Analyse automatique des données scripturales prétraitées par des outils de visualisation

Hélène-Sarah Bécotte-Boutin^{a,1,2,3}, Gilles Caporossi^{2,3,4}, Alain Hertz^{1,3} et Christophe Leblay^{2,5}

¹ Polytechnique Montréal, 2900, boul. Édouard-Montpetit Campus de l'Université de Montréal 2500, chemin de Polytechnique, Montréal (Québec) H3T 1J4, Canada

² Institut des textes et manuscrits modernes (ITEM), CNRS : École normale supérieure [ENS] 45, rue d'Ulm 75005 Paris, France

³ GERAD, HEC Montréal 3000, chemin de la Côte-Sainte-Catherine, Montréal (Québec) H3T 2A7, Canada

⁴ HEC Montréal, 3000 chemin de la Côte-Sainte-Catherine Montréal (Québec) Canada H3T 2A7, Canada

⁵ University of Turku, 20014 Turun yliopisto, Finlande

Résumé. Plusieurs méthodes pour analyser le processus d'écriture ont été utilisés afin de comprendre les stratégies des scripteurs. L'outil principal pour analyser le processus d'écriture est le fichier log, qui contient de façon exhaustive et détaillée l'ensemble des opérations effectuées par le scripteur lors de la rédaction d'un texte. Les données qui y sont emmagasinées sont de quantité considérable et lorsqu'elles ne sont pas préalablement traitées, elles sont hostiles à être analysées par l'humain. Parmi les outils d'analyse utilisés, les représentations du processus d'écriture permettent l'agrégation des données grâce à un pré-traitement. Les structures sous-jacentes des données ainsi représentées sont généralement plus propices à l'analyse que les données brutes. Cet article vise à démontrer différentes méthodes d'analyse automatique pouvant être appliquées à ces structures afin de trouver ou confirmer des structures et tendances à travers les données.

Abstract. Several methods to analyze the writing process were used in order to understand the strategies of the writers. The main tool to analyze the writing process is the log file which contains all the operations performed by the writer when writing a text, in a comprehensive and detailed way. The data stored in it is of considerable amount and when not previously treated, it is not made to be analyzed by humans. Among the analytical tools used, the representations of the writing process allow aggregation of data through a pre-treatment. The underlying data structures as shown by these tools are generally conducive to analyzing the raw data afterwards. This article aims to demonstrate various automatic analysis methods that can be applied to these structures to find or confirm the structures and trends through data.

^a Auteur de correspondance : helene.becotte@polymtl.ca

1 Introduction

Les techniques d'analyse des textes produits foisonnent et de nouvelles disciplines sont entièrement créées autour de ce concept. Les techniques utilisées sont variées et proviennent de plusieurs domaines [1]. Les *statistiques*, les *tests A/B*, la *fusion et intégration de données*, la *régression*, la *classification*, les *règles d'association*, le *data mining*, la *visualisation*, le *text mining*, l'*analyse de sentiments*, les *réseaux de neurones*, l'*analyse de réseaux*, le *traitement du langage naturel*, la *simulation*, l'*analyse de séries chronologiques* sont toutes des techniques ou groupes de techniques qui s'entrecroisent [2].

Bien que ces techniques soient utilisées sur des textes produits et des données statiques et que le processus d'écriture soit pour sa part composé de données dynamiques multidimensionnelles, il est possible de s'en inspirer pour bénéficier des possibilités de telles techniques. C'est ce que nous nous proposons de faire dans ce travail, en nous concentrant sur la visualisation.

L'étude du processus d'écriture est complexe pour plusieurs raisons. La toute première raison est due à la singularité du processus individuel qui fait en sorte qu'il n'existe aucune méthode unique pour rédiger un texte de qualité [3]. La seconde est qu'il est composé principalement de données qui varient dans le temps. Ainsi, les opérations d'écriture, telles qu'elles ont été décrites par la génétique textuelle dans un environnement numérique [4], peuvent être listées en fonction de leur ordre chronologique d'apparition. Ces opérations et les pauses qui y sont attachées, sont difficiles à interpréter par l'humain lorsqu'elles ne sont pas combinées et contextualisées pour former des unités identifiables par l'homme : des mots, des phrases ou une section de ce qui deviendra du texte.

Parmi la panoplie disponible des outils d'analyse disponibles, les représentations du processus scriptural sont les plus utilisées grâce à leur facilité à traiter l'information.

Avec l'avènement de technologies dédiées à l'analyse du processus depuis les années 90, les possibilités d'automatisation se sont multipliées. Malgré cela, peu de méthodes permettent aujourd'hui aux chercheurs d'accélérer leur analyse et de se concentrer sur l'interprétation et la compréhension des mécanismes de révisions qui pourraient permettre à un scripteur d'améliorer la qualité de ses textes.

Dans ce travail, nous nous proposons de nous intéresser, tout d'abord, aux structures sous-jacentes aux données impliquées dans le processus scriptural en distinguant nettement entre les représentations réalisées dans le cadre de la linguistique cognitive et celles réalisées dans le cadre de la linguistique génétique. Ce qui nous permettra d'introduire à la méthodologie suivie et d'exposer les modalités d'enregistrement. Nous finirons par illustrer notre propos à l'aide d'un exemple tiré de notre corpus afin de préciser les objectifs poursuivis.

2 Prétraitement & structures des données scripturales

Depuis les années 80, plusieurs méthodes pour analyser le processus d'écriture ont été utilisées [5]. L'outil principal pour analyser le processus d'écriture est le fichier d'enregistrement, appelé log, qui contient de façon exhaustive et détaillée l'ensemble des opérations effectuées par le scripteur lors de la rédaction d'un texte [6]. Les données qui y sont emmagasinées sont considérables, et lorsqu'elles ne sont pas préalablement traitées, elles sont hostiles à l'analyse humaine [7]. Ce traitement préalable nous semble donc déterminant dans l'accès aux données scripturales enregistrées. Autrement dit, il est impossible d'exploiter les programmes d'enregistrement de l'écriture (dits aussi de *temps réel*), sans procéder, au préalable, à un redéploiement visuel des données obtenues. Les structures sous-jacentes des données ainsi représentées sont généralement plus propices à l'analyse que les données brutes inexploitable. La principale raison est que lorsqu'on cherche à explorer des données, plusieurs étapes sont nécessaires avant de pouvoir utiliser les techniques d'analyse à proprement parler. Le processus

simplifié par lequel les données sont acquises, traitées et analysées se résume ainsi : 1) *enregistrement*, 2) *partage*, 3) *tri*, 4) *recherche*, 5) *représentation visuelle des données* et 6) *analyse* [2]. Plusieurs de ces actions ont déjà été étudiées de manière individuelle et ont été l'objet de création de logiciels visant à enregistrer le processus d'écriture et à le représenter dans l'une ou l'autre de ses dimensions [8]. Ces logiciels sont tous différents et généralement destinés à un projet de recherche particulier [6]. Certains sont conçus pour traiter l'information et la retransmettent ensuite à l'utilisateur de façon simplifiée, sous la forme d'une représentation visuelle.

L'utilisation de représentations graphiques de données, appelées de façon plus générale *visualisations*, consiste à explorer et essayer de comprendre les grands ensembles de données [9, p. xvi]. Elles permettent notamment d'identifier des tendances, structures, irrégularités et relations entre les données sur une certaine période temporelle [10, p. 110]. Le format compact de la visualisation agrège les données et utilise les capacités cognitives de l'humain [11]. Le but principal de l'utilisation de ces images représentant le processus scriptural est de laisser l'œil trouver des structures sous-jacentes parmi les données [12].

Il est important cependant de considérer que l'utilisation de représentations visuelles des données comporte certains défauts. Puisque les conclusions d'analyse seront effectuées par l'utilisateur, les capacités de traitement de celui-ci influencent fortement les résultats obtenus [13]. Tout d'abord, l'être humain a une capacité limitée à traiter des quantités d'informations et il aurait aussi parfois tendance à percevoir des tendances et associations alors qu'il n'y en a pas [1]. Pour cette raison et pour maximiser le potentiel de conclusions possibles, il est pertinent de pouvoir confirmer mathématiquement ces conclusions grâce à des modèles d'analyse de données et de s'assurer qu'il existe effectivement une corrélation forte entre différentes variables [14].

Les techniques d'analyse utilisées sont ensuite dépendantes du type de pré-traitement effectué sur les données. Structurer les données d'une certaine manière plutôt qu'une autre influence donc nécessairement les résultats d'analyse obtenus par la suite. Conséquemment, lors du choix d'une technique particulière d'analyse, ce n'est pas la technique elle-même qui est importante mais l'utilisation finale des résultats obtenus qu'il faut considérer [15, p. 134]. Dans le but d'obtenir par exemple une compréhension du système sous-jacent aux données comme dans le cas de l'analyse du temps de l'écriture, il est important de s'assurer que le modèle utilisé et les résultats obtenus soient propice à l'interprétation [16, p. 209].

Il existe deux grands modèles de représentations qui prétraitent les données de manière à pouvoir appliquer des techniques d'analyse linguistique de données. Le tout premier, attaché aux travaux réalisés par une linguistique cognitive, est celui qui a été développé dans le cadre des *Systèmes d'Information Géographique* (dorénavant SIG). Le second, en réaction au premier, a été développé en associant, de manière pluridisciplinaire, la linguistique génétique et la théorie mathématique des graphes. A ces deux modèles, nous proposerons, dans ce travail, d'introduire un tout nouveau modèle, qui est en fait le mixte des deux grands modèles : la *représentation progressive*.

2.1 Linguistique cognitive et systèmes d'information géographiques (SIG)

Les SIG sont un type de représentation utilisé initialement pour traiter des données géographiques. Les principaux attributs sont l'identification spatiale et temporelle d'activités, qui ont une ressemblance avec les données issues de l'écriture [17]. Ces attributs sont utilisés en tant que coordonnées. En géographie, ces coordonnées sont souvent la latitude et la longitude [18, p. 1] mais peuvent également contenir l'élévation [18, p. 73]. En analyse du processus,

dans la majorité des cas, ces coordonnées vont être composées d'une mesure de temps et d'une quantité de révisions effectuées [17].

Voici un exemple d'une représentation effectuée sur un logiciel dédié pour les SIG [19] :

Fig. 1. Représentation SIG.

Notre propos n'étant pas de rentrer dans l'analyse de ces représentations, nous nous contenterons, dans le cadre de ce travail, d'en noter les principales caractéristiques. Ainsi, comme il est possible de le voir dans la figure 1, la contribution clé des SIG réside dans la structure qu'ils donnent aux données et aux possibilités de les analyser considérant l'emplacement physique des données dans le texte. Cette dimension du processus d'écriture se nomme *spatialité*. Des méthodes d'analyse spécifiques sont adaptées à ces structures de données de manière à mettre en relief certaines problématiques. Les principales techniques comprennent (1) l'analyse des *chevauchements entre différentes caractéristiques*, (2) *l'identification de sous-groupes*, (3) l'exploration de *tendances spatiales selon certaines variables* et 4) l'exploration de la *relation entre plusieurs variables en fonction de leur emplacement spatial* [20, p. 2].

Ont utilisé ce mode de représentation (SIG) principalement les études suivantes : *LS Graph* [21, 17], *Genèse du texte* [4], *GIS Graph* [19], *Timeline* [22] et *Inputlog* [23]. Toutes ces représentations ont en commun l'utilisation des axes X et Y, comme la figure 1 l'illustre, chacune exploitant des caractéristiques différentes : chronologie, topologie, avec un jeu de couleurs.

Les méthodes les plus pertinentes dans le cas de l'analyse du processus de l'écriture sont sans contredit celles qui permettent de mettre de l'avant l'aspect spatial des données. La visualisation est considérée comme étant la première étape pour n'importe quelle analyse spatiale [20, p. 19 ch.2]. Par la suite, la requête de données est la méthode la plus simple pour débiter l'analyse. Il s'agit d'une façon d'*agréger* les données à un niveau supérieur pour obtenir des informations concentrées sur une problématique particulière. Les statistiques sont l'étape suivante pour comprendre les caractéristiques des données.

Finalement, plusieurs sources d'erreur sont possibles et un choix de modélisation de données inapproprié est aussi considéré critique dans le cas des SIG [20, p. 18 ch.2].

2.2 Linguistique génétique et représentation par graphes

La représentation par les graphes permet de mettre davantage en relief l'aspect dynamique de l'écriture. Celle-ci est orientée sur la chronologie du processus d'écriture [24]. L'une des particularités est de bien gérer la transformation et les mouvements du texte grâce aux nœuds représentant des parties de texte (ajoutés, supprimés) reliés ensemble par des arcs définissant leur relation, soit chronologique ou spatiale [25]. Il est possible de voir le contenu des nœuds. Telle que mentionnée par Caporossi et Leblay [8], cette représentation montre néanmoins l'aspect temporel de la rédaction, soit le moment exact où le scripteur a effectué chacune des opérations représentées.

Voici l'exemple d'une représentation par graphes produite par Caporossi et Leblay [8] :

Fig. 2. Représentation génétique par graphes.

Cette représentation qui reprend une description génétique à partir des opérations élémentaires d'écriture (Ajout *versus* Suppression) doit être lue de la manière suivante:

- (0): Il me
- (1): <SUPPR> em II
- (2): <AJOUT> <CR><CR>Je ne vo
- (3): <AJOUT> lis
- (4): <SUPPR> sil
- (5): <AJOUT> is pas un lieu
- (6): <AJOUT> q
- (7): <SUPPR> q
- (8): <AJOUT> précis comme élément essentiel d
- (9): <SUPPR> d leitesse tmemélé emmoc sicérp
- (10): <AJOUT> ou
- (11): <AJOUT> en end
- (12): <SUPPR> dne ne
- (13): <AJOUT> un endroit pr
- (14): <AJOUT> +ec

(15): <SUPPR> ce+

(16): <AJOUT> écis comme élément essentiel d

Ce mode de représentation n'oublie pas l'importance du texte produit : il est possible de le suivre au moyen des traits pointillés rouges. Il est facile de noter tout le travail d'écriture réalisé par le scripteur, soit les 2/3 de ce qui est écrit (11 nœuds sur 16) n'apparaîtra pas dans le texte produit.

Mais, au delà de la simple appréhension de cette représentation mathématique, et afin de bien en comprendre les subtilités, nous proposons un succinct rappel des ressources théoriques des graphes.

Un graphe est une représentation visuelle simple constituée de points reliés ensemble par des lignes [26, p. 1]. Leur structure est utilisée dans des domaines variés puisqu'elle représente simplement des *relations entre objets* [27, p. 1]. La terminologie est plutôt intuitive. Les points sont appelés *nœuds* tandis que les lignes sont appelées *arêtes*. Les nœuds peuvent représenter des villes sur une carte, des individus dans un réseau social ou encore simplement des informations reliées entre elles par les arêtes [28, p. 2]. Deux nœuds sont dits *adjacents* ou *voisins* s'ils sont reliés par une arête. Finalement, le degré d'un nœud est le nombre de voisins que possède celui-ci [29, p. 4]. De manière plus formelle, un graphe $G = (V, E)$ est constitué d'un ensemble fini V de nœuds et d'un ensemble E d'arêtes. Un graphe peut être *orienté*, dans lequel cas $G = (V, A)$ a un ensemble A d'arcs. Un arc qui relie un sommet x à un sommet y sera noté $[x, y]$ tandis qu'un arc qui relie un sommet y à un sommet x sera noté $[y, x]$ ¹.

De manière appliquée, en prenant l'exemple des réseaux sociaux, un graphe non orienté serait utilisé pour décrire une plateforme telle que *Facebook*, pour décrire la relation mutuelle que deux individus partagent en étant *amis*. Un graphe orienté plutôt serait utile dans un cas comme *Twitter* où la relation n'est pas nécessairement réciproque. Un usager peut très bien *suivre* un autre usager, sans que l'inverse soit vrai. S'il y a une relation de précédence entre des actions, là encore l'orientation des arcs peut être pertinente [2, p. 29].

L'analyse de réseaux, que certains auteurs appellent plutôt *graph mining* [30], est relative à la théorie des graphes et est une technique d'analyse de données utilisée en *big data* [2, p. 29]. Cette technique permet en effet d'utiliser de bons algorithmes pour l'analyse et le traitement des données complexes pouvant être représentées sous forme de graphes [31]. Depuis une dizaine d'années, le développement d'algorithmes de *data mining* dédiés aux graphes a été source d'une augmentation d'intérêt de la part des milieux scientifiques et industriels [32]. Parmi les techniques adaptées les plus fréquentes, on retrouve la *classification* et l'*identification de tendances* [30]. Nous nous sommes donc posés la question de l'adaptabilité de cette théorie mathématique aux sciences humaines, et plus particulièrement à la linguistique de l'écrit.

Ainsi, les possibilités d'analyse de ces structures ont déjà été étudiées dans le contexte du processus de l'écriture. La mise en évidence de sous-graphes particuliers représentant les *patterns* des opérations les plus fréquentes (ajout vs insertion, suppressions immédiate et différée, remplacement), a déjà été réalisée. L'identification de ceux-ci est utile dans l'analyse du graphe représentant le temps de l'écriture [7]. Il serait aussi possible d'avoir une visée plus exploratoire pour l'analyse et chercher par exemple des tendances ou autres sous-structures dans les données.

La recherche de ce type de structures peut se faire sur un seul large graphe pour y trouver des structures qui sont récurrente un nombre minimal de fois. Elle peut également se faire sur un groupe de graphes, pour lequel l'objectif est de trouver des *structures qui sont récurrentes* dans une certaine fraction des graphes faisant partie de ce groupe [30]. En raison de la différence fondamentale entre les caractéristiques sous-jacentes aux données et à la formulation du problème, les algorithmes qui sont développés pour l'analyse d'un groupe de graphes ne

peuvent être appliqués à l'analyse des caractéristiques d'un seul graphe. Inversement, les algorithmes développés pour l'analyse à l'intérieur d'un large graphe peuvent être facilement adaptés pour l'analyse d'un ensemble de graphes [32].

Dans l'analyse du temps de l'écriture, les deux approches sont également pertinentes, pour d'abord repérer les *stratégies récurrentes* d'un scripteur, et pour ensuite comparer les stratégies des différents scripteurs entre eux. Les méthodes d'analyse seront appliquées sur les différents niveaux de graphes pour tenter de détecter des stratégies ou opérations nouvelles.

L'utilisation de ces propriétés en analyse de l'écriture permettrait ainsi de comparer plusieurs textes d'un même auteur pour consolider et mieux comprendre ses propres stratégies de rédactions sous forme de sous-ensembles de données. Il serait également possible de prendre plusieurs textes d'un ensemble de scripteurs ayant un niveau de maîtrise d'écriture semblable et ainsi de tenter de découvrir les récurrences entre des stratégies individuelles de rédaction.

3 Corpus et méthodologie

Nous nous proposons maintenant de préciser le cadre méthodologique dans lequel a été enregistré notre corpus.

3.1 Méthodologie

Ce projet est centré sur l'écriture, comme processus descriptible linguistiquement, s'inscrivant dans le temps et l'espace, c'est-à-dire dans l'avant d'un texte finalement stabilisé. Ce choix s'appuie sur le constat méthodologique, selon lequel le produit n'est pas l'image fidèle de la production. En adoptant une approche génétique du texte, associée à un logiciel d'enregistrement de l'écriture, nous cherchons à décrire comment divers scripteurs donnent lieu à diverses rédactions et révisions, invisibles dans le texte achevé. Trois orientations émergent alors: (1) le *rôle des opérations d'écriture-réécriture qui sous-tendent toute activité scripturale*, (2) la *chronologie qui caractérise les écritures experte, novice, universitaire et scolaire*, et (3) *l'impact de la visualisation de cette chronologie sur la description et la reproduction de ces écritures*. L'objectif est de démontrer que la temporalité est une dimension essentielle de l'écriture, tout autant que la spatialité. L'étude, bien trop souvent négligée de cette temporalité, est intimement liée aux modes de représentations et de visualisations numériques. Ce que nous nous proposons d'étudier, afin de répondre à une interrogation contemporaine élémentaire: comment donner à voir ce qui se passe précisément sur un écran, pendant le temps de l'écriture. L'équipe² a développé une méthodologie propre, à partir d'un outil mathématique (graphes) adéquat, qui permet, en parallèle aux modèles de représentations (SIG) déjà consacrés par la linguistique cognitive, une vraie cartographie génétique de la dynamique scripturale. Ce qui a nécessité la mise en place d'un logiciel dédié proche des attentes de la linguistique génétique.

3.2 Un nouveau programme d'enregistrement

Le programme *GenographiX* (GGX), a été mis en place spécifiquement pour mettre en évidence le travail génétique de réécriture des scripteurs. Voici l'environnement graphique offert par le programme :

Fig. 3. Le programme *GenographiX*.

Comme la figure 3 le montre, le programme *GenoGraphiX* est expérimental, fonctionnant uniquement, pour l'instant du moins, dans un environnement fermé. Le logiciel permet d'enregistrer le détail de l'activité du scripteur, et de représenter cette activité par l'entremise d'un graphe mathématique tel qu'il a été décrit précédemment. Les données brutes (*log*) sont enregistrées dans un format similaire à un programme de type *Scriptlog*. La liste des événements est alors analysée par le logiciel qui regroupe ces événements en *cellules* (ou *nœuds*, ou *arêtes*) et identifie les *liens* (ou *arcs*) qui peuvent exister entre les cellules. Dans la version courante du logiciel, chaque cellule est représentée par un rectangle dans lequel on peut lire le texte correspondant (les versions antérieures représentaient chaque cellule par un cercle, mais le texte correspondant n'était pas visible). Ces cellules peuvent être de trois natures : 1) du texte ajouté que le scripteur a supprimé avant la fin de la rédaction (en jaune), 2) du texte ajouté qui restera dans le document final (en rouge) et 3) du texte supprimé (en bleu). Des liens entre cellules représentent une relation temporelle ou spatiale. Comme nous l'avons déjà signalé, pour des raisons méthodologiques, une attention particulière est portée sur le texte final qui peut toujours être lu en suivant les cellules et les liens rouges. Il est toutefois aussi possible de suivre le travail de manière chronologique en suivant les liens en trait continu (les autres liens sont en pointillés). Afin de marquer la relation particulière qu'il y a entre l'ajout d'une portion de texte et sa suppression ultérieure, la cellule d'ajout et la cellule de suppression correspondant sont reliées par un lien vert. Un exemple de graphe est représenté par la figure 5.

3.3 Une double tâche

Il s'agit d'un corpus numérique d'écritures enregistrées, composé de deux tâches successives. La première, de nature narrative, se présente comme suit :

« Racontez une histoire. Vous devez intégrer dans l'ordre et sans les modifier les trois phrases :

-Elle habitait dans cette maison depuis longtemps.

-Il se retourna en entendant ce grand bruit.

-Depuis cette aventure, les enfants ne sortent plus la nuit. »

La seconde tâche, de nature argumentative, est la suivante :

« Qu'est-ce qui est important pour vivre ensemble ? ». Un sujet qui se prête à tous les âges, sans reposer sur une structure de type pour/contre représentative d'un modèle dominant des textes argumentatifs.

Précisons que la première reprend un protocole qui a déjà fait l'objet de recueil de corpus de textes auprès d'enfants et de futurs enseignants en formation, proposé au départ par M. Charolles dans le cadre d'études de la cohérence textuelle et repris notamment dans Garcia-

Debanc et Bonnemaïson [33]. L'analyse des processus d'écriture trouvera ainsi des points d'appui dans les études existantes des textes produits, étant donné que ces études, comme nous l'avons déjà souligné dans notre introduction, sont déjà très installées. Ce corpus est déjà partiellement enregistré ; il se compose comme suit :

Tableau 1. Corpus d'écritures enregistrées.

PUBLICS			SCRIPTEURS	TÂCHES (20 minutes/tâche)	
				Narration	Argument
Enfants	-	-	-	-	-
Adultes	Etudiants (Langue étrangère)	Licence	3	3	3
		Master	1	1	1
		Doctorat	-	-	-
	Etudiants (Langue maternelle)	Licence	6	6	1
		Master	3	3	3
		Doctorat	2	2	2
	Enseignants-chercheurs	Doctorat	3	3	3
TOTAL en heures et minutes				= 6h	= 4h20

Il a été fait le choix de consignes à même de fonctionner auprès de publics variés (âge, langues maternelle vs étrangère et niveau d'expertise), la distinction enfant vs adulte recoupant principalement des lieux d'enregistrement. Le corpus est donc composé, pour l'heure actuelle, de 14 scripteurs en langue maternelle et de 4 scripteurs en langue étrangère, pour un total de 10h 20 minutes d'enregistrement. Ce qui semble déjà caractéristique est le fait que ce soit le nombre d'heures d'écriture qui importe, et non plus, le nombre de pages rédigés. Cette distinction reste élémentaire entre des corpus de produits et des corpus de production.

4. Un entre-deux: la représentation mixte

Pour illustrer notre propos, nous proposons de suivre l'écriture d'une adulte de 30 ans. La représentation progressive est un hybride entre la représentation génétique par les graphes et les SIG [34]. Elle a été créée dans le but de combiner les avantages de ces deux représentations et de pouvoir représenter plus de variables en s'inspirant de bonnes pratiques en matière de visualisations. Voici, avant de donner l'exemple d'une représentation mixte, nommée progressive par Becotte-Boutin, Caporossi et Hertz [34], le texte produit (Lil-Gon-30-F-T1-Fi-Fr.txt). Il s'agit d'un enregistrement de 20 minutes :

Fig. 4. Texte produit.

Comme il est possible de le voir dans l'éditeur de GGX, le texte produit apparaît sous le format *txt*. Voyons maintenant les 20 premières secondes de cet enregistrement, sous la forme d'un graphe :

Fig. 5. Représentation mixte, dite *progressive*.

Ce graphe est *mixte* parce qu'il reprend une représentation inspirée des axes X (position spatiale) et Y (chronologie). Dans ce modèle mixte, il faut noter que les trois premières opérations effectuées, soit l'insertion d'un L, e et d'un espace sont en rouge. Ceci signifie qu'elles sont présentes dans la version finale du texte. Par la suite, la présence des opérations successives « ' tait » en jaune et leur double en bleu qui se situe au même emplacement sur l'axe des abscisses signifie que ces opérations ont d'abord été effectuées (en jaune) mais par la suite supprimées (en bleu). Le lien bleu pointillé qui relie l'opération jaune avec son double bleu signifie qu'elles occupent le même emplacement spatial dans le texte. Au bas de la représentation il est aisé de remarquer les opérations successives « était une », qui sont en rouge tout comme le début et sont d'ailleurs liées à l'espace en position 3, ce qui signifie que ces opérations sont successives dans la version finale du texte. En lisant successivement les opérations en rouge, liées chronologiquement par des liens de couleur rouge également, il est possible de constater que le texte final de cette portion représentée, peut se lire « Il était une », conformément à ce qu'il est indiqué à la figure 4.

Puisque les SIG ont cette particularité de bien mettre en relation l'aspect temporel et spatial du temps de l'écriture [17], ils ont en contrepartie le désavantage de n'être qu'une vue d'ensemble du processus. La conséquence de cet attribut est de rendre difficile pour l'analyste de le lier avec opérations et le contexte d'écriture [8]. Pour sa part, la représentation par les graphes met davantage l'accent sur le lien entre les opérations et leur séquence [7]. Il est cependant difficile de connaître le contenu exact de chaque nœud et la position temporelle des opérations.

Cette représentation reprend d'abord l'idée des graphes génétiques mais est incluse dans un graphique ayant comme axe des abscisses la position et comme axe des ordonnées le temps. Contrairement aux SIG dont le déroulement du temps se lit de bas en haut, cette représentation se lit de gauche à droite et de haut en bas, en suivant le modèle logique d'un texte [34]. Par

rapport à la représentation par les graphes, dans ce cas-ci, chaque nœud représente une opération individuelle.

La sous-structure des données de cette représentation possède entièrement les attributs des deux représentations. Dans un graphe, il est possible d'avoir des attributs complémentaires, souvent considérés comme des poids [26, p. 15]. Chaque sommet représente ainsi une opération effectuée. L'opération peut être soit une insertion ou une suppression. De façon similaire aux fichiers log, chaque opération possède les attributs *temps*, *position* et *type*. Les arcs possèdent également des attributs. Nécessairement, ils ont une *origine* et une *destination* mais dans ce cas-ci ils ont aussi un *type* [34]. Cette structure permet ainsi d'appliquer les notions mathématiques et modèles propres à la théorie des graphes pour tenter de déceler des tendances et sous-structures.

Les attributs des sommets peuvent dans un deuxième temps être utilisés plus simplement en tant que coordonnées et reprendre les méthodes propres aux SIG.

5. Conclusion

Au sein des approches intéressées par les processus d'écriture-réécriture sur écran, les outils numériques d'évaluation sont en train de devenir un véritable objet de recherche, tout autant que les médiums de productions textuelles. Parmi ces outils d'évaluation, la visualisation de données occupe une place de choix. Se situant directement dans la tradition des transcriptions réalisées par la génétique (schématiquement diplomatiques versus linéaires), la visualisation des données scripturales prolonge ces paradigmes en proposant des données dynamiques, et non plus statiques. Il n'est alors plus seulement question d'évaluer par l'image ce qui vient d'être écrit, mais bien d'évaluer par l'image ce qui est en train d'être écrit, d'être réécrit sur écran, dans des séquentialités singulières.

Le principal avantage d'utiliser des méthodes découlant d'une structure de données propre à un type de représentation du processus d'écriture est de pouvoir à la fois bénéficier des capacités cognitives humaines à détecter des tendances et récurrences dans des données visualisées et de confirmer ces hypothèses avec des méthodes mathématiques. Parmi les différentes représentations explorées, il semble que l'utilisation de la représentation mixte, dite progressive, permette de structurer les données en ayant accès à un bassin plus large de méthodes. Il serait tout de même intéressant de comparer ces représentations et certaines de leurs méthodes d'analyse respectives afin de confirmer leurs avantages et élaborer un cadre méthodologique pour permettre aux chercheurs d'avoir un accès facile à ces compléments d'analyse.

Dans l'état actuel de notre projet, les attentes sont les suivantes: 1) Affiner notre outil d'enregistrement (GGX) afin de pouvoir aussi enregistrer dans des environnements écologiques comme, par exemple, *Word* ou *Open Office*. Pour l'instant, cet outil reste confidentiel. 2) Mettre en place une base de données, étendue, composée d'heures d'enregistrement. Une telle base de données n'existe pas actuellement en langue française. 3) Pouvoir rendre accessible cette base de données sur la plate-forme dédiée (TEMPUS), l'idée étant bien de mettre à disposition des chercheurs des données temporelles variées. 4) Développer les concepts théoriques nécessaires. Et 5) de procéder à des descriptions affinées en littérature débutante et experte.

Références

- [1] D. Boyd et K. Crawford, «Critical questions for big data,» *Information, Communication & Society*, vol. 15, n° 15, pp. 662-679, 2012.
- [2] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh et A. Hung Byers, «Big data: the next frontier for innovation, competition, and productivity,» McKinsey Global Institute, 2011.
- [3] S. Plane, D. Alamargot et J.-L. Lebrave, «Temporalité de l'écriture et rôle du texte produit dans l'activité rédactionnelle,» *Langages*, vol. 177, n° 11, pp. 11-32, 2010.
- [4] C. Doquet-Lacoste, *Étude Génétique de l'Écriture sur Traitement de Texte d'Élèves de Cours Moyen 2, Année 1995-1996*, Paris: Université Sorbonne nouvelle, 2003, p. 633.
- [5] K. S. Miller et K. P. Sullivan, «Keystroke Logging: An introduction,» chez *Computer keystroke logging and writing*, Oxford, Elsevier, 2006, pp. 1-10.
- [6] K. P. Sullivan et E. Lindgren, «La révision en production écrite enregistrée,» chez *Temps de l'écriture: enregistrements et représentations*, C. Leblay et G. Caporossi, Éd., Louvain-la-Neuve, Academia, 2014, pp. 71-92.
- [7] G. Caporossi et C. Leblay, «Outils de visualisation de données enregistrées,» chez *Temps de l'écriture: enregistrements et représentations*, C. Leblay et G. Caporossi, Éd., Louvain-la-Neuve, Academia, 2014, pp. 147-166.
- [8] G. Caporossi et C. Leblay, «Online Writing Data Representation : A Graph Theory Approach,» chez *Lecture Notes in Computer Sciences 7014*, 2011, pp. 80-89.
- [9] N. Yau, *Visualize this: the flowing data guide to design, visualization and statistics*, Indianapolis: Wiley Publishing, 2011, p. 358.
- [10] M. Minelli, M. Chambers et A. Dhiraj, *Big data, big analytics : Emerging business intelligence and analytic trends for today's businesses*, Wiley, 2013, p. 224.
- [11] F. Blanchard, *Visualisation et classification de données multidimensionnelles application aux images multicomposantes*, Université de Reims Champagne Ardenne , 2005.
- [12] M. Tory et T. Moller, «Human factors in visualization research,» *IEEE Transactions on visualization and computer graphics*, vol. 10, n° 11, pp. 72-84, 2004.
- [13] G. Dzemyda, O. Kurasova et J. Zilinskas, «Multidimensional Data Visualization: Methods and Applications,» Springer, Vilnius, 2013.
- [14] T. Moller, B. Hamann et R. Russell, «Preface,» chez *Mathematical foundation of scientific visualization, computer graphics, and massive data exploration*, Berlin, Springer, 2009, pp. V-VII.
- [15] H. A. Karimi, «Big data :Techniques and technologies in geoinformatics,» Boca Taton: CRC Press, 2014.

- [16] S. Tufféry, Data mining et statistique décisionnelle: l'intelligence des données, Éditions Technip, 2010, p. 705.
- [17] E. Lindgren et K. P. Sullivan, «The LS Graph : A Methodology for Visualizing Writing Revision,» *Language Learning*, vol. 52, n° 13, pp. 565-595, 2002.
- [18] J. Van Sickle, Basic GIS coordinates, Second Edition, Boca Raton: CRC Press, 2010.
- [19] E. Lindgren, K. P. H. Sullivan, U. Lindgren et K. Spelman Miller, «GIS for Writing: Applying Geographical Information Systems Techniques to Data Mine Writings' Cognitive Processes,» chez *Writing and Cognition*, Amsterdam, Elsevier, 2007, pp. 83-96.
- [20] C. D. Lloyd, Spatial data analysis: an introduction for GIS users, New York: Oxford University Press, 2010.
- [21] M. Leijten et L. Van Waes, «Inputlog : New Perspectives on the Logging of On-Line Writing Processes,» chez *Computer Keystroke Logging and Writing*, K. P. S. a. E. Lindgren, Éd., Elsevier, 2006, pp. 73-94.
- [22] A. Wengelin, M. Torrance, K. Holmqvist, S. Simpson, D. Galbraith, V. Johansson et R. Johansson, «Combined eyetracking and keystroke-logging methods for studying cognitive processes in text production,» *Behavior Research Methods*, vol. 41, n° 12, pp. 337-351, 2009.
- [23] M. Leijten et L. Van Waes, «Keystroke Logging in Writing Research : Using Inputlog to Analyze and Visualize Writing Processes,» vol. 30, n° 13, pp. 358-392, 2013.
- [24] C. Leblay, «Le Temps de l'Écriture. Genèse, durée, représentations,» 2011. [En ligne]. Available: <https://www.jyu.fi/ajankohtaista/arkisto/2011/11/tiedote-2011-11-04-10-14-59-722468>. [Accès le 15 12 2013].
- [25] V. Southavilay, K. Yacef, P. Reimann et R. A. Calvo, «Analysis of Collaborative Writing Processes Using Revision Maps and Probabilistic Topic Models,» *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pp. 38-47, 2013.
- [26] A. J. Bondy et U. S. R. Murty, Graph theory with applications, New York: Elsevier Science Publishing Co. Inc., 1982.
- [27] R. K. Ahuja, T. L. Magnanti et J. B. Orlin, Network flows: Algorithms and applications, Upper Saddle River: Prentice Hall, 1993.
- [28] E. J. Henley et R. A. Williams, Graph theory in modern engineering, Houston: Academic Press, 1973.
- [29] S. Saha Ray, Graph theory with algorithms and its applications in applied science and technology, Rourkela, India: Springer India, 2013.
- [30] C. C. Aggarwal et H. Wang, «Graph data management and mining: a survey of algorithms and applications,» chez *Managing and mining graph data*, C. C. Aggarwal et H. Wang, Éd., New York, Springer, 2010, pp. 13-68.
- [31] L. Takac et M. Zabovsky, «Data analysis in public social networks,» Lomza, Poland, 2012.

- [32] M. Kuramochi et G. Karypis, «Finding frequent patterns in a large sparse graph,» *Data mining and knowledge discovery*, vol. 11, n° %13, pp. 243-271, 2005.
- [33] C. Garcia-Deban et K. Bonnemaïson, «La gestion de la cohésion textuelle par des élèves de 11-12 ans: réussites et difficultés,» chez *CMLF 2014*, Toulouse.
- [34] H.-S. Becotte-Boutin, G. Caporossi et A. Hertz, «The progressive visualization, a new tool for analyzing the writing process,» *Cahiers du GERAD*, vol. 131, 2015.
- [35] H. Zha, Y. Yang, J. Wang et L. Wen, «Transforming XPDL to Petri Nets,» chez *Business Process Management Workshops*.
- [36] Y. Wu, L. Wang, J. Ren et W. Ding, «Mining sequential patterns eith periodic wildcard gaps,» *Applied intelligence*, vol. 41, n° %11, pp. 99-116, 2014.
- [37] WritingPro, «Writing Pro,» 2014. [En ligne]. Available: <http://www.writingpro.eu/>. [Accès le 02 08 2014].
- [38] I. Witten, «Text mining,» Hamilton, New Zealand, 2004.
- [39] G. Weikum, J. Hoffart, N. Nakashole, M. Spaniol, F. Suchanek et M. A. Yosef, «Big data methods for computational linguistics,» *IEEE Data Engineering Bulletin*, vol. 35, pp. 46-55, 2012.
- [40] J. Veronis, «Computerized Correction of Phonographic Errors,» *Computers and the Humanities*, vol. 22, n° %11, pp. 43-56, 1988.
- [41] L. Van Waes et P. J. Schellens, «Writing Profiles : The Effect of the Writing Mode on Pausing and Revision Patterns of Experienced Writers,» *Journal of Pragmatics*, vol. 35, pp. 829-853, 2003.
- [42] L. Van Waes et M. Leijten, «Inputlog 6.0: State of the art.,» chez *Paper presented at the Keystroke logging training school*, Antwerp, 2014.
- [43] L. Van Waes et M. Leijten, «Inputlog 6.0: Pause and fluency analysis.,» chez *Paper presented at the keystroke logging training school*, Antwerp, 2014.
- [44] R. E. Tarjan, «Problems in data structures and algorithms,» chez *Graph theory, combinatorics and algorithms*, M. C. Golumbic et I. B. Hartman, Éd., Springer, 2005, pp. 17-39.
- [45] S. Stromqvist, K. Holmqvist, V. Johansson, H. Karlsson et A. Wengelin, «What Keystroke Logging can Reveal about Writing,» chez *Computer Keystroke Logging and Writing*, K. P. S. & E. Lindgren, Éd., Elsevier, 2006, pp. 45-71.
- [46] K. Spelman Miller et K. P. Sullivan, «Keystroke Logging : An Introduction,» chez *Computer Keystroke Logging and Writing*, Elsevier, 2006, pp. 1-10.
- [47] K. Severinson Eklundh et P. Kollberg, «A Computer Tool and Franework for Analysing On-Line Revisions,» chez *The Science of Writing : Theories, Methods, Individual Differences, and Applications*, Mahwah, NJ, Lawrence Erlbaum, 1996, pp. 163-188.
- [48] D. H. Roen et R. Willey, «The Effects of Audience Awareness on Drafting and Revising,» *Research in the Teaching of English*, vol. 22, n° %11, pp. 75-88, 1988.

- [49] S. Plane, D. Alamargot et J.-L. Levrabe, «Temporalité de l'Écriture et Rôle du Texte Produit dans l'Activité Rédactionnelle,» *Langages*, vol. 177, n° 11, pp. 11-32, 2010.
- [50] D. Perrin, «Progression Analysis (PA): Investigating Writing Strategies at the Workplace,» *Journal of Pragmatics*, vol. 35, pp. 907-921, 2003.
- [51] R. D. Owston, S. Murphy et H. H. Wideman, «The Effects of Word Processing on Student's Writing Quality and Revision Strategies,» *Research in the Teaching of English*, vol. 26, n° 13, pp. 249-276, 1992.
- [52] E. New, «Computer-Aided Writing in French as a Foreign Language : A Qualitative and Quantitative Look at the Process of Revision,» *The Modern Language Journal*, vol. 83, n° 1, pp. 80-97, 1999.
- [53] G. Miner, J. Elder, B. Nibset, D. Delen, A. Fast et T. Hill, Practical text mining and statistical analysis for non-structured text data applications, Academic Press éd., Saint-Louis, 2012, p. 1094.
- [54] Merriam-Webster, «Apophenia, Merriam-Webster,» 2014. [En ligne]. Available: <http://www.merriam-webster.com/dictionary/apophenia>. [Accès le 03 08 2014].
- [55] J. Leskovec, J. Kleinberg et C. Faloutsos, «Graph evolution: Densification and shrinking diameters,» *ACM Transactions on Knowledge Discovery from Data*, vol. 1, n° 1, p. Article 2, 2007.
- [56] M. Leijten et L. Van Waes, «Inputlog features,» 2014. [En ligne]. Available: http://www.inputlog.net/description_features.html. [Accès le 25 07 2014].
- [57] J.-L. Lebrave et A. Grésillon, «Linguistique et génétique des textes : un décalogue,» 23 03 2009. [En ligne]. Available: <http://item.ens.fr/index.php?id=434571>. [Accès le 14 12 2013].
- [58] J.-L. Lebrave, «Comment écriront-ils?,» *Diogène*, vol. 196, n° 14, pp. 163-171, 2001.
- [59] C. Leblay et G. Caporossi, «Introduction aux données temporelles de l'écriture,» chez *Temps de l'écriture: enregistrements et représentations*, C. Leblay et G. Caporossi, Éd., Louvain-la-Neuve, Academia, 2014, pp. 5-15.
- [60] C. Leblay, «En deçà du bien et du mal écrire,» 14 06 2012. [En ligne]. Available: <http://www.item.ens.fr/index.php?id=578258>. [Accès le 07 07 2014].
- [61] M. M. A. Latif, «A State-of-the-Art Review of the Real-Time Computer-Aided Study of the Writing Process,» *International Journal of English Studies*, vol. 8, n° 1, pp. 29-50, 2008.
- [62] P. Kollberg, «S-notation as a tool for analysing the episodic nature of revisions,» Barcelona, 1996.
- [63] R. Karp, «Optimization problems related to internet congestion control,» chez *Graph theory, combinatorics and algorithms*, M. C. Golumbic et I. B. Hartman, Éd., Springer, 2005, pp. 1-16.
- [64] G. Kanawaty, Introduction à l'étude du travail 3e édition, Bureau International du Travail éd., Genève, 1996.

- [65] S. Helbing et S. Baliatti, «From social data mining to forecasting socio-economic crises,» *The European Physical Journal Special Topics*, vol. 195, pp. 3-68, 2011.
- [66] C. Haas, «How the Writing Medium Shapes the Writing Process : Effects of Word Processing on Planning,» *Research in the Teaching of English*, vol. 23, n° 12, pp. 181-207, 1989.
- [67] D. Foucambert et J. Foucambert, «Gestes d'écriture et caractéristiques linguistiques des textes achevés,» chez *Temps de l'écriture: enregistrements et représentations*, C. Leblay et G. Caporossi, Éd., Louvain-la-Neuve, Academia, 2014, pp. 43-70.
- [68] C. Doquet, «Pour une approche linguistique de l'écriture enregistrée,» chez *Temps de l'écriture: enregistrements et représentations*, Academia, Éd., LouvainLa-Neuve, 2014, pp. 21-42.
- [69] M. Dehmer et S. C. Basak, *Statistical and machine learning approaches for network analysis*, Somerset: John Wiley & Sons, 2012.
- [70] C. De Looze, *Analyse et Interprétation de l'Empan Temporel des Variations Prosodiques en Français et en Anglais*, Marseille: Université de Provence, 2010.
- [71] M. Cox, C. Ortmeier-Hopper et K. E. Tirabassi, «Teaching Writing for the "Real World": Community and Workplace Writing,» *The English Journal*, vol. 98, n° 15, pp. 72-80, 2009.
- [72] J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein et C. Welton, «MAD skills: new analysis practices for big data,» *Proceedings of the VLDB Endowment* 2.2, pp. 1481-1492, 2009.
- [73] D. Chakrabarti et C. Faloutsos, *Graph mining: laws, tools, and case studies*, Morgan & Claypool Publishers series, 2012.
- [74] A. Carbone et M. Gromov, «Mathematical sclices of molecular biology,» *Gazette des mathématiciens, édition spéciale*, vol. 88, pp. 11-80, 2001.
- [75] I. Breetvelt, H. Van Den Bergh et G. Rijlaarsdam, «Relations between Writing Processes and Text Quality : When and How,» *Cognition and Instruction*, vol. 12, n° 12, pp. 103-123, 1994.
- [76] M. Bramer, *Principles of data mining*, 2nd éd., Portsmouth: Springer, 2013.
- [77] M. R. Berthold, «Bisociative knowledge discovery,» chez *Advances in intelligent data analysis X*, J. Gama, E. Bradley et J. Hollmen, Éd., Porto, Springer, 2011, pp. 1-7.
- [78] A.-L. Barabasi, R. Albert et H. Jeong, «Scale-free characteristics of random networks: the topology of the world-wide web,» *Physica A*, vol. 281, pp. 69-77, 2000.
- [79] D. Alamargot, G. Caporossi, D. Chesnet et C. Ros, «What Makes a Skilled Writer? Working Memory and Audience Awareness During Text Composition,» *Learning and Individual Differences*, vol. 21, pp. 505-516, 2011.
- [80] R. Agrawal, T. Imielinski et A. Swami, «Mining association rules between sets of items in large databases,» New York, 1993.
- [81] L. Flower et J. R. Hayes, «A cognitive process theory of writing,» *College Composition*

and Communication, vol. 32, n° 14, pp. 365-387, 1981.

- [82] P. Kollberg, «Rules for the S-notation: a computer-based method for representing revisions,» IPLab, Royal Institute of Technology (KTH), Stockholm, Sweden, 1996.
- [83] D. Alamargot et J.-L. Lebrave, «The study of professional writing: A joint contribution from cognitive psychology and genetic criticism,» *European Psychologist*, n° 14, doi:10.1027/1016-9040/a000001, 2009.
- [84] T. Olive, J.-L. Lebrave, J.-M. Passerault et N. Le Bigot, «La dimension visuo-spatiale de la production de textes: approches de psychologie cognitive et de critique génétique,» *Langages*, vol. 177, n° 11, pp. 29-55, 2010.
- [85] W. Daelemans, P. Berck et S. Gillis, «Data mining as a Method for Linguistic Analysis: Dutch Diminutives,» *Dutch Diminutives, Folia Linguistica, XXXI/1-2*, pp. 57-75, 1997.
- [86] P. A. Bath, C. Craigs, R. Maheswaran, J. Raymond et P. Willett, «Use of graph theory to identify patterns of deprivation and high morbidity and mortality in public health data sets,» *Journal of the American Medical Informatics Association*, vol. 12, n° 16, pp. 630-641, Nov-Dec 2005.
- [87] A. Markowetz, K. Blasiewicz, C. Montag, C. Switala et E. S. Thomas, «Psycho-Informatics: Big Data shaping modern psychometrics,» *Medical Hypotheses*, vol. 82, n° 14, pp. 405-411, April 2014.
- [88] F. J. Ohlhorst, *Big Data Analytics : Turning Big Data into Big Money*, Wiley, 2013, p. 176.
- [89] E. Lindgren et K. P. Sullivan, «Writing and the Analysis of Revision: An Overview,» chez *Computer Keystroke Logging and Writing*, Elsevier, 2006, pp. 31-44.
- [90] ITEM, «Enjeux de recherche,» 7 mai 2014. [En ligne]. Available: <http://www.item.ens.fr/index.php?identifiant=item>. [Accès le 14 juin 2014].
- [91] I. Breetvelt, H. v. d. Bergh et G. Rijlaarsdam, «Relations between writing processes and text quality: When and how?,» *Cognition and Instruction*, vol. 12, n° 12, pp. 103-123, 1994.
- [92] C. Lynch, «How do your data grow?,» *Nature*, vol. 455, n° 14, pp. 28-29, 2008.
- [93] E. Midgette, P. Haria et C. MacArthur, «The effects of content and audience awareness goals for revision on the persuasive essays of fifth- and eighth-grade students,» *Reading and Writing*, vol. 21, pp. 131-151, 2008.
- [94] E. W. Bethel, P. Prabhat, S. Byna, O. Rubel, K. J. Wu et M. Wehner, «Why high performance visual data analytics is both relevant and difficult,» Burlingame, California, USA, 2013.
- [95] F. A. Abukhodair, B. E. Riecke, H. I. Erhan et C. D. Shaw, «Does interactive animation control improve exploratory data analysis of animated trend visualization?,» Burlingame, California, USA, 2013.
- [96] L. Bartram, «Perceptual and interpretative properties of motion for information visualization,» *Proceedings of the 1997 workshop on new paradigms in information visualization and manipulation*, pp. 3-7, 1997.

- [97] D. Alamargot, G. Caporossi, D. Chesnet et C. Ros, «What makes a skilled writer? Working memory and audience awareness during text composition,» *Learning and Individual Differences*, vol. 21, n° 15, pp. 505-516, 2011.
- [98] V. M. Baaijen, D. Galbraith et K. d. Glopper, «Keystroke analysis: reflections on procedures and measures,» *Written Communication*, vol. 29, n° 13, pp. 246-277, 2012.
- [99] A. Kirk, Data visualization: a successful design process [electronic book], Packt Pub., 2012.
- [100] A. Wengelin, «Temps et pauses dans l'écriture au clavier,» chez *Temps de l'écriture: enregistrements et représentations*, Louvain-la-Neuve, Academia, 2014, pp. 97-124.
- [101] W. Aigner, S. Miksch, H. Schumann et C. Tominski, «Visualization of Time-Oriented Data,» Springer, London, 2011.
- [102] D. Perrin et S. Laemmel, «Application à l'écriture journalistique,» chez *Temps de l'écriture, enregistrements et représentations*, Louvain-la-Neuve, Academia-L'Harmattan s.a., 2014, pp. 171-192.
- [103] A. Unwin, C.-h. Chen et W. K. Hardle, «Introduction,» chez *Handbook of Data Visualization*, Berlin, Springer, 2008, pp. 3-14.
- [104] A. Vathy-Fogarassy et J. Abonyi, Graph-Based clustering and data visualization, Springer, 2013.
- [105] D. Chesnet et D. Alamargot, «Eye and Pen 2 manuel de l'utilisateur,» Poitiers, 2011.
- [106] E. R. Tufte, The visual display of quantitative information, Second éd., Cheshire, Connecticut: Graphics Press, 2001.

¹ Deux graphes, ou sous-graphes, sont dits *isomorphiques* lorsqu'ils représentent clairement des diagrammes semblables [26, p. 4]. De manière plus formelle, un graphe $G_1 = (V_1, E_1)$ est isomorphe à un graphe $G_2 = (V_2, E_2)$ s'il y a une correspondance entre les ensembles de sommets V_1 et V_2 et entre les ensembles d'arêtes E_1 et E_2 de telle manière que si e_1 est une arête joignant les sommets u_1 et v_1 , dans le graphe G_1 , alors l'arête correspondante e_2 dans G_2 joint les sommets u_2 et v_2 , ce qui correspond à u_1 et v_1 respectivement [29]. Les algorithmes pour la recherche de structures semblables dans les graphes se basent donc sur cette propriété [73].

² L'équipe TERS (Temps de l'Écriture et ses Représentations) est une équipe de l'Institut des Textes et Manuscrits Modernes (ITEM/ENS-CNRS).