamargot, D. (2011, 10). Eye and Pen 2 manuel de l'utilisateur. Poitiers.

Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J. M., & Welton, C. (2009). MAD skills: new analysis practices for big data. *Proceedings of the VLDB Endowment 2.2* , 1481-1492.

Cox, M., Ortmeier-Hopper, C., & Tirabassi, K. E. (2009). Teaching Writing for the "Real World": Community and Workplace Writing. *The English Journal , 98* (5), 72-80.

Daelemans, W., Berck, P., & Gillis, S. (1997). Data mining as a Method for Linguistic Analysis: Dutch Diminutives. *Dutch Diminutives, Folia Linguistica, XXXI/I -2* , pp. 57-75.

De Looze, C. (2010). *Analyse et Interprétation de l'Empan Temporel des Variations Prosodiques en Français et en Anglais.* Marseille: Université de Provence.

Dehmer, M., & Basak, S. C. (2012). *Statistical and machine learning approaches for network analysis.* Somerset: John Wiley & Sons.

Doquet, C. (2014). Pour une approche linguistique de l'écriture enregistrée. In Academia (Ed.), *Temps de l'écriture: enregistrements et représentations* (pp. 21-42). LouvainlLa-Neuve.

Doquet-Lacoste, C. (2003). *Étude Génétique de l'Écriture sur Traitement de Texte d'Élèves de Cours Moyen 2, Année 1995-1996.* Paris: Université Sorbonne nouvelle.

Doquet, C & Leblay, C. 2014. Temporalité de l'écriture et génétique textuelle: vers un autre métalangage? *In* F. Neveu *et alii* (Ed.) *Actes numériques du 4ème Congrès Mondial de Linguistique Française*, Berlin, 19-23 Juillet 2014. [http://www.shs-conferences.org/articles/shsconf/pdf/2014/05/shsconf_cmlf14_01204.pdf]

Dzemyda, G., Kurasova, O., & Zilinskas, J. (2013). Multidimensional Data Visualization: Methods and Applications. *Springer Optimization and Its Applications 75* . Vilnius: Springer.

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

Foucambert, D., & Foucambert, J. (2014). Gestes d'écriture et caractéristiques linguistiques des textes achevés. In C. Leblay, & G. Caporossi (Eds.), *Temps de l'écriture: enregistrements et représentations* (pp. 43-70). Louvain-la-Neuve: Academia.

Haas, C. (1989). How the Writing Medium Shapes the Writing Process : Effects of Word Processing on Planning. *Research in the Teaching of English , 23* (2), 181-207.

Helbing, S., & Balietti, S. (2011). From social data mining to forecasting socio-economic crises. *The European Physical Journal Special Topics , 195*, 3-68.

Henley, E. J., & Williams, R. (1973). *Graph theory in modern engineering.* Houston: Academic Press.

ITEM. (2014, mai 7). *Enjeux de recherche*. Retrieved juin 14, 2014 from ITEM: http://www.item.ens.fr/index.php?identifier=l-item

Kanawaty, G. (1996). *Introduction à l'étude du travail 3e édition* (Bureau International du Travail ed.). Genève.

Karimi, H. A. (2014). Big data :Techniques and technologies in geoinformatics. Boca Taton: CRC Press. From http://www.crcnetbase.com/isbn/9781466586550

Karp, R. (2005). Optimization problems related to internet congestion control. In M. C. Golumbic, & I. B.-A. Hartman (Eds.), *Graph theory, combinatorics and algorithms* (pp. 1-16). Springer.

Kirk, A. (2012). *Data visualization: a successful design process [electronic book].* Packt Pub.

Kollberg, P. (1996). *Rules for the S-notation: a compuyter-based method for representing revisions.* Stockholm, Sweden: IPLab, Royal Institute of Technology (KTH).

Kollberg, P. (1996). S-notation as a tool for analysing the episodic nature of revisions. Barcelona: European Writing Conferences.

Kuramochi, M., & Karypis, G. (2005). Finding frequent patterns in a large sparse graph. *Data mining and knowledge discovery , 11* (3), 243-271.

Latif, M. M. (2008). A State-of-the-Art Review of the Real-Time Computer-Aided Study of the Writing Process. *International Journal of English Studies , 8* (1), 29-50.

Leblay, C. (2012, 06 14). *En deçà du bien et du mal écrire*. Retrieved 07 07, 2014 from ITEM: http://www.item.ens.fr/index.php?id=578258

Leblay, C. (2011). *Le Temps de l'Écriture. Genèse, durée, représentations*. Retrieved 12 15, 2013 from https://www.jyu.fi/ajankohtaista/arkisto/2011/11/tiedote-2011-11-04-10-14-59-722468

Leblay, C. (2009). La question du déjà écrit dans le processus d'écriture observé en temps réel. une contribution de la génétique à la didactique. In I. Fenoglio & J.-M. Adam (éds.) Génétique de la production écrite et linguistique, *Modèles linguistiques*, Tome XXX, volume 59, 153-176.

Leblay, C., & Caporossi, G. (2014). Introduction aux données temporelles de l'écriture. In C. Leblay, & G. Caporossi (Eds.), *Temps de l'écriture: enregistrements et représentations* (pp. 5-15). Louvain-la-Neuve: Academia.

Lebrave, J.-L. (2001). Comment écriront-ils? *Diogène , 196* (4), 163-171.

Lebrave, J.-L., & Grésillon, A. (2009, 03 23). *Linguistique et génétique des textes : un décalogue.* Retrieved 12 14, 2013 from Item: http://item.ens.fr/index.php?id=434571

Leijten, M., & Van Waes, L. (2006). Inputlog : New Perspectives on the Logging of On-Line Writing Processes. In K. P. Lindgren (Ed.), *Computer Keystroke Logging and Writing* (pp. 73-94). Elsevier.

Leijten, M., & Van Waes, L. (2014). *Inputlog features*. Retrieved 07 25, 2014 from http://www.inputlog.net/description_features.html

Leijten, M., & Van Waes, L. (2013). Keystroke Logging in Writing Research : Using Inputlog to Analyze and Visualize Writing Processes. *30* (3), 358-392.

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data , 1* (1), Article 2.

Lindgren, E., & Sullivan, K. P. (2002). The LS Graph : A Methodology for Visualizing Writing Revision. *Language Learning , 52* (3), 565-595.

Lindgren, E., & Sullivan, K. P. (2006). Writing and the Analysis of Revision: An Overview. In E. Lindgren, & K. P. Sullivan, *Computer Keystroke Logging and Writing* (pp. 31-44). Elsevier.

Lindgren, E., Sullivan, K. P., Lindgren, U., & Spelman Miller, K. (2007). GIS for Writing: Applying Geographical Information Systems Techniques to Data Mine Writings' Cognitive Processes. In G. Rijlaarsdam, D. Galbraith, M. Torrance, & L. van Waes, *Writing and Cognition* (pp. 83-96). Amsterdam: Elsevier.

Lynch, C. (2008). How do your data grow? *Nature , 455* (4), 28-29.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., et al. (2011). *Big data: the next frontier for innovation, competition, and productivity.* McKinsey Global Institute.

Markowetz, A., Blasziewicz, K., Montag, C., Switala, C., & Thomas, E. S. (2014, April). Psycho-Informatics: Big Data shaping modern psychometrics. *Medical Hypotheses , 82* (4), pp. 405-411.

Merriam-Webster. (2014). *Apophenia, Merriam-Webster*. Retrieved 08 03, 2014 from http://www.merriam-webster.com/dictionary/apophenia

Midgette, E., Haria, P., & MacArthur, C. (2008). The effects of content and audience awareness goals for revision on the persuasive essays of fifth- and eighth-grade students. *Reading and Writing , 21*, 131-151.

Minelli, M., Chambers, M., & Dhiraj, A. (2013). *Big data, big analytics : Emerging business intelligence and analytic trends for today's businesses.* Wiley.

Miner, G., Elder, J., Nibset, B., Delen, D., Fast, A., & Hill, T. (2012). *Practical text mining and statistical analysis for non-structured text data applications* (Academic Press ed.). Saint-Louis.

New, E. (1999). Computer-Aided Writing in French as a Foreign Language : A Qualitative and Quantitative Look at the Process of Revision. *The Modern Language Journal , 83* (i), 80-97.

Ohlhorst, F. J. (2013). *Big Data Analytics : Turning Bi g Data into Big Money.* Wiley.

Olive, T., Lebrave, J.-L., Passerault, J.-M., & Le Bigot, N. (2010). La dimension visuo-spatiale de la production de textes: approches de psychologie cognitive et de critique génétique. *Langages , 177* (1), pp. 29-55.

Owston, R. D., Murphy, S., & Wideman, H. H. (1992). The Effects of Word Processing on Student's Writing Quality and Revision Strategies. *Research in the Teaching of English , 26* (3), 249-276.

Perrin, D. (2003). Progression Analysis (PA): Investigating Writing Strategies at the Workplace. *Journal of Pragmatics , 35*, 907-921.

Perrin, D., & Laemmel, S. (2014). Application à l'écriture journalistique. In C. Leblay, & G. Caporossi, *Temps de l'écriture, enregistrements et représentations* (pp. 171-192). Louvain-la-Neuve: Academia-L'Harmattan s.a.

Plane, S., Alamargot, D., & Lebrave, J.-L. (2010). Temporalité de l'écriture et rôle du texte produit dans l'activité rédactionnelle. *Langages , 177* (1), pp. 11-32.

Plane, S., Alamargot, D., & Levrabe, J.-L. (2010). Temporalité de l'Écriture et Rôle du Texte Produit dans l'Activité Rédactionnelle. *Langages , 177* (1), 11-32.

Roen, D. H., & Willey, R. (1988). The Effects of Audience Awareness on Drafting and Revising. *Research in the Teaching of English , 22* (1), 75-88.

Saha Ray, S. (2013). *Graph theory with algorithms and its applications in applied science and technology.* Rourkela, India: Springer India.

Severinson Eklundh, K., & Kollberg, P. (1996). A Computer Tool and Franework for Analysing On-Line Revisions. In *The Science of Writing : Theories, Methods, Individual Differences, and Applications* (pp. 163-188). Mahwah, NJ: Lawrence Erbaum.

Southavilay, V., Yacef, K., Reimann, P., & Calvo, R. A. (2013). Analysis of Collaborative Writing Processes Using Revision Maps and Probabilistic Topic Models. *Proceedings of the Third International Conference on Learning Analytics and Knowledge* , 38-47.

Spelman Miller, K., & Sullivan, K. P. (2006). Keystroke Logging : An Introduction. In *Computer Keystroke Logging and Writing* (pp. 1-10). Elsevier.

Stromqvist, S., Holmqvist, K., Johansson, V., Karlsson, H., & Wengelin, A. (2006). What Keystroke Logging can Reveal about Writing. In K. P. Lindgren (Ed.), *Computer Keystroke Logging and Writing* (pp. 45-71). Elsevier.

Sullivan, K. P., & Lindgren, E. (2014). La révision en production écrite enregistrée. In C. Leblay, & G. Caporossi (Eds.), *Temps de l'écriture: enregistrements et représentations* (pp. 71-92). Louvain-la-Neuve: Academia.

Takac, L., & Zabovsky, M. (2012). Data analysis in public social networks. Lomza, Poland: International Scientific Conference & International Workshop Present Day Trend of Innocations 2012.

Tarjan, R. E. (2005). Problems in data structures and algorithms. In M. C. Golumbic, & I. B.-A. Hartman (Eds.), *Graph theory, combinatorics and algorithms* (pp. 17-39). Springer.

Tufféry, S. (2010). *Data mining et statistique décisionnelle: l'intelligence des données.* Éditions Technip.

Unwin, A., Chen, C.-h., & Hardle, W. K. (2008). Introduction. In A. Unwin, C.-h. Chen, & W. K. Hardle, *Handbook of Data Visualization* (pp. 3-14). Berlin: Springer.

Van Waes, L., & Leijten, M. (2014). Inputlog 6.0: Pause and fluency analysis. . *Paper presented at the keystroke logging training school.* Antwerp.

Van Waes, L., & Leijten, M. (2014). Inputlog 6.0: State of the art. *Paper presented at the Keystroke logging training school.* Antwerp.

Van Waes, L., & Schellens, P. J. (2003). Writing Profiles : The Effect of the Writing Mode on Pausing and Revision Patterns of Experienced Writers. *Journal of Pragmatics , 35*, 829-853.

Vathy-Fogarassy, A., & Abonyi, J. (2013). *Graph-Based clusterting and data visualization.* Springer.

Veronis, J. (1988). Computerized Correction of Phonographic Errors. *Computers and the Humanities , 22* (1), 43-56.

Weikum, G., Hoffart, J., Nakashole, N., Spaniol, M., Suchanek, F., & Yosef, M. A. (2012). Big data methods for computational linguistics. *IEEE Data Engineering Bulletin , 35*, pp. 46-55.

Wengelin, A. (2014). Temps et pauses dans l'écriture au clavier. In C. Leblay, & G. Caporossi, *Temps de l'écriture: enregistrements et représentations* (pp. 97-124). Louvain-la-Neuve: Academia.

Wengelin, A., Torrance, M., Holmwvist, K., Simpson, S., Galbraith, D., Johansson, V., et al. (2009). Combined eyetracking and keystroke-logging methods for studying cognitive processes in text production. *Behavior Research Methods , 41* (2), 337-351.

Witten, I. (2004). Text mining. Hamilton, New Zealand. From http://www.cs.waikato.ac.nz/~ihw/papers/04-IHW-Textmining.pdf

WritingPro. (2014). *Writing Pro*. Retrieved 08 02, 2014 from http://www.writingpro.eu/

Wu, Y., Wang, L., Ren, J., & Ding, W. (2014). Mining sequential patterns eith periodic wildcard gaps. *Applied intelligence , 41* (1), 99-116.

Yau, N. (2011). *Visualize this: the flowing data guide to design, visualization and statistics.* Indianapolis: Wiley Publishing.

Zha, H., Yang, Y., Wang, J., & Wen, L. Transforming XPDL to Petri Nets. In *Business Process Management Workshops.*