! "#$%&' (#) *+, #*-.) ' (/ ' &(/ +" " 0' &(&12-3*) 2#$ &(320*2#-*0' &(3#2(/ ' &(+) *-$&(/ ' (4-&) #$5#*-+" ((

HŽlène-Sarah BŽcotte-Boutin[a,!,2,3], **Gilles Caporossi**[2,3,4], **Alain Hertz**[1,3] **et Christophe Leblay**[2,5]

[1] **Polytechnique MontrŽal, 2900, boul. ƒdouard-Montpetit Campus de l'UniversitŽ de MontrŽal 2500, chemin de Polytechnique, MontrŽal (QuŽbec) H3T 1J4, Canada**
[2] **Institut des textes et manuscrits modernes (ITEM), CNRS : ƒcole normale supŽrieure [ENS] 45, rue d'Ulm 75005 Paris, France**
[3] **GERAD, HEC MontrŽal 3000, chemin de la C™te-Sainte-Catherine, MontrŽal (QuŽbec) H3T 2A7, Canada**
[4] **HEC MontrŽal, 3000 chemin de la C™te-Sainte-Catherine MontrŽal (QuŽbec) Canada H3T 2A7, Canada**
[5] **University of Turku, 20014 Turun yliopisto, Finlande**

60&) , 0. Plusieurs mŽthodes pour analyser le processus d'Žcriture ont ŽtŽ utilisŽs afin de comprendre les stratŽgies des scripteurs. L'outil principal pour analyser le processus d'Žcriture est le fichier log, qui contient de fa•on exhaustive et dŽtaillŽe l'ensemble des opŽrations effectuŽes par le scripteur lors de la rŽdaction d'un texte. Les donnŽes qui y sont emmagasinŽes sont de quantitŽ considŽrable et lorsqu'elles ne sont pas prŽalablement traitŽes, elles sont hostiles ^ •tre analysŽes par l'humain. Parmi les outils d'analyse utilisŽs, les reprŽsentations du processus d'Žcriture permettent l'agrŽgation des donnŽes grâce ^ un prŽ-traitement. Les structures sous-jacentes des donnŽes ainsi reprŽsentŽes sont gŽnŽralement plus propices ^ l'analyse que les donnŽes brutes. Cet article vise ^ dŽmontrer diffŽrentes mŽthodes d'analyse automatique pouvant •tre appliquŽes ^ ces structures afin de trouver ou confirmer des structures et tendances ^ travers les donnŽes.

! 7&*2#1*8(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.(

" **Auteur de correspondance : helene.becotte@polymtl.ca**

9(:"*2+/) 1*-+"(

Les techniques dôanalyse des textes produits foisonnent et de nouvelles disciplines sont enti•rement crŽes autour de ce concept. Les techniques utilisŽes sont variŽes et proviennent de plusieurs domaines [1]. Les statistiques**, les** tests A/B**, la** fusion **et** int□gration de donn□es**, la** r□gression**, la** classification**, les** r□gles d□association**, le** data mining**, la** visualisation**, le** text mining**,** l□analyse de sentiments**, les** r□seaux de neurones**,** lôanalyse de r□seaux**, le** traitement du langage naturel**, la** simulation**,** lôanalyse de s□ries chronologiques **sont toutes des techniques ou groupes de techniques qui sôentrecroisent [2].**

Bien que ces techniques soient utilisŽes sur des textes produits et des donnŽes statiques et que le processus dôŽcriture soit pour sa part composŽ de donnŽes dynamiques multidimensionnelles, il est possible de sôen inspirer pour bŽnŽficier des possibilitŽs de telles techniques. Côest ce que nous nous proposons de faire dans ce travail, en nous concentrant sur la visualisation.

LôŽtude du processus dôŽcriture est complexe pour plusieurs raisons. La toute premi•re raison est due ^ la singularitŽ du processus individuel qui fait en sorte quôil nôexiste aucune mŽthode unique pour rŽdiger un texte de qualitŽ [3]. La seconde est quôil est composŽ principalement de donnŽes qui varient dans le temps. Ainsi, les opŽrations dôŽcriture, telles quôelles ont ŽtŽ dŽcrites par la gŽnŽtique textuelle dans un environnement numŽrique [4], peuvent •tre listŽes en fonction de leur ordre chronologique dôapparition. Ces opŽrations et les pauses qui y sont attachŽes, sont difficiles ^ interprŽter par lôhumain lorsquôelles ne sont pas combinŽes et contextualisŽes pour former des unitŽs identifiables par lôhomme : des mots, des phrases ou une section de ce qui deviendra du texte.

Parmi la panoplie disponible des outils dôanalyse disponibles, les reprŽsentations du processus scriptural sont les plus utilisŽes gr‰ce ^ leur facilitŽ ^ traiter lôinformation.

Avec lôav•nement de technologies dŽdiŽes ^ lôanalyse du processus depuis les annŽes 90, les possibilitŽs dôautomatisation se sont multipliŽes. MalgrŽ cela, peu de mŽthodes permettent aujourdôhui aux chercheurs dôaccŽlŽrer leur analyse et de se concentrer sur lôinterprŽtation et la comprŽhension des mŽcanismes de rŽvisions qui pourraient permettre ^ un scripteur dôamŽliorer la qualitŽ de ses textes.

Dans ce travail, nous nous proposons de nous intŽresser, tout dôabord, aux structures sous-jacentes aux donnŽes impliquŽes dans le processus scriptural en distinguant nettement entre les reprŽsentations rŽalisŽes dans le cadre de la linguistique cognitive et celles rŽalisŽes dans le cadre de la linguistique gŽnŽtique. Ce qui nous permettra dôintroduire ^ la mŽthodologie suivie et dôexposer les modalitŽs dôenregistrement. Nous finirons par illustrer notre propos ^ lôaide dôun exemple tirŽ de notre corpus afin de prŽciser les objectifs poursuivis.

; (<20*2#-*' , ' "*(=(&*2) 1*) 2' &(/ ' &(/ +"" 0' &(&12-3*) 2#$ &(

Depuis les annŽes 80, plusieurs mŽthodes pour analyser le processus dôŽcriture ont ŽtŽ utilisŽes [5]. Lôoutil principal pour analyser le processus dôŽcriture est le fichier dôenregistrement, appelŽ log, qui contient de fa•on exhaustive et dŽtaillŽe lôensemble des opŽrations effectuŽes par le scripteur lors de la rŽdaction dôun texte [6]. Les donnŽes qui y sont emmagasinŽes sont considŽrables, et lorsquôelles ne sont pas prŽalablement traitŽes, elles sont hostiles ^ lôanalyse humaine [7]. Ce traitement prŽalable nous semble donc dŽterminant dans lôacc•s aux donnŽes scripturales enregistrŽes. Autrement dit, il est impossible dôexploiter les programmes dôenregistrement de lôŽcriture (dits aussi de temps r□el**), sans procŽder, au prŽalable, ^ un redŽploiement visuel des donnŽes obtenues. Les structures sous-jacentes des donnŽes ainsi reprŽsentŽes sont gŽnŽralement plus propices ^ lôanalyse que les donnŽes brutes inexploitables. La principale raison est que lorsquôon cherche ^ explorer des donnŽes, plusieurs Žtapes sont nŽcessaires avant de pouvoir utiliser les techniques dôanalyse ^ proprement parler. Le processus**

simplifiŽ par lequel les donnŽes sont acquises, traitŽes et analysŽes se rŽsume ainsi : 1) enregistrement**, 2)** partage**, 3)** tri**, 4)** recherche**, 5)** repr□sentation visuelle des donn□es **et 6)** analyse **[2]. Plusieurs de ces actions ont dŽj^ ŽtŽ Žtudižes de mani•re individuelle et ont ŽtŽ lÔobjet de crŽation de logiciels visant ^ enregistrer le processus dÔŽcriture et ^ le reprŽsenter dans lÔune ou lÔautre de ses dimensions [8]. Ces logiciels sont tous diffŽrents et gŽnŽralement destinŽs ^ un projet de recherche particulier [6]. Certains sont con•us pour traiter lÔinformation et la retransmettent ensuite ^ lÔutilisateur de fa•on simplifiže, sous la forme dÔune reprŽsentation visuelle.**

LÔutilisation de reprŽsentations graphiques de donnŽes, appelŽes de fa•on plus gŽnŽrale visualisations**, consiste ^ explorer et essayer de comprendre les grands ensembles de donnŽes [9, p. xvi]. Elles permettent notamment dÔidentifier des tendances, structures, irrŽgularitŽs et relations entre les donnŽes sur une certaine pŽriode temporelle [10, p. 110]. Le format compact de la visualisation aggr•ge les donnŽes et utilise les capacitŽs cognitives de lÔhumain [11]. Le but principal de lÔutilisation de ces images reprŽsentant le processus scriptural est de laisser lÔ Ï il trouver des structures sous-jacentes parmis les donnŽes [12].**

Il est important cependant de considŽrer que lÔutilisation de reprŽsentations visuelles des donnŽes comporte certains dŽfauts. Puisque les conclusions dÔanalyse seront effectuŽes par lÔutilisateur, les capacitŽs de traitement de celui-ci influencent fortement les rŽsultats obtenus [13]. Tout dÔabord, lÔ•tre humain a une capacitŽ limitže ^ traiter des quantitŽs dÔinformations et il aurait aussi parfois tendance ^ percevoir des tendances et associations alors quÔil nÔy en a pas [1]. Pour cette raison et pour maximiser le potentiel de conclusions possibles, il est pertinent de pouvoir confirmer mathŽmatiquement ces conclusions gr‰ce ^ des mod•les dÔanalyse de donnŽes et de sÔassurer quÔil existe effectivement une corrŽlation forte entre diffŽrentes variables [14].

Les techniques dÔanalyse utilisŽes sont ensuite dŽpendantes du type de prŽ-traitement effectuŽ sur les donnŽes. Structurer les donnŽes dÔune certaine mani•re plut™t quÔune autre influence donc nŽcessairement les rŽsultats dÔanalyse obtenus par la suite. ConsŽquement, lors du choix dÔune technique particuli•re dÔanalyse, ce nÔest pas la technique elle-m•me qui est importante mais lÔutilisation finale des rŽsultats obtenus quÔil faut considŽrer [15, p. 134]. Dans le but dÔobtenir par exemple une comprŽhension du syst•me sous-jacent aux donnŽes comme dans le cas de lÔanalyse du temps de lÔŽcriture, il est important de sÔassurer que le mod•le utilisŽ et les rŽsultats obtenus soient propice ^ lÔinterprŽtation [16, p. 209].

Il existe deux grands mod•les de reprŽsentations qui prŽtraitent les donnŽes de mani•re ^ pouvoir appliquer des techniques dÔanalyse linguistique de donnŽes. Le tout premier, attachŽ aux travaux rŽalisŽs par une linguistique cognitive, est celui qui a ŽtŽ dŽveloppŽ dans le cadre des Syst□mes d□Information G□ographique **(dorŽnavant SIG). Le second, en rŽaction au premier, a ŽtŽ dŽveloppŽ en associant, de mani•re pluridisciplinaire, la linguistique gŽnŽtique et la thŽorie mathŽmatique des graphes. A ces deux mod•les, nous proposerons, dans ce travail, dÔintroduire un tout nouveau mod•le, qui est en fait le mixte des deux grands mod•les : la** repr□sentation progressive**.**

; 89(>-"?) -&*-.) ' (1+?"-*-4' (' *(&%&*@ ' &(/ A"B-2, #*-+"(?0+?2#3C-.) ' &(DE:F G;

Les SIG sont un type de reprŽsentation utilisŽ initialement pour traiter des donnŽes gŽographiques. Les principaux attributs sont lÔidentification spatiale et temporelle dÔactivitŽs, qui ont une ressemblance avec les donnŽes issues de lÔŽcriture [17]. Ces attributs sont utilisŽs en tant que coordonnŽes. En gŽographie, ces coordonnŽes sont souvent la latitude et la longitude [18, p. 1] mais peuvent Žgalement contenir lÔŽlŽvation [18, p. 73]. En analyse du processus,

dans la majoritŽ des cas, ces coordonnŽes vont •tre composŽes dÔune mesure de temps et dÔune quantitŽ de rŽvisions effectuŽes [17].

Voici un exemple dÔune reprŽsentation effectuŽe sur un logiciel dŽdiŽ pour les SIG [19] :

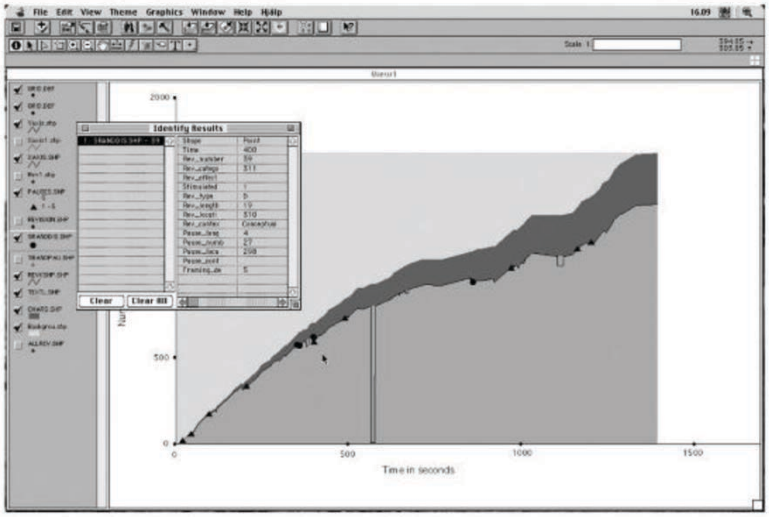

Fig. 1. **ReprŽsentation SIG.**

Notre propos nÔŽtant pas de rentrer dans lÔanalyse de ces reprŽsentations, nous nous contenterons, dans le cadre de ce travail, dÔen noter les principales caractŽristques. Ainsi, comme il est possible de le voir dans la figure 1, la contribution clŽ des SIG rŽside dans la structure quÔils donnent aux donnŽes et aux possibilitŽs de les analyser considŽrant lÔemplacement physique des donnŽes dans le texte. Cette dimension du processus dÔŽcriture se nomme spatialit**. Des mŽthodes dÔanalyse spŽcifiques sont adaptŽes ^ ces structures de donnŽes de mani•re ^ mettre en relief certaines problŽmatiques. Les principales techniques comprennent (1) lÔanalyse des** chevauchements entre diff rentes caract ristiques**, (2)** l identification de sous-groupes**, (3) lÔexploration de** tendances spatiales selon certaines variables **et 4) lÔexploration de la** relation entre plusieurs variables en fonction de leur emplacement spatial **[20, p. 2].**

Ont utilisŽ ce mode de reprŽsentation (SIG) principalement les Žtudes suivantes : LS Graph **[21, 17],** Gen se du texte **[4],** GIS Graph **[19],** Timeline **[22] et** Inputlog **[23]. Toutes ces reprŽsentations ont en commun lÔutilisation des axes X et Y, comme la figure 1 lÔillustre, chacune exploitant des caractŽristiques diffŽrentes : chronologie, topologie, avec un jeu de couleurs.**

Les mŽthodes les plus pertinentes dans le cas de lÔanalyse du processus de lÔŽcriture sont sans contredit celles qui permettent de mettre de lÔavant lÔaspect spatial des donnŽes. La visualisation est considŽrŽe comme Žtant la premi•re Žtape pour nÔimporte quelle analyse spatiale [20, p. 19 ch.2]. Par la suite, la requ•te de donnŽes est la mŽthode la plus simple pour dŽbuter lÔanalyse. Il sÔagit dÔune fa•on dÔagr ger les donnŽes ^ un niveau supŽrieur pour obtenir des informations concentrŽes sur une problŽmatique particuli•re. Les statistiques sont lÔŽtape suivante pour comprendre les caractŽristiques des donnŽes.

Finalement, plusieurs sources dÔerreur sont possibles et un choix de modŽlisation de donnŽes inappropriŽ est aussi considŽrŽ critique dans le cas des SIG [20, p. 18 ch.2].

; 8 (>-" ?) -&*-.) ' (?0" 0*-.) ' (' *(2' 320&' "*#*-+" (3#2(?2#3C' &(

La reprŽsentation par les graphes permet de mettre davantage en relief lÔaspect dynamique de lÔŽcriture. Celle-ci est orientŽe sur la chronologie du processus dÔŽcriture [24]. LÔune des particularitŽs est de bien gŽrer la transformation et les mouvements du texte grâce aux nŒuds reprŽsentant des parties de texte (ajoutŽs, supprimŽs) reliŽs ensemble par des arcs dŽfinissant leur relation, soit chronologique ou spatiale [25]. Il est possible de voir le contenu des nŒuds. Telle que mentionnŽe par Caporossi et Leblay [8], cette reprŽsentation montre nŽanmoins lÔaspect temporel de la rŽdaction, soit le moment exact où le scripteur a effectuŽ chacune des opŽrations reprŽsentŽes.

Voici lÔexemple dÔune reprŽsentation par graphes produite par Caporossi et Leblay [8] :

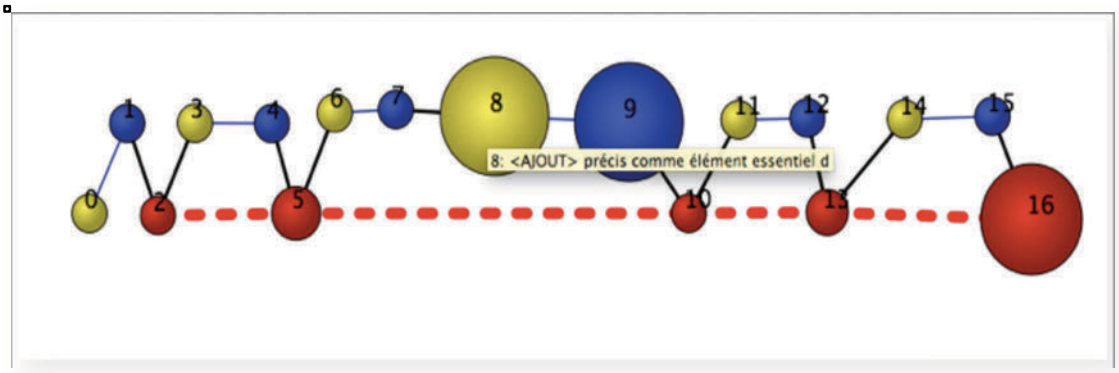

Fig. 2. **ReprŽsentation gŽnŽtique par graphes.**

Cette representation qui reprend une description gŽnŽtique ^ partir des opŽrations ŽlŽmentaires dÔŽcriture (Ajout versus **Suppression) doit •tre lue de la mani•re suivante:**

(0): Il me

(1): <SUPPR> em Il

(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> prŽcis comme ŽlŽment essentiel d

(9): <SUPPR> d leitnesse tnemŽlŽ emmoc sicŽrp

(10): <AJOUT> ou

(11: <AJOUT> en end

(12): <SUPPR> dne ne

(13): <AJOUT> un endroit pr

(14): <AJOUT> +ec

(15): <SUPPR> ce+

(16): <AJOUT> Žcis comme ŽlŽment essentiel d

Ce mode de reprŽsentation n'oublie pas lÔimportance du texte produit : il est possible de le suivre au moyen des traits pointillŽs rouges. Il est facile de noter tout le travail dÔŽcriture rŽalisŽ par le scripteur, soit les 2/3 de ce qui est Žcrit (11 n Ïuds sur 16) nÔapparaitra pas dans le texte produit.

Mais, au delˆ de la simple apprŽhension de cette reprŽsentation mathŽmatique, et afin de bien en comprendre les subtilitŽs, nous proposons un succinct rappel des ressources thŽoriques des graphes.

Un graphe est une reprŽsentation visuelle simple constituŽe de points reliŽs ensemble par des lignes [26, p. 1]. Leur structure est utilisŽe dans des domaines variŽs puisquÔelle reprŽsente simplement des relations entre objets **[27, p. 1]. La terminologie est plut™t intuitive. Les points sont appelŽs** n˜uds **tandis que les lignes sont appelŽes** ar□tes**. Les n Ïuds peuvent reprŽsenter des villes sur une carte, des individus dans un rŽseau social ou encore simplement des informations reliŽes entre elles par les ar•tes [28, p. 2]. Deux n Ïuds sont dits** adjacents **ou** voisins **sÔils sont reliŽs par une ar•te. Finalement, le degrŽ dÔun n Ïud est le nombre de voisins que poss•de celui-ci [29, p. 4]. De mani•re plus formelle, un graphe** ! " #$%&' **est constituŽ dÔun ensemble fini** $ **de n Ïuds et dÔun ensemble** (& **dÔar•tes. Un graphe peut •tre** orient□**, dans lequel cas** ! " #$%)' **a un ensemble**) **dÔarcs. Un arc qui relie un sommet** * **ˆ un sommet** + **sera notŽ** [*%+] **tandis quÔun arc qui relie un sommet** + **ˆ un sommet** * **sera notŽ** [+%*]**.**

De mani•re appliquŽe, en prenant lÔexemple des rŽseaux sociaux, un graphe non orientŽ serait utilisŽ pour dŽcrire une plateforme telle que Facebook, **pour dŽcrire la relation mutuelle que deux individus partagent en Žtant** amis**. Un graphe orientŽ plut™t serait utile dans un cas comme** Twitter **o• la relation nÔest pas nŽcessairement rŽciproque. Un usager peut tr•s bien** suivre **un autre usager, sans que lÔinverse soit vrai. SÔil y a une relation de prŽcŽdence entre des actions, lˆ encore lÔorientation des arcs peut •tre pertinente [2, p. 29].**

LÔanalyse de rŽseaux, que certains auteurs appellent plut™t graph mining **[30], est relative ˆ la thŽorie des graphes et est une technique dÔanalyse de donnŽes utilisŽe en** big data **[2, p. 29]. Cette technique permet en effet dÔutiliser de bons algorithmes pour lÔanalyse et le traitement des donnŽes complexes pouvant •tre reprŽsentŽes sous forme de graphes [31]. Depuis une dizaine dÔannŽes, le dŽveloppement dÔalgorithmes de** data mining **dŽdiŽs aux graphes a ŽtŽ source dÔune augmentation dÔintŽr•t de la part des milieux scientifiques et industriels [32]. Parmi les techniques adaptŽes les plus frŽquentes, on retrouve la** classification **et lÔ**identification de tendances **[30]. Nous nous sommes donc posŽs la question de lÔadaptabilitŽ de cette thŽorie mathŽmatique aux sciences humaines, et plus particuli•rement ˆ la linguistique de lÔŽcrit.**

Ainsi, les possibilitŽs dÔanalyse de ces structures ont dŽjˆ ŽtŽ ŽtudiŽes dans le contexte du processus de lÔŽcriture. La mise en Žvidence de sous-graphes particuliers reprŽsentant les patterns **des opŽrations les plus frŽquentes (ajout vs insertion, suppressions immŽdiate et diffŽrŽe, remplacement), a dŽjˆ ŽtŽ rŽalisŽe. LÔidentification de ceux-ci est utile dans lÔanalyse du graphe reprŽsentant le temps de lÔŽcriture [7]. Il serait aussi possible dÔavoir une visŽe plus exploratoire pour lÔanalyse et chercher par exemple des tendances ou autres sous-structures dans les donnŽes.**

La recherche de ce type de structures peut se faire sur un seul large graphe pour y trouver des structures qui sont rŽcurrente un nombre minimal de fois. Elle peut Žgalement se faire sur un groupe de graphes, pour lequel lÔobjectif est de trouver des structures qui sont r□currentes **dans une certaine fraction des graphes faisant partie de ce groupe [30]. En raison de la diffŽrence fondamentale entre les caractŽristiques sous-jacentes aux donnŽes et ˆ la formulation du probl•me, les algorithmes qui sont dŽveloppŽs pour lÔanalyse dÔun groupe de graphes ne**

peuvent •tre appliquŽs ^ lÔanalyse des caractŽristiques dÔun seul graphe. Inversement, les algorithmes dŽveloppŽs pour lÔanalyse ^ lÔintŽrieur dÔun large graphe peuvent •tre facilement adaptŽs pour lÔanalyse dÔun ensemble de graphes [32].

Dans lÔanalyse du temps de lÔŽcriture, les deux approches sont Žgalement pertinentes, pour dÔabord repŽrer les stratégies récurrentes **dÔun scripteur, et pour ensuite comparer les stratŽgies des diffŽrents scripteurs entre eux. Les mŽthodes dÔanalyse seront appliquŽes sur les diffŽrents niveaux de graphes pour tenter de dŽtecter des stratŽgies ou opŽrations nouvelles.**

LÔutilisation de ces propriŽtŽs en analyse de lÔŽcriture permettrait ainsi de comparer plusieurs textes dÔun m•me auteur pour consolider et mieux comprendre ses propres stratŽgies de rŽdactions sous forme de sous-ensembles de donnŽes. Il serait Žgalement possible de prendre plusieurs textes dÔun ensemble de scripteurs ayant un niveau de ma"trise dÔŽcriture semblable et ainsi de tenter de dŽcouvrir les rŽcurrences entre des stratŽgies individuelles de rŽdaction.

H(I +23) &(' *(, 0*C+/ +$-?-' (

Nous nous proposons maintentant de prŽciser le cadre mŽthodologique dans lequel a ŽtŽ enregistrŽ notre corpus.

H89((((J 0*C+/ +$-?-' (

Ce projet est centrŽ sur lÔŽcriture, comme processus descriptible linguistiquement, sÔinscrivant dans le temps et lÔespace, cÔest-^-dire dans lÔavant dÔun texte finalement stabilisŽ. Ce choix sÔappuie sur le constat mŽthodologique, selon lequel le produit nÔest pas lÔimage fid•le de la production. En adoptant une approche gŽnŽtique du texte, associŽe ^ un logiciel dÔenregistrement de lÔŽcriture, nous cherchons ^ dŽcrire comment divers scripteurs donnent lieu ^ diverses rŽdactions et rŽvisions, invisibles dans le texte achevŽ. Trois orientations Žmergent 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 dŽmontrer que la temporalitŽ est une dimension essentielle de l'Žcriture, tout autant que la spatialitŽ. LÔŽtude, bien trop souvent nŽgligŽe de cette temporalitŽ, est intimement liŽe aux modes de reprŽsentations et de visualisations numŽriques. Ce que nous nous proposons dÔŽtudier, afin de rŽpondre ^ une interrogation contemporaine ŽlŽmentaire: comment donner ^ voir ce qui se passe prŽcisŽment sur un Žcran, pendant le temps de lÔŽcriture. LÔŽquipe[2] a dŽveloppŽ une mŽthodologie propre, ^ partir dÔun outil mathŽmatique (graphes) adŽquat, qui permet, en parall•le aux mod•les de reprŽsentations (SIG) dŽj^ consacrŽs par la linguistique cognitive, une vraie cartographie gŽnŽtique de la dynamique scripturale. Ce qui a nŽcessitŽ la mise en place dÔun logiciel dŽdiŽ proche des attentes de la linguistique gŽnŽtique.**

H8; ((((K" (" +) 4' #) (32+?2#, , ' (/ A" 2' ?-&*2' , ' " *(

Le programme GenographiX **(GGX), a ŽtŽ mis en place spŽcifiquemet pour mettre en Švidence le travail gŽnŽtique de rŽŽcriture des scripteurs. Voici lÔenvironnement graphique offert par le programme :**

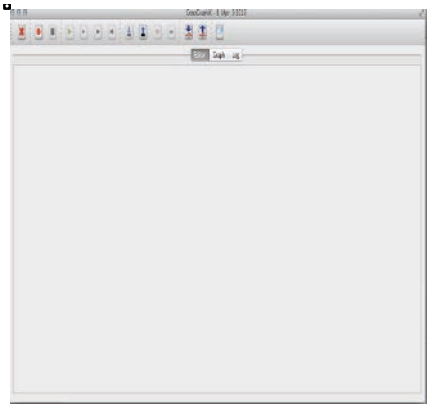

Fig. 3. **Le programme** GenographiX.

Comme la figure 3 le montre, le programme GenoGraphiX **est expŽrimental, fonctionnant uniquement, pour lÔinstant du moins, dans un environnement fermŽ. Le logiciel permet dÔenregistrer le dŽtail de lÔactivitŽ du scripteur, et de reprŽsenter cette activitŽ par lÔentremise dÔun graphe mathŽmatique tel quÔil a ŽtŽ dŽcrit prŽcŽdemment. Les donnŽes brutes (**log**) sont enregistrŽes dans un format similaire ˆ un programme de type** Scriptlog. **La liste des ŽvŽnements est alors analysŽe par le logiciel qui regroupe ces ŽvŽnements en** cellules **(ou** n˘uds**, ou** ar□tes**) et identifie les** liens **(ou** arcs**) qui peuvent exister entre les cellules. Dans la version courante du logiciel, chaque cellule est reprŽsentŽe par un rectangle dans lequel on peut lire le texte correspondant (les versions antŽrieures reprŽsentaient chaque cellule par un cercle, mais le texte correspondant nÔŽtait pas visible). Ces cellules peuvent •tre de trois natures : 1) du texte ajoutŽ que le scripteur a supprimŽ avant la fin de la rŽdaction (en jaune), 2) du texte ajoutŽ qui restera dans le document final (en rouge) et 3) du texte supprimŽ (en bleu). Des liens entre cellules reprŽsentent une relation temporelle ou spatiale. Comme nous lÔavons dŽjˆ signalŽ, pour des raisons mŽthodologiques, une attention particuli•re est portŽe sur le texte final qui peut toujours •tre lu en suivant les cellules et les liens rouges. Il est toutefois aussi possible de suivre le travail de mani•re chronologique en suivant les liens en trait continu (les autres liens sont en pointillŽs). Afin de marquer la relation particuli•re quÔil y a entre lÔajout dÔune portion de texte et sa suppression ultŽrieure, la cellule dÔajout et la cellule de suppression correspondant sont reliŽes par un lien vert. Un exemple de graphe est reprŽsentŽ par la figure 5.**

H8-H(K"' ' (/ +) 7$ (*L1C' (

Il sÔagit dÔun corpus numŽrique dÔŽcritures enregistrŽes, composŽ de deux t‰ches successives. La premi•re, de nature narrative, se prŽsente comme suit :
Ç Racontez une histoire. Vous devez int□grer 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 t‰che, de nature argumentative, est la suivante :
Ç Qu□est-ce qui est important pour vivre ensemble ?È**. Un sujet qui se pr•te ˆ tous les ‰ges, sans reposer sur une structure de type pour/contre reprŽsentative dÔun mod•le dominant des textes argumentatifs.**

PrŽcisons que la premi•re reprend un protocole qui a dŽjˆ fait lÔobjet de recueil de corpus de textes aupr•s dÔenfants et de futurs enseignants en formation, proposŽ au dŽpart par M. Charolles dans le cadre dÔŽtudes de la cohŽrence textuelle et repris notamment dans Garcia-

Debanc et Bonnemaison [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.**

M&K"(' "*2' N') OF$#(2' 320&' "*#*-+"(, -O*' (

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 :

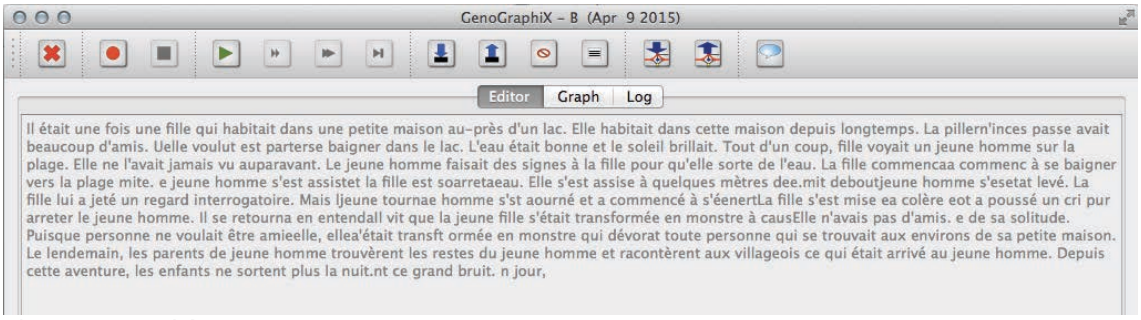

Fig. 4. **Texte produit.**

Comme il est possible de le voir dans lÕŽditeur de GGX, le texte produit appara”t sous le format txt**. Voyons maintenant les 20 premi•res secondes de cet enregistrement, sous la forme dÔun graphe :**

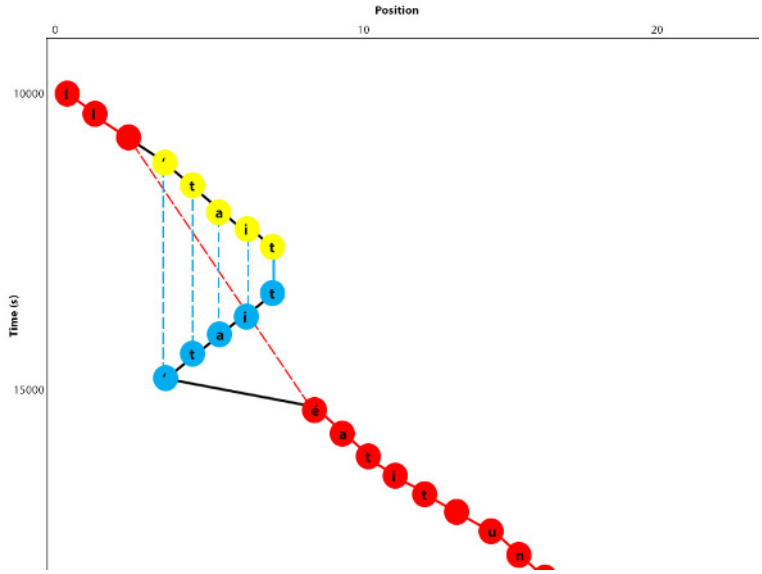

Fig. 5. **ReprŽsentation mixte, dite** progressive.

Ce graphe est mixte **parce quÕil reprend une reprŽsentation inspirŽe des axes X (position spatiale) et Y (chronologie). Dans ce mod•le mixte, il faut noter que les trois premi•res opŽrations effectuŽes, soit lÔinsertion dÔun L, e et dÔun espace sont en rouge. Ceci signifie quÔelles sont prŽsentes dans la version finale du texte. Par la suite, la prŽsence des opŽrations successives Ç Õ tait È en jaune et leur double en bleu qui se situe au m•me emplacement sur lÔaxe des abscisses signifique que ces opŽrations ont dÔabord ŽtŽ effectuŽes (en jaune) mais par la suite supprimŽes (en bleu). Le lien bleu pointillŽ qui relie lÔopŽration jaune avec son double bleu signifie quÔelles occupent le m•me emplacement spatial dans le texte. Au bas de la reprŽsentation il est aisŽ de remarquer les opŽrations successives Ç Žtait une È, qui sont en rouge tout comme le dŽbut et sont dÔailleurs liŽes ^ lÔespace en position 3, ce qui signifie que ces opŽrations sont successives dans la version finale du texte. En lisant successivement les opŽrations en rouge, liŽes chronologiquement par des liens de couleur rouge Žgalement, il est possible de constater que le texte final de cette portion reprŽsentŽe, peut se lire Ç Il Žtait une È, conformŽment ^ ce quÔil est indiquŽ ^ la figure 4.**

Puisque les SIG ont cette particularitŽ de bien mettre en relation lÔaspect temporel et spatial du temps de lÕŽcriture [17], ils ont en contrepartie le dŽsavantage de nÔ•tre quÔune vue dÔensemble du processus. La consŽquence de cet attribut est de rendre difficile pour lÔanalyste de le lier avec opŽrations et le contexte dÔŽcriture [8]. Pour sa part, la reprŽsentation par les graphes met davantage lÔaccent sur le lien entre les opŽrations et leur sŽquence [7]. Il est cependant difficile de conna”tre le contenu exact de chaque n Ï ud et la position temporelle des opŽrations.

Cette reprŽsentation reprend dÔabord lÔidŽe des graphes gŽnŽtiques mais est incluse dans un graphique ayant comme axe des abscisses la position et comme axe des ordonnŽes le temps. Contrairement aux SIG dont le dŽroulement du temps se lit de bas en haut, cette reprŽsentation se lit de gauche ^ droite et de haut en bas, en suivant le mod•le logique dÔun texte [34]. Par

rapport ^ la reprŽsentation par les graphes, dans ce cas-ci, chaque nÏud reprŽsente une opŽration individuelle.

La sous-structure des donnŽes de cette reprŽsentation poss•de enti•rement les attributs des deux reprŽsentations. Dans un graphe, il est possible dôavoir des attributs complŽmentaires, souvent considŽrŽs comme des poids [26, p. 15]. Chaque sommet reprŽsente ainsi une opŽration effectuŽe. Lôopration peut •tre soit une insertion ou une suppression. De fa•on similaire aux fichiers log, chaque opŽration poss•de les attributs temps, position **et** type**. Les arcs poss•dent Žgalement des attributs. NŽcessairement, ils ont une** origine **et une** destination **mais dans ce cas-ci ils ont aussi un** type **[34]. Cette structure permet ainsi dôappliquer les notions mathŽmatiques et mod•les propres ^ la thŽorie des graphes pour tenter de dŽceler des tendances et sous-structures.**

Les attributs des sommets peuvent dans un deuxi•me temps •tre utilisŽs plus simplement en tant que coordonnŽes et reprendre les mŽthodes propres aux SIG.

Q8(l +"1$) &+"((

Au sein des approches int□ress□es par les processus d'□criture-r□□criture sur □cran, les outils num□riques d'□valuation sont en train de devenir un v□ritable objet de recherche, tout autant que les m□diums de productions textuelles. Parmi ces outils d□□valuation, la visualisation de donn□es occupe une place de choix. Se situant directement dans la tradition des transcriptions r□alis□es par la g□n□tique (sch□matiquement diplomatiques versus lin□aires), la visualisation des donn□es scripturales prolonge ces paradigmes en proposant des donn□es dynamiques, et non plus statiques. Il n'est alors plus seulement question d□□valuer par l'image ce qui vient d'□tre □crit, mais bien d'□valuer par l'image ce qui est en train d'□tre □crit, d□□tre r□□crit sur □cran, dans des s□quentialit□s singuli□res.

Le principal avantage dôutiliser des mŽthodes dŽcoulant dôune structure de donnŽes propre ^ un type de reprŽsentation du processus dôŽcriture est de pouvoir ^ la fois bŽnŽficier des capacitŽs cognitives humaines ^ dŽtecter des tendances et rŽcurrences dans des donnŽes visualisŽes et de confirmer ces hypoth•ses avec des mŽthodes mathŽmatiques. Parmis les diffŽrentes reprŽsentations explorŽes, il semble que lôutilisation de la reprŽsentation mixte, dite progressive, permette de structurer les donnŽes en ayant acc•s ^ un bassin plus large de mŽthodes. Il serait tout de m•me intŽressant de comparer ces reprŽsentations et certaines de leurs mŽthodes dôanalyse respectives afin de confirmer leurs avantages et Žlaborer un cadre mŽthodologique pour permettre aux chercheurs dôavoir un acc•s facile ^ ces complŽments dôanalyse.

Dans l□□tat actuel de notre projet, les attentes sont les suivantes: 1) Affiner notre outil d□enregistrement (GGX) afin de pouvoir aussi enregistrer dans des environnements □cologiques comme, par exemple, *Word* ou *Open Office*. Pour l'instant, cet outil reste confidentiel. 2) Mettre en place une base de donn□es, □tendue, compos□e d'heures d'enregistrement. Une telle base de donn□es n'existe pas actuellement en langue fran□aise. 3) Pouvoir rendre accessible cette base de donn□es sur la plate-forme d□di□e (TEMPUS), l'id□e □tant bien de mettre □ disposition des chercheurs des donn□es temporelles vari□es. 4) D□velopper les concepts th□oriques n□cessaires. Et 5) de proc□der □ des descriptions affin□es en litt□ratie d□butante et experte.

(

6 0B92' "1' &((

(

[1] **D. Boyd et K. Crawford, ÇCritical questions for big data,È** 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, ÇBig data: the next frontier for innovation, competition, and productivity,È McKinsey Global Institute, 2011.**

[3] **S. Plane, D. Alamargot et J.-L. Lebrave, ÇTemporalitŽ de l'Žcriture et r™le du texte produit dans l'activitŽ rŽdactionnelle,È** Langages, **vol. 177, n¡ %11, pp. 11-32, 2010.**

[4] **C. Doquet-Lacoste, *f*tude GŽnŽtique de l'*f*criture sur Traitement de Texte d'*f*lŠves de Cours Moyen 2, AnnŽe 1995-1996, Paris: UniversitŽ Sorbonne nouvelle, 2003, p. 633.**

[5] **K. S. Miller et K. P. Sullivan, ÇKeystroke Logging: An introduction,È chez** Computer keystroke logging and writing**, Oxford, Elsevier, 2006, pp. 1-10.**

[6] **K. P. Sullivan et E. Lindgren, ÇLa rŽvision en production Žcrite enregistrŽe,È chez** Temps de l'écriture: enregistrements et représentations**, C. Leblay et G. Caporossi, *f*ds., Louvain-la-Neuve, Academia, 2014, pp. 71-92.**

[7] **G. Caporossi et C. Leblay, ÇOutils de visualisation de donnŽes enregistrŽes,È chez** Temps de l'écriture: enregistrements et représentations**, C. Leblay et G. Caporossi, *f*ds., Louvain-la-Neuve, Academia, 2014, pp. 147-166.**

[8] **G. Caporossi et C. Leblay, ÇOnline Writing Data Representation : A Graph Theory Approach,È 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, **UniversitŽ de Reims Champagne Ardenne , 2005.**

[12] **M. Tory et T. Moller, ÇHuman factors in visualization research,È** IEEE Transactions on visualization and computer graphics, **vol. 10, n¡ %11, pp. 72-84, 2004.**

[13] **G. Dzemyda, O. Kurasova et J. Zilinskas, ÇMultidimensional Data Visualization: Methods and Applications,È Springer, Vilnius, 2013.**

[14] **T. Moller, B. Hamann et R. Russell, ÇPreface,È chez** Mathematical foundation of scientific visualization, computer graphics, and massive data exploration**, Berlin, Springer, 2009, pp. V-VII.**

[15] **H. A. Karimi, ÇBig data :Techniques and technologies in geoinformatics,È Boca Taton: CRC Press, 2014.**

[16] **S. TuffŽry, Data mining et statistique dŽcisionnelle: l'intelligence des donnŽes, ƒditions Technip, 2010, p. 705.**

[17] **E. Lindgren et K. P. Sullivan, ÇThe LS Graph : A Methodology for Visualizing Writing Revision,È** 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, ÇGIS for Writing: Applying Geographical Information Systems Techniques to Data Mine Writings' Cognitive Processes,È chez** Writing and Cognition**, Amsterdam, Elsevier, 2007, pp. 83-96.**

[20] **C. D. Lloyd, Spatial data analysys: an introduction for GIS users, New York: Oxford University Press, 2010.**

[21] **M. Leijten et L. Van Waes, ÇInputlog : New Perspectives on the Logging of On-Line Writing Processes,È chez** Computer Keystroke Logging and Writing**, K. P. S. a. E. Lindgren, ƒd., Elsevier, 2006, pp. 73-94.**

[22] **A. Wengelin, M. Torrance, K. Holmwvist, S. Simpson, D. Galbraith, V. Johansson et R. Johansson, ÇCombined eyetracking and keystroke-logging methods for studying cognitive processes in text production,È** Behavior Research Methods, **vol. 41, n¡ %12, pp. 337-351, 2009.**

[23] **M. Leijten et L. Van Waes, ÇKeystroke Logging in Writing Research : Using Inputlog to Analyze and Visualize Writing Processes,È vol. 30, n¡ %13, pp. 358-392, 2013.**

[24] **C. Leblay, ÇLe Temps de l'ƒcriture. Gen•se, durŽe, reprŽsentations,È 2011. [En ligne]. Available: https://www.jyu.fi/ajankohtaista/arkisto/2011/11/tiedote-2011-11-04-10-14-59-722468. [Acc•s le 15 12 2013].**

[25] **V. Southavilay, K. Yacef, P. Reimann et R. A. Calvo, ÇAnalysis of Collaborative Writing Processes Using Revision Maps and Probabilistic Topic Models,È** 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, ÇGraph data management and mining: a survey of algorithms and applications,È chez** Managing and mining graph data**, C. C. Aggarwal et H. Wang, ƒds., New York, Springer, 2010, pp. 13-68.**

[31] **L. Takac et M. Zabovsky, ÇData analysis in public social networks,È Lomza, Poland, 2012.**

[32] **M. Kuramochi et G. Karypis, ÇFinding frequent patterns in a large sparse graph,È** Data mining and knowledge discovery, **vol. 11, n¡ %13, pp. 243-271, 2005.**

[33] **C. Garcia-Debanc et K. Bonnemaison, ÇLa gestion de la cohŽsion textuelle par des ŽI•ves de 11-12 ans: rŽussites et difficultŽs,È chez** CMLF 2014**, Toulouse.**

[34] **H.-S. Becotte-Boutin, G. Caporossi et A. Hertz, ÇThe progressive visualization, a new tool for analyzing the writing process,È** Cahiers du GERAD, **vol. 131, 2015.**

[35] **H. Zha, Y. Yang, J. Wang et L. Wen, ÇTransforming XPDL to Petri Nets,È chez** Business Process Management Workshops**.**

[36] **Y. Wu, L. Wang, J. Ren et W. Ding, ÇMining sequential patterns eith periodic wildcard gaps,È** Applied intelligence, **vol. 41, n¡ %11, pp. 99-116, 2014.**

[37] **WritingPro, ÇWriting Pro,È 2014. [En ligne]. Available: http://www.writingpro.eu/. [Acc•s le 02 08 2014].**

[38] **I. Witten, ÇText mining,È Hamilton, New Zealand, 2004.**

[39] **G. Weikum, J. Hoffart, N. Nakashole, M. Spaniol, F. Suchanek et M. A. Yosef, ÇBig data methods for computational linguistics,È** IEEE Data Engineering Bulletin, **vol. 35, pp. 46-55, 2012.**

[40] **J. Veronis, ÇComputerized Correction of Phonographic Errors,È** Computers and the Humanities, **vol. 22, n¡ %11, pp. 43-56, 1988.**

[41] **L. Van Waes et P. J. Schellens, ÇWriting Profiles : The Effect of the Writing Mode on Pausing and Revision Patterns of Experienced Writers,È** Journal of Pragmatics, **vol. 35, pp. 829-853, 2003.**

[42] **L. Van Waes et M. Leijten, ÇInputlog 6.0: State of the art.,È chez** Paper presented at the Keystroke logging training school**, Antwerp, 2014.**

[43] **L. Van Waes et M. Leijten, ÇInputlog 6.0: Pause and fluency analysis.,È chez** Paper presented at the keystroke logging training school**, Antwerp, 2014.**

[44] **R. E. Tarjan, ÇProblems in data structures and algorithms,È chez** Graph theory, combinatorics and algorithms**, M. C. Golumbic et I. B. Hartman, *f*ds., Springer, 2005, pp. 17-39.**

[45] **S. Stromqvist, K. Holmqvist, V. Johansson, H. Karlsson et A. Wengelin, ÇWhat Keystroke Logging can Reveal about Writing,È chez** Computer Keystroke Logging and Writing**, K. P. S. &. E. Lindgren, *f*d., Elsevier, 2006, pp. 45-71.**

[46] **K. Spelman Miller et K. P. Sullivan, ÇKeystroke Logging : An Introduction,È chez** Computer Keystroke Logging and Writing**, Elsevier, 2006, pp. 1-10.**

[47] **K. Severinson Eklundh et P. Kollberg, ÇA Computer Tool and Franework for Analysing On-Line Revisions,È chez** The Science of Writing : Theories, Methods, Individual Differences, and Applications**, Mahwah, NJ, Lawrence Erbaum, 1996, pp. 163-188.**

[48] **D. H. Roen et R. Willey, ÇThe Effects of Audience Awareness on Drafting and Revising,È** Research in the Teaching of English, **vol. 22, n¡ %11, pp. 75-88, 1988.**

[49] **S. Plane, D. Alamargot et J.-L. Levrabe, ÇTemporalitŽ de l'ƒcriture et R™le du Texte Produit dans l'ActivitŽ RŽdactionnelle,È** Langages, **vol. 177, n¡ %11, pp. 11-32, 2010.**

[50] **D. Perrin, ÇProgression Analysis (PA): Investigating Writing Strategies at the Workplace,È** Journal of Pragmatics, **vol. 35, pp. 907-921, 2003.**

[51] **R. D. Owston, S. Murphy et H. H. Wideman, ÇThe Effects of Word Processing on Student's Writing Quality and Revision Strategies,È** Research in the Teaching of English, **vol. 26, n¡ %13, pp. 249-276, 1992.**

[52] **E. New, ÇComputer-Aided Writing in French as a Foreign Language : A Qualitative and Quantitative Look at the Process of Revision,È** The Modern Language Journal, **vol. 83, n¡ %1i, 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 Žd., Saint-Louis, 2012, p. 1094.**

[54] **Merriam-Webster, ÇApophenia, Merriam-Webster,È 2014. [En ligne]. Available: http://www.merriam-webster.com/dictionary/apophenia. [Acc•s le 03 08 2014].**

[55] **J. Leskovec, J. Kleinberg et C. Faloutsos, ÇGraph evolution: Densification and shrinking diameters,È** ACM Transactions on Knowledge Discovery from Data, **vol. 1, n¡ %11, p. Article 2, 2007.**

[56] **M. Leijten et L. Van Waes, ÇInputlog features,È 2014. [En ligne]. Available: http://www.inputlog.net/description_features.html. [Acc•s le 25 07 2014].**

[57] **J.-L. Lebrave et A. GrŽsillon, ÇLinguistique et gŽnŽtique des textes : un dŽcalogue,È 23 03 2009. [En ligne]. Available: http://item.ens.fr/index.php?id=434571. [Acc•s le 14 12 2013].**

[58] **J.-L. Lebrave, ÇComment Žcriront-ils?,È** Diog□ne, **vol. 196, n¡ %14, pp. 163-171, 2001.**

[59] **C. Leblay et G. Caporossi, ÇIntroduction aux donnŽes temporelles de l'Žcriture,È chez** Temps de l'□criture: enregistrements et repr□sentations**, C. Leblay et G. Caporossi, ƒds., Louvain-la-Neuve, Academia, 2014, pp. 5-15.**

[60] **C. Leblay, ÇEn de•ˆ du bien et du mal Žcrire,È 14 06 2012. [En ligne]. Available: http://www.item.ens.fr/index.php?id=578258. [Acc•s le 07 07 2014].**

[61] **M. M. A. Latif, ÇA State-of-the-Art Review of the Real-Time Computer-Aided Study of the Writing Process,È** International Journal of English Studies, **vol. 8, n¡ %11, pp. 29-50, 2008.**

[62] **P. Kollberg, ÇS-notation as a tool for analysing the episodic nature of revisions,È Barcelona, 1996.**

[63] **R. Karp, ÇOptimization problems related to internet congestion control,È chez** Graph theory, combinatorics and algorithms**, M. C. Golumbic et I. B. Hartman, ƒds., Springer, 2005, pp. 1-16.**

[64] **G. Kanawaty, Introduction ˆ l'Žtude du travail 3e Ždition, Bureau International du Travail Žd., Gen•ve, 1996.**

[65] **S. Helbing et S. Balietti, ÇFrom social data mining to forecasting socio-economic crises,È** The European Physical Journal Special Topics, **vol. 195, pp. 3-68, 2011.**

[66] **C. Haas, ÇHow the Writing Medium Shapes the Writing Process : Effects of Word Processing on Planning,È** Research in the Teaching of English, **vol. 23, n¡ %12, pp. 181-207, 1989.**

[67] **D. Foucambert et J. Foucambert, ÇGestes d'Žcriture et caractŽristiques linguistiques des textes achevŽs,È chez** Temps de l'criture: enregistrements et repr sentations**, C. Leblay et G. Caporossi, ƒds., Louvain-la-Neuve, Academia, 2014, pp. 43-70.**

[68] **C. Doquet, ÇPour une approche linguistique de l'Žcriture enregistrŽe,È chez** Temps de l'criture: enregistrements et repr sentations**, Academia, ƒd., LouvainlLa-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 InterprŽtation de l'Empan Temporel des Variations Prosodiques en Fran•ais et en Anglais, Marseille: UniversitŽ de Provence, 2010.**

[71] **M. Cox, C. Ortmeier-Hopper et K. E. Tirabassi, ÇTeaching Writing for the "Real World": Community and Workplace Writing,È** The English Journal, **vol. 98, n¡ %15, pp. 72-80, 2009.**

[72] **J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein et C. Welton, ÇMAD skills: new analysis practices for big data,È** 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, ÇMathematical sclices of molecular biology,È** Gazette des math maticiens, dition sp ciale, **vol. 88, pp. 11-80, 2001.**

[75] **I. Breetvelt, H. Van Den Bergh et G. Rijlaarsdam, ÇRelations between Writing Processes and Text Quality : When and How,È** Cognition and Instruction, **vol. 12, n¡ %12, pp. 103-123, 1994.**

[76] **M. Bramer, Principles of data mining, 2nd Žd., Portsmouth: Springer, 2013.**

[77] **M. R. Berthold, ÇBisociative knowledge discovery,È chez** Advances in intelligent data analysis X**, J. Gama, E. Bradley et J. Hollmen, ƒds., Porto, Springer, 2011, pp. 1-7.**

[78] **A.-L. Barabasi, R. Albert et H. Jeong, ÇScale-free characteristics of random networks: the topology of the world-wide web,È** Physica A, **vol. 281, pp. 69-77, 2000.**

[79] **D. Alamargot, G. Caporossi, D. Chesnet et C. Ros, ÇWhat Makes a Skilled Writer? Working Memory and Audience Awareness During Text Composition,È** Learning and Individual Differences, **vol. 21, pp. 505-516, 2011.**

[80] **R. Agrawal, T. Imielinski et A. Swami, ÇMining association rules between sets of items in large databases,È New York, 1993.**

[81] **L. Flower et J. R. Hayes, ÇA cognitive process theory of writing,È** College Composition

and Communication, **vol. 32, n¡ %14, pp. 365-387, 1981.**

[82] P. Kollberg, ÇRules for the S-notation: a compuyter-based method for representing revisions,È IPLab, Royal Institute of Technology (KTH), Stockholm, Sweden, 1996.

[83] D. Alamargot et J.-L. Lebrave, ÇThe study of professional writing: A joint contribution from cognitive psychology and genetic criticism,È European Psychologist, **n¡ %1doi:10.1027/1016-9040/a000001, 2009.**

[84] T. Olive, J.-L. Lebrave, J.-M. Passerault et N. Le Bigot, ÇLa dimension visuo-spatiale de la production de textes: approches de psychologie cognitive et de critique gŽnŽtique,È Langages, **vol. 177, n¡ %11, pp. 29-55, 2010.**

[85] W. Daelemans, P. Berck et S. Gillis, ÇData mining as a Method for Linguistic Analysis: Dutch Diminutives,È Dutch Diminutives, Folia Linguistica, XXXI/I -2, **pp. 57-75, 1997.**

[86] P. A. Bath, C. Craigs, R. Maheswaran, J. Raymond et P. Willett, ÇUse of graph theory to identify patterns of deprivaton and high morbidity and mortality in public health data sets,È Journal of the American Medical Informatics Association, **vol. 12, n¡ %16, pp. 630-641, Nov-Dec 2005.**

[87] A. Markowetz, K. Blasziewicz, C. Montag, C. Switala et E. S. Thomas, ÇPsycho-Informatics: Big Data shaping modern psychometrics,È 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, ÇWriting and the Analysis of Revision: An Overview,È chez Computer Keystroke Logging and Writing**, Elsevier, 2006, pp. 31-44.**

[90] ITEM, ÇEnjeux de recherche,È 7 mai 2014. [En ligne]. Available: http://www.item.ens.fr/index.php?identifier=l-item. [Acc•s le 14 juin 2014].

[91] I. Breetvelt, H. v. d. Bergh et G. Rijlaarsdam, ÇRelations between writing processes and text quality: When and how?,È Cognition and Instruction, **vol. 12, n¡ %12, pp. 103-123, 1994.**

[92] C. Lynch, ÇHow do your data grow?,È Nature, **vol. 455, n¡ %14, pp. 28-29, 2008.**

[93] E. Midgette, P. Haria et C. MacArthur, ÇThe effects of content and audience awareness goals for revision on the persuasive essays of fifth- and eighth-grade students,È 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, ÇWhy high performance visual data analytics is both relevant and difficult,È Burlingame, California, USA, 2013.

[95] F. A. Abukhodair, B. E. Riecke, H. I. Erhan et C. D. Shaw, ÇDoes interactive animation control improve exploratory data analysis of animated trend visualization?,È Burlingame, California, USA, 2013.

[96] L. Bartram, ÇPerceptual and interpretative properties of motion for information visualization,È 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, ÇWhat makes a skilled writer? Working memory and audience awareness during text composition,È Learning and Individual Differences, **vol. 21, n¡ %15, pp. 505-516, 2011.**

[98] V. M. Baaijen, D. Galbraith et K. d. Glopper, ÇKeystroke analysis: reflections on procedures and measures,È 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, ÇTemps et pauses dans l'Žcriture au clavier,È chez Temps de l'criture: enregistrements et repr sentations**, Louvain-la-Neuve, Academia, 2014, pp. 97-124.**

[101] W. Aigner, S. Miksch, H. Schumann et C. Tominski, ÇVisualization of Time-Oriented Data,È Springer, London, 2011.

[102] D. Perrin et S. Laemmel, ÇApplication ˆ l'Žcriture journalistique,È chez Temps de l'criture, enregistrements et repr sentations**, Louvain-la-Neuve, Academia-L'Harmattan s.a., 2014, pp. 171-192.**

[103] A. Unwin, C.-h. Chen et W. K. Hardle, ÇIntroduction,È chez Handbook of Data Visualization**, Berlin, Springer, 2008, pp. 3-14.**

[104] A. Vathy-Fogarassy et J. Abonyi, Graph-Based clusterting and data visualization, Springer, 2013.

[105] D. Chesnet et D. Alamargot, ÇEye and Pen 2 manuel de l'utilisateur,È Poitiers, 2011.

[106] E. R. Tufte, The visual display of quantitative information, Second Žd., Cheshire, Connecticut: Graphics Press, 2001.

---

[1] **Deux graphes, ou sous-graphes, sont dits** isomorphiques **lorsquÕils reprŽsentent clairement des diagrammes semblables [26, p. 4]. De mani•re plus formelle, un graphe** ! " ($ &)**est isomorphique ˆ un graphe** ! " ($ &)**sÕil y a une correspondance entre les ensemble de sommets** $ **et** $ **(et entre les ensembles dÕar•tes** & **et** & **(de telle mani•re que si** / **(est une ar•te joignant les sommets** 0 **(et** 1 **, dans le graphe** ! **, alors lÕar•te correspondante** / **(dans(**! **joins les sommets** 0 **(et** 1 **, ce qui corresponds ˆ** 0 **(et** 1 **respectivement [29]. Les algorithmes pour la recherche de structures semblables dans les graphes se basent donc sur cette propriŽtŽ [73].**

[2] **LÕŽquipe TERs (Temps de lÕEcriture et ses ReprŽsentations) est une Žquipe de lÕInstitut des Textes et Manuscrits Modernes (ITEM/ENS-CNRS).**