Online Writing Data Representation: A Graph Theory Approach

Gilles Caporossi[1] and Christophe Leblay[2]

[1] GERAD and HEC Montréal, Montréal, Canada
[2] ITEM, Paris, France and SOLKI, Jyväskylä, Finland
gilles.caporossi@gerad.ca, Christophe.Leblay@kolumbus.fi

Abstract. There are currently several systems to collect online writing data in keystroke logging. Each of these systems provides reliable and very precise data. Unfortunately, due to the large amount of data recorded, it is almost impossible to analyze except for very limited recordings. In this paper, we propose a representation technique based upon graph theory that provides a new viewpoint to understand the writing process. The current application is aimed at representing the data provided by ScriptLog although the concepts can be applied in other contexts.

1 Introduction

The recent approaches of the study of writing based upon online record contrast with those based upon paper versions. The later ones are oriented on the page space and the former ones emphasis on the temporal dimension.

The models based upon online records have been developed in the 80s, with the pioneering work of Matsuhashi [13]. Returning to a bipolar division, Matsuhashi suggests distinguishing between the conceptual level (semantics, grammar and spelling) and sequential plan (the planning and phrasing).

The work that will follow will be, for the vast majority of them, related to the software for recording. Thus, the work of Ahlsen & Strömqvist [2] and Wengelin [19] are directly related to software applications ScriptLog, those of Sullivan & Lindgren [16] applications JEdit, those of Van Waes & Schellens [17], or Van Waes & Leijten [18] software InPutLog, those of Jakobsen [9] Translog software, and finally those of Chesnet & Alamargot [6] Eye and Pen software.

Without having to deny all work on the revision seen as a product (final text), these new approaches are all pointing the finger that writing is primarily a temporal activity. The multitude of software approaches developed for online recording of the writing activity shows a clear interest in the study of the process of writing. Would it be from the cognitive psychology or from the didactic point of view, analyzing the writing activity as a process is very important for researchers. What all these studies share is a common concern: that the record of writing is associated with its representation. As the raw data collected by the software is so detailed, it is difficult to analyze without preprocessing and without a proper representation that may be used conveniently by the researcher.

J. Gama, E. Bradley, and J. Hollmén (Eds.): IDA 2011, LNCS 7014, pp. 80–89, 2011.

In this paper, we propose a new representation technique that visually allows identification of basic operations (such as insertion, deletion, etc.), but also to identify portions in the document according to the processing activity of the writer. Each time this new representation technique was presented to psychologists or linguists it received a very positive feedback.

The paper is organized as follows : The next section describes the data involved and the third exposes some visualizing techniques. In the fourth and fifth section, we describe the proposed methodology and some transpositions from classical linguistic transformations of texts. The concepts proposed in the paper are illustrated by examples from the recording of the writing a short essay of 15 minutes under ScriptLog 1.4 (Mac version). This corpus composed entirely of Finnish-speaking writers [10] includes novice writers (first year of college Finnish, French, Core) and expert writers (university teachers).

2 Data

Elementary events that are recorded by a software such as ScriptLog are keyboard keystrokes or mouse clicks, as shown on figure 1. They represent the basic units that technically suffice to represent the whole writing process. Using them with no previous treatment may however not be a convenient way to look at the data. For instance, studying the pauses, their lengths and locations does not require the same level of information as studying the text revision process. In both cases, the same basic information is used but the researcher needs to apply agregation to a different level depending on his needs. Preprocessing to the data sometimes cannot be avoided. Given the large amount of data produced by the system (the log file corresponding to the data that was recorded in 15 minutes may yield up to 2000 lines), this preprocessing should, if possible, be automated to avoid errors.

time	type	from	to	key
0.00	10	1	0	<START>
4.21	7	0	0	L
4.46	7	1	1	I
4.75	7	2	2	E
5.05	7	3	3	U
5.26	7	4	4	
5.70	7	5	5	I
5.86	7	6	6	D
8.08	7	7	7	…
8.36	7	8	8	A
8.53	7	9	9	L
8.81	7	10	10	
12.28	5	11	11	<DELETE>
12.45	5	10	10	<DELETE>
12.61	5	9	9	<DELETE>
12.78	5	8	8	<DELETE>

Fig. 1. Excerpt from a log file (log) obtained using ScriptLog

3 Visualization Techniques

Visualization is an important part of this writing process but few visualization techniques actually exists. One of those techniques is the so-called linear representation in which any character written is displayed. Should a portion be deleted, it is crossed out instead of being deleted in order to show the text production in its process and not only as a final product. Cursor movements by arrows or mouse are also identified so that it is possible to follow the text construction process. An example of linear representation is given on figure 2. Such a representation has the advantage to display the text but remains difficult to understand in the case of a complex creation.

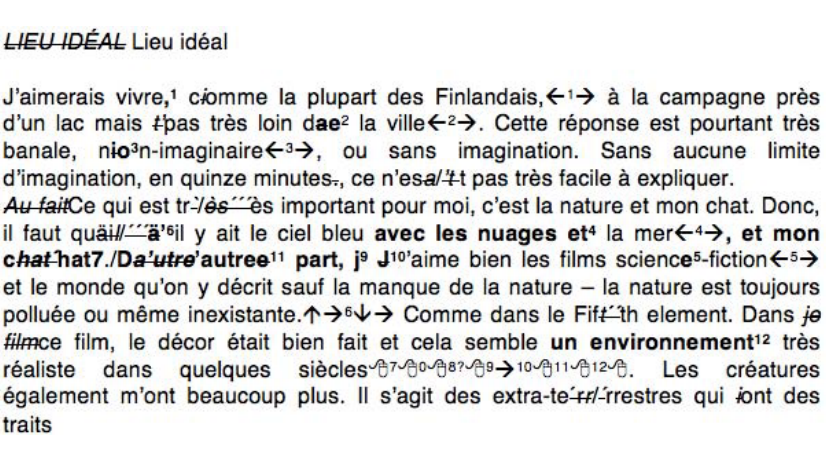

Fig. 2. Linear representation of a short 15 minutes text

Another way to visualize the creation process is based upon few values representing the text produced so far. Au fil de la plume in Genèse du texte [5] displays the position of the cursor as well as the total length of the text as a function of time. Such an approach indicates zones where the writer modifies already written text (see figure 3). This type of representation is also referred to as "GIS representation" and is used in various software such as InputLog [16].

One weakness of the "GIS representation" is that the position of visible text may not be correct as soon as an insertion or deletion occurs at a prior position. The position of any character being altered, it is difficult to figure out the part of the text involved by any subsequent modification. Another weak point is the lack of reference to the text corresponding to points on the graphic.

4 Method : Graph Representation

To assess the problem of the moving position of written text after revision, we propose here a slightly different approach in which each character is not described by the absolute position when written. Instead, we use a relative position which

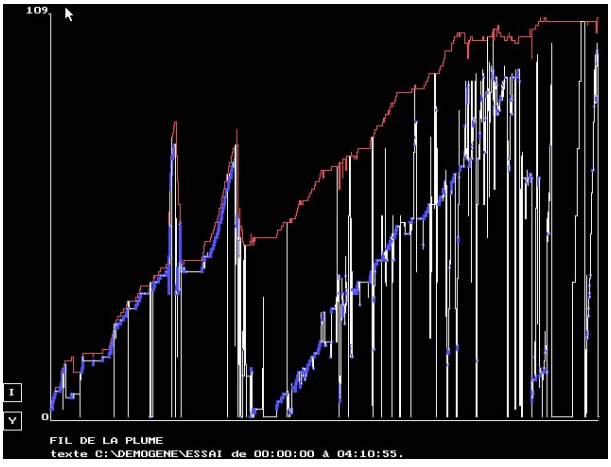

Fig. 3. Au fil de la plume - Genèse du Texte

proves to be more suitable to represent the dynamic aspect involved in the writing activity. Sequences of keystroke are merged together to form an entity involved in the conception of the text, which is represented by a node in the graph. Should two nodes interact, either by a chronological or spatial relation, they are joined by an edge, or link, showing this relation.

Graphs are mathematical tools based on nodes or vertices that are possibly connected by links or edges. Some application fields may be more or less related to graph theory. Chemistry is closely related to graph theory as some graph theoretical results directly apply to chemistry [3][4]. Some other applications refer to the algorithmic part underlying graph theory and networks, such as transportation, scheduling and communication. Since the 1990s, graphs are also used for the representation purpose in human sciences by the means of concept maps [14]. In the present paper, we propose to use graph representation to visualize a new kind of data.

Some examples of graph representations of the writing process are drawn on figure 4, for a novice production, and figure 5, for an expert.

- The size of a vertex is related to the number of elementary operations it represents. In the case of the novice writer, there are few large nodes, which shows a higher frequency of errors or typos. The text corresponding to each node could be displayed within the node, which would provide a representation close to the linear representation, but we decided not to do so here in order to keep the representations as simple as possible.
- The structure of the graph is also very informative, the structure of the graph of the novice is almost linear while a portion (in the middle of the graph) of the graph of the expert is much more complicated. This complex portion, between nodes 37 and 79 represents a part of the production that was rewritten and changed on a higher level, clearly not just from the lexical point of view.

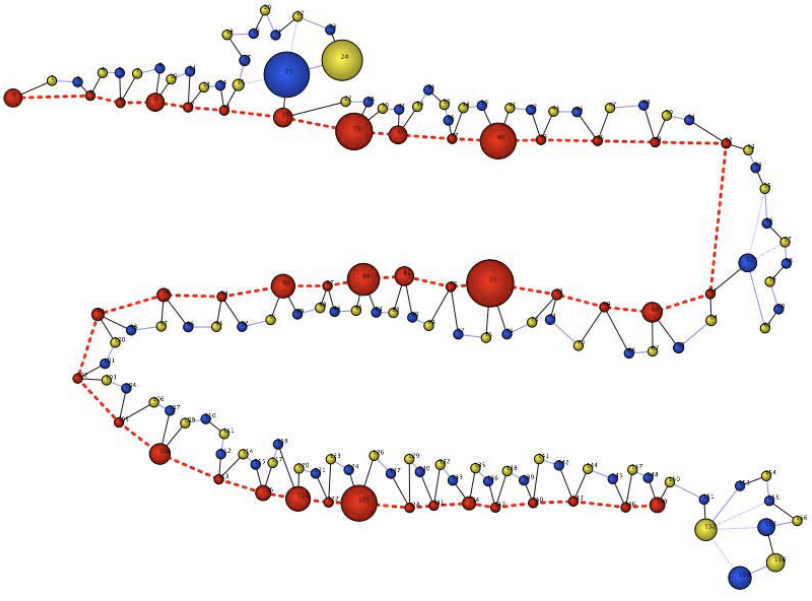

Fig. 4. Graph visualization : an example of novice writer

The key or mouse events found in the records are of three types: (i) additions or insertions a character or a space (ii) deletions of characters or spaces, and (iii) cursor moves by the mean of arrows or mouse. Spatially and temporally contiguous sequences are merged and represented by the nodes of the graph.

4.1 Nodes

The size and color of each node is an indicator of the number of elementary events it represents and their nature respectively. An addition that has later been removed appears in yellow, an addition that remains until the final text is drawn in red and deletions are displayed in blue. The final text thus appears in red while modifications that do not appear in the final text are either yellow or blue depending on their nature. The nodes are numbered according to their creation sequence.

4.2 Links

The nodes are connected by links or edges representing a spatial or temporal relation. The shape and color of edges indicate the nature of this relation. A solid line represents the chronological link (solid lines draw a path from the node 0 to the last node thru all nodes in the chronological order). Other links between nodes necessarily correspond to spatial relations. The link between an addition node and its deletion counterpart is drawn in blue and the spatial link between nodes that are part of the final text appear in red. Reading the content of nodes

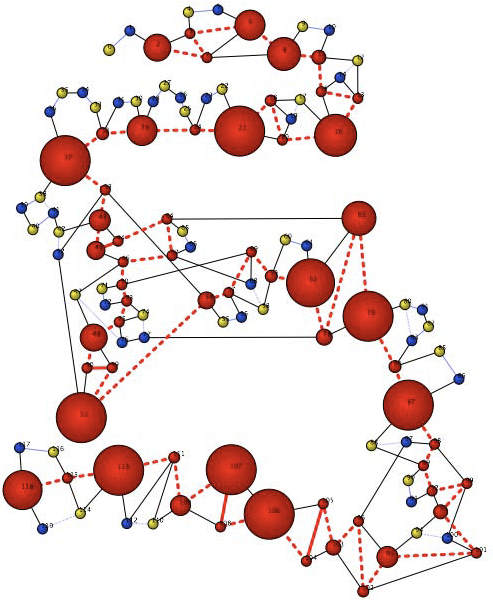

Fig. 5. Graph visualization : an example of expert writer

along the red link will therefore display the whole text in its final version. Note that the path describing the final text is composed of red nodes that are linked by red links.

5 Analysis of Graphical Patterns

Different avenues are possible to the analysis of graphs and we will concentrate here on the most useful ones. We will first identify patterns that correspond to some classical operations involved in the writing process. From a technical point of view, some operations will correspond to special subgraphs which could easily be recognized. The identification of these subgraphs is useful to analyze the graph as a representation of the writing process.

5.1 Additions and Insertions

Adding text could occur in three ways : (i) adding text at the end of the node that is being written will not be represented by any special pattern; (ii) inserting text in the node currently being written, but not at the end will cause this node to be split and a triangle with a solid red line appears as illustrated on figure 6. This solid red line is crossed in a way or the other depending if we follow the spatial or the chronological order. (iii) Inserting text in a node that is already written will cause this node to split and the corresponding configuration is shown on figure 7. From the graphic and linguistic standpoints, insertions corresponds

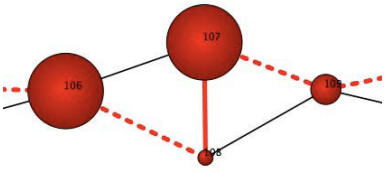

Fig. 6. Insertion in current node

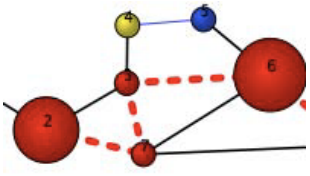

Fig. 7. Insertion

to addition inside (ii - iii) while the addition is the development of the text at its end (i).

5.2 Deletions

In the case of deletions as additions, different subgraphs are found depending if we are erasing the end of the last node, a part of the last node or a part of a node that was already written. The case of an immediate suppression (e.g. after a typing error) is shown in figure 8 where the text from node 4 is immediately removed by node 5, a deletion in the last node but not at its end is presented on figure 9 where node 110 is removed by node 112. A delayed removal will result in the subgraph shown in figure 10.

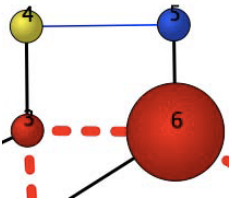

Fig. 8. Immediate deletion

5.3 Substitutions

In addition to these simple operations, some more complex operations may be viewed as sequences of these simple operations, but they nevertheless correspond to special subgraphs that may easily be recognized. For instance, replacement

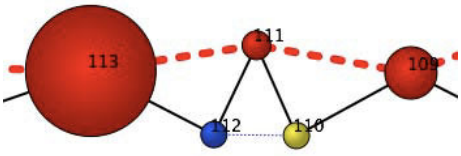

Fig. 9. Deletion of a part of the last node

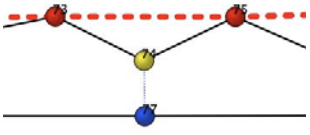

Fig. 10. Delayed elimination

may be viewed as a deletion immediately followed by an insertion at the same place. Figure 11 represents the subgraph corresponding to a replacement in a node that was already written. The replacement of a string in the last node, but not at its end is shown on figure 12. The interpretation of the replacement at the end of the last node is more complex because from a technical point of view, it is impossible to know whether the addition comes instead of the deleted portion or after. The deletion is not bounded and the addition may extend beyond the replacement, which is difficult to identify. In this case, some other information must be used by the researcher to interpret this sequence in a way or the other.

6 Summary and Future Research Directions

In this paper, we propose a new representation technique of written language production in which the problem of moving text position is handled. This technique allows the researcher to easily identify the portion of the document the writer is modifying. According to the reaction of researchers from linguistics working on the writing process, this representation is easier to understand than those previously available. An important aspect being the capability to visualize modification patterns from a spacial and temporal point of view on the same representation. It also seems that the intuition is more stimulated by a graph representation than it could be by linear or GIS representations.

Some important aspects of the graph representation in writing need further investigations.

- Emphasis on the temporal aspect by inserting nodes corresponding to long pauses (the definition of the minimum duration of a pause may be defined by the user), or indicating the time and duration corresponding to each node.
- Distinguish the various levels of text improvements as defined by Faigley and Witte [7] by distinguishing surface modification (correction of typos,

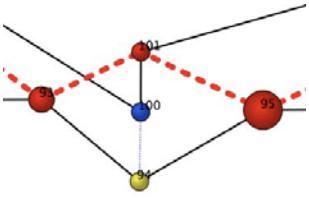

Fig. 11. Delayed substitution

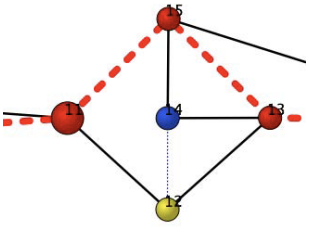

Fig. 12. Substitution in the last node

orthographical adjustment..) from text-based modification (reformulation, syntactic..). A first step in this direction would be to differentiate nodes involving more than a single word, which may be identified by the presence of a space before and after visible characters. Indeed, a second step would be to improve the qualification of the nature of the transformation represented by a node. This last part requires tools from computational linguistic.

– The graph drawing aspect is actually achieved by hand. Devising an algorithm that would automatically place vertices in such a way that (i) patterns are easy to recognize and (ii) the spatial aspect is preserved as much as possible, so that following the writing process remains easy.

References

1. Alamargot, D., Chanquoy, L.: Through the Models of Writing. Studies in Writing. Kluwer Academic Pubishers, Dordrecht (2001)
2. Ahlsén, E., Strömqvist, S.: ScriptLog: A tool for logging the writing process and its possible diagnostic use. In: Loncke, F., Clibbens, J., Arvidson, H., Lloyd, L. (eds.) Argumentative and Alternative Communication: New Directions in Research and Practice, pp. 144–149. Whurr Publishers, London (1999)
3. Caporossi, G., Cvetković, D., Gutman, I., Hansen, P.: Variable Neighborhood Search for Extremal Graphs. 2. Finding Graphs with Extremal Energy. J. Chem. Inf. Comput. Sci. 39, 984–996 (1999)

4. Caporossi, G., Gutman, I., Hansen, P.: Variable Neighborhood Search for Extremal Graphs. 4. Chemical Trees with Extremal Connectivity Index. Computers and Chemistry 23, 469–477 (1999)
5. Chenouf, Y., Foucambert, J., Violet, M.: Genèse du texte.Technical report 30802 - Institut National de Recherche Pédagogique (1996)
6. Chesnet, D., Alamargot, D.: Analyse en temps réel des activités oculaires et graphomotrices du scripteur: intérêts du dispositif Eye and Pen. L'année Psychologique 32, 477–520 (2005)
7. Faigley, L., Witte, S.: Analysing revision. College composition and communication 32, 400–414 (1981)
8. Jakobsen, A.L.: Logging target text production with Translog. In: Hansen, G. (ed.) Probing the Process in Translation. Methods and Results, Samfundslitteratur, Copenhagen, pp. 9–20 (1999)
9. Jakobsen, A.L.: Research Methods in Translation: Translog. In: Sullivan, K.P.H., Lindgren, E. (eds.) Computer Key-Stroke Logging and Writing: Methods and Applications, pp. 95–105. Elsevier, Amsterdam (2006)
10. Leblay, C.: Les invariants processuels. En deçà du bien et du mal écrire. Pratiques 143/144, 153–167 (2009)
11. Leijten, M., Van Waes, L.: Writing with speech recognition: the adaptation process of professional writers. Interacting with Computers 17, 736–772 (2005)
12. Lindgren, E., Sullivan, K.P.H., Lindgren, U., Spelman Miller, K.: GIS for writing: applying geographic information system techniques to data-mine writing's cognitive processes. In: Ri- Jlaarsdam, G. (series ed.), Torrance, M., Van Waes, L., Galbraith, D. (vol. eds.) Writing and Cognition: Research and Applications, pp. 83–96. Elsevier, Amsterdam (2007)
13. Matsuhashi, A.: Revising the plan and altering the text. In: Matsuhashi, A. (ed.) Writing in Real Time, pp. 197–223. Ablex Publishing Corporation, Norwood (1987)
14. Novak, J.D.: Concept maps and Vee diagrams: Two metacognitive tools for science and mathematics education. Instructional Science 19, 29–52 (1990)
15. Strömqvist, S., Karlsson, H.: ScriptLog for Windows - User's manual. Technical report - University of Lund: Department of Linguistic and University College of Stavanger: Centre for Reading Research (2002)
16. Sullivan, K.P.H., Lindgren, E. (eds.): Computer Keystroke Logging and Writing: Methods and Applications. Elsevier, Amsterdam (2006)
17. Van Waes, L., Schellens, P.J.: Writing profiles: The effect of the writing mode on pausing and revision patterns of experienced writers. Journal of Pragmatics 35(6), 829–853 (2003)
18. Van Waes, L., Leijten, M.: Inputlog: New Perspectives on the Logging of On-Line Writing Processes in a Windows Environment. In: Sullivan, K.P.H., Lindgren, E. (eds.) Computer Key-Stroke Logging and Writing: Methods and Applications, pp. 73–93. Elsevier, Amsterdam (2006)
19. Wengelin, Ä.: Examining pauses in writing: Theories, methods and empirical data. In: Sullivan, K.P.H., Lindgren, E. (eds.) Computer Key-Stroke Logging and Writing: Methods and Applications, pp. 107–130. Elsevier, Amsterdam (2006)