Intégration de la visualisation dans l'analyse de processus complexes : écritures et réécritures dans un corpus multilingue universitaire

Denis Foucambert[1,*], *Tracy* Heranic[2], *Christophe* Leblay[3], *Maarit* Mutta[3], et *Minjing* Zhong[4]

[1] Département de linguistique, Université du Québec à Montréal, Canada
[2] Département d'études françaises, Université de Concordia, Canada
[3] School of Languages and Translation studies, University of Turku, Finland
[4] Department of French, Xiangtan University, Chine

Résumé. Cet article traite des processus d'écriture d'apprenants universitaires de français langue étrangère (L2/L3). Ces écritures ont été recueillies à l'aide du programme *GenoGraphiX-Log* qui est construit sur les exigences de la génétique textuelle et de la théorie mathématique des graphes. Notre corpus consiste en 44 écritures produites en français par des locuteurs ayant comme L1 soit l'anglais, soit le mandarin soit le finnois. Le premier objectif est de mieux décrire, en fonction des L1 des participants, les opérations d'écriture mises en œuvre en français lors de la réalisation d'une même tâche. Le second objectif est d'évaluer comment la visualisation, basée sur la théorie des graphes et sur des méthodes statistiques inductives, soutient cette analyse des processus d'écriture. Les résultats se basent sur deux analyses issues des graphes : une analyse en composantes principales (ACP) et la visualisation des écritures exemplaires (les plus proches des centres de gravité de chaque groupe). Ces deux analyses complémentaires nous permettent de mesurer les spécificités des trois groupes et d'approfondir qualitativement l'analyse de signatures scripturales. Notre corpus et nos analyses montrent l'intérêt de l'utilisation des méthodes mixtes dans l'analyse des processus complexes d'écriture à l'aide d'outils de visualisation.

Abstract. Exploiting visualization techniques in analyzing complex processes: writing and rewriting operations in a multilingual academic corpus. This article discusses the writing/rewriting processes of university-level students of French as a foreign language (L2/L3). The analyzed writing samples were acquired through the use of a keystroke logging program, GenoGraphiX-Log, software particularly suited to the needs of Text Genetics research. The data collected through the software allows for the analysis of the writing process through various visualization techniques. Our corpus consists of 44 keystroke recordings by participants with different L1s (English, Chinese and Finnish) during a writing task in

* Corresponding author : foucambert.denis@uqam.ca

French. The first objective is to better understand the ways in which a text is crafted in French by writers faced with the same writing task but coming from different L1s. The second objective is to evaluate how visualization, based on graph theory and statistical inductive methods, supports this analysis of writing processes. The results are based on two analyses derived from the graphs: principal component analysis (PCA) and graphic visualization of exemplary writing (the closest to the centers of gravity of each participant group). The two complementary analyses allow us to measure the specificities of the three groups and to qualitatively deepen the analysis of the scriptural signatures. Our corpus and our analyses show the interest of using mixed methods in the analysis of complex writing processes by using visualization tools.

1 Introduction

Ce que donne à voir la technologie contemporaine de l'écriture n'est pas vain : aux considérations concernant l'espace nouveau d'une écriture numérisée se superpose le temps direct, appelé parfois à tort *temps réel*, de l'écriture. Ce dernier, enregistré par différents logiciels dédiés, autorise enfin à voir le temps de l'écriture se dérouler devant nos yeux dans des visualisations : diverses scriptions (comme les ajouts, les insertions, les suppressions et les remplacements), mouvements du curseur, jets textuels (Lindgren & Sullivan, 2019). Ces approches récentes du temps de l'écriture (Cislaru & Olive, 2018) dont les travaux précurseurs datent des années 2000 (Doquet-Lacoste, 2003; Lindgren, 2005; Spelman Miller, 2000) sont mises en contraste avec celles qui relèvent des comparaisons de versions produites sur papier (Fabre, 1987), davantage orientées sur le déroulement spatial de la page, offrent de l'écriture une description de plus en plus précise et exhaustive et ôtent définitivement à l'écrit sur papier l'exclusivité qu'il avait pour l'étude de l'écriture et de ses processus (Latif, 2008).

Nous parlerons alors d'*écriture(s) enregistrée(s)* pour décrire ce que la génétique textuelle nomme *réécriture(s)*, où sont mises en évidence toutes les possibilités gestuelles offertes par l'utilisation d'un clavier associé à des périphériques (souris, pavé tactile principalement). Dans cette perspective, il semble judicieux de considérer la notion d'*acquisitio*n en production écrite : celle-ci a tous les traits d'une procédure, c'est-à-dire d'une *représentation motrice* (ou *gestuelle*), qui relève de l'*implicite*, de la *spontanéité* et de la *durée*; ce mode de l'acquisition relève surtout de la difficulté de l'*observation* et demande alors une approche autrement outillée que celle du papier-crayon.

Des trois méthodes qui se présentent à qui veut s'attacher à recueillir, décrire et analyser un corpus numérique de productions écrites (comparaison de produits, analyse chronométrique et interviews), nous nous concentrerons principalement sur l'analyse chronométrique à l'aide d'un logiciel d'enregistrement des écritures *GenoGraphiX-Log* 2.0 (Usoof *et al*., 2020) (dorénavant *GGXLog)* totalement adapté aux exigences génétiques (Lebrave, 2011). Le but est alors d'enregistrer le déroulement de l'écriture *en acte* sans prédéterminer le choix délicat et incontournable de la transcription. Parce qu'il s'agit bien d'enregistrer pour mieux étudier. Dès 2010, Plane, Alamargot et Lebrave soulignent déjà, avec raison, à quel point la temporalité de la production écrite reste un phénomène complexe, en ces termes: « Il s'agira […] de circonscrire les différentes «strates» temporelles de l'activité rédactionnelle, en montrant en quoi les différents processus se déroulent dans des temporalités différentes. » Il s'agit bien de cela, de pouvoir noter, transcrire, représenter ou visualiser les différentes temporalités.

1.1 Génétique textuelle & visualisation des écritures

Le travail génétique effectué, depuis l'année 1976, au moment où est créé le CAM (Centre d'Analyse des Manuscrits) lequel deviendra, en 1982, l'ITEM (Institut des Textes et Manuscrits Modernes (www.item.ens.fr), a été marqué par une volonté de se démarquer de la critique philologique, en refusant le jugement de valeur du temps qui s'écoule dans une perspective de dégradation; à la différence de la philologie, la génétique du texte ne considère pas le *temps écoulé* comme une perte de qualité, la qualité textuelle venant souvent de la succession de versions du même texte. Parallèlement, la critique génétique a aussi pris position contre l'idée de la *clôture du texte produit* que la mouvance structuraliste réclamait. L'opposition est donc forte entre un produit fini, achevé, stabilisé, édité et une production non finie, non achevée, non stabilisée et non éditée (Anokhina & Pétillon 2009). Aux acquis donc méthodologiques, et descriptifs comme les opérations d'écriture (voir plus loin) de la génétique textuelle, nous avons associé la théorie mathématique des graphes qui donne à voir ce qui restait caché soit même ce qui disparaissait. La visualisation des processus d'écriture, ainsi modélisée, apporte des précisions sur la compétence stratégique des scripteurs multilingues et aide à identifier des écritures individuelles.

Selon la théorie des graphes appliquée à l'écriture, les différentes activités scripturales telles que la succession des segments, la durée des pauses, par exemple, peuvent être visualisées à l'aide de jeux de couleurs, et de formes (Bécotte-Boutin et al., 2019; Leblay & Caporossi, 2015). Ainsi, dans la représentation graphique de ces derniers, la couleur rouge peut être associée à l'inscription, les couleurs jaune et bleue à la suppression, tandis que le vert peut donner à voir, avec précision, la succession et la durée des pauses. Contrairement aux autres programmes d'écritures enregistrées, *GGXLog* n'utilise pas de représentations dites issues de Systèmes d'Information Géographique (SIG) sur deux axes (*x* et *y*), mais bien des formes, librement colorées, composées de cercles, appelées *nœuds* et de traits, appelés *liens*. Ce jeu entre ces nœuds et liens offre de grandes possibilités de visualisations.

D'une manière globale, la visualisation par graphe est une combinaison de *scénarios* qui peuvent se produire au cours d'une session d'écriture naturelle. Les modèles (*pattern*) permettent d'identifier facilement le scénario qui s'est produit à un moment donné et comment celui-ci affecte ce qui va devenir du texte, ainsi que sa relation avec d'autres événements. Nous précisons néanmoins que nous présentons les graphes sous la forme de saisies d'écran qui permettent de donner à voir le jeu des formes et des couleurs, mais qui perdent le côté dynamique offert par le programme. Ainsi, ne peuvent pas être montrés les nœuds et les liens qui apparaissent progressivement, lors de la construction des segments qui sont affichés lorsque la souris est déplacée sur un nœud.

Parmi les différentes options visuelles présentées sur le programme, nous avons choisi celle qui est nommée *graphe progressif* (*Progressive Graph*) qui permet de visualiser le graphe à différentes étapes de la session d'écriture. Le texte concurrent est alors simultanément affiché dans la zone de texte située sous le graphe. Conformément aux descriptions offertes par la *génétique textuelle*, descriptions auxquelles le programme *GGXLog* est fortement attaché, s'ensuit une liste théorique des différentes opérations d'écriture adaptées aux divers scénarios d'écriture (*insertion*, *suppression*, *déplacement*), mais aussi des opérations numériques de *copier-coller, couper-coller* et *pause* (Usoof *et al*., 2020). Voici, d'une manière théorique, sans support textuel, des éléments descriptifs:

Ainsi, une *insertion*, opération d'ajout effectuée par retour dans le *déjà écrit*, divisera un nœud existant en deux parties avec des numéros consécutifs (nœuds 1 et 2) et l'insertion, en tant que telle aura un numéro de nœud plus grand que les deux nœuds précédents (nœud 3), alors que l'ordre chronologique sera maintenu dans un lien continu.

Une *suppression,* effectuée à l'aide de la touche d'effacement arrière (*backspace key*), est visualisée d'une manière spécifique qui a pour effet de diviser le nœud en trois parties (nœuds 1, 2 et 3) qui maintiendront l'ordre chronologique initial. La séquence produite

exclura alors le segment supprimé (nœud 3), qui sera typiquement coloré d'une nuance différente. En outre, l'événement de suppression (nœud 4) est inclus dans l'ordre chronologique pour souligner le moment où la suppression a été effectuée. D'une manière pratique, l'ordre inversé des lettres dans le nœud de suppression indique que la touche d'effacement arrière a été utilisée pour la suppression. Une *suppression,* effectuée à l'aide de la touche de suppression (*delete key*) ou de la mise en surbrillance (*highlight*) est aussi prise en compte et visualisée d'une manière spécifique. Ainsi, le nœud est divisé en trois parties (nœuds 1, 2 et 3) pour maintenir l'ordre chronologique initial. La séquence produite exclura le segment supprimé (nœud 3), qui sera typiquement coloré dans une nuance différente. En outre, l'événement de suppression (nœud 4) est inclus dans l'ordre chronologique pour souligner le moment où la suppression a été effectuée. Une mise en surbrillance et une suppression peuvent être affichées lorsqu'un segment existant est mis en surbrillance à l'aide de la souris ou du clavier et qu'il est ensuite supprimé à l'aide principalement des touches d'espacement arrière ou de suppression ou lorsque ce segment est remplacé à l'aide d'une action clavier ou souris (couper-copier-coller).

L'opération d'écriture nommée *copier-coller* est aussi prise en compte, même si celle-ci n'est pas un objet de la description génétique traditionnelle. Dans l'exemple ci-dessous, une partie du texte existant (nœud 1) est copiée puis suivie d'un ajout (nœud 2). Ensuite, le texte copié est collé (nœud 3) et enfin suivi d'un autre ajout au texte (nœud 4). Ici, le lien discontinu entre les nœuds 1 et 3 représente l'endroit d'où le segment collé a été copié. Cet exemple illustre également un scénario dans lequel l'ordre chronologique des événements est similaire à celui de la séquence produite.

L'opération de *déplacement* ou l'opération de *couper-coller* est prise en compte. Ce scénario est traité d'une manière similaire à celle d'une suppression suivie d'un ajout ou d'une insertion. Comme dans le cas d'une suppression, le nœud existant est divisé en 3 (nœuds 1, 2 et 3) en conservant l'ordre chronologique d'écriture. Le segment déplacé est représenté dans une nuance différente (nœud 3), et l'événement de la suppression est représenté par un nœud de suppression (nœud 4). Le segment déplacé est maintenant représenté comme un ajout ou une insertion, mais dans sa propre teinte pour mettre en évidence le déplacement. La ligne discontinue entre le nœud déplacé et le nœud de suppression montre la relation entre cette suppression et l'ajout ou l'insertion.

L'opération de *pause*, sentie comme une opération de degré zéro, devient aussi un objet visuel. Lorsque l'option "Afficher la pause" est sélectionnée dans les paramètres (cf. Annexe 3), les pauses supérieures ou égales à une valeur de pause choisie (par exemple 2000 ms) sont représentées par un nœud associé à une couleur spécifique (nœud 2). Le temps de pause est alors indiqué à l'intérieur du nœud.

1.2 Objectifs de la présente étude

Deux objectifs en interaction sont à l'origine du travail présenté ici. Notre premier objectif est de mieux comprendre les moyens mis en œuvre par des locuteurs ayant des L1 différentes lorsqu'ils se trouvent confrontés à une consigne d'écriture identique dans une langue étrangère (L2 ou L3) identique (ici le français). Les questions auxquelles la littérature scientifique a commencé à apporter quelques réponses s'articulent notamment autour de la recherche de différences individuelles et des régularités observables dans le processus de productions des apprenants avancés de L2 (Breuer, 2019; Lindgren et al., 2008), même si les travaux sur le français L2 sont quasiment encore inexistants. Cet objectif cherche *in fine* à mesurer et à mieux comprendre les différences dans les processus d'écriture selon des profils langagiers (notamment dus à la langue première), entre idiosyncrasie et homogénéité des comportements scripturaux.

Notre second objectif est d'évaluer comment la visualisation, basée sur la théorie des graphes, soutient cette analyse des processus d'écriture. De ces visualisations sont extraites

des mesures spécifiques qui doivent permettre de chercher des régularités et des dissonances dans les écritures. Peut-on montrer visuellement ce qui regroupe et oppose les écritures réalisées par des participants appartenant à des groupes bien définis ayant des L1 différentes?

2 Méthode

Notre corpus est composé des enregistrements produits par 44 scripteurs, tous étudiants universitaires de français langue étrangère (L2 ou L3), avec des niveaux de compétences linguistiques variées, du niveau intermédiaire au niveau avancé.

Les protocoles se sont déroulés dans 3 universités différentes (Université de Concordia, Université de Turku et Université de Xiangtan), dans trois pays différents (Canada, Finlande et Chine).

2.1 Population

Les 13 participants canadiens sont âgés de 18 à 39 ans (moyenne d'âge 24,1 ans). Ils partagent la même L1, l'anglais d'Amérique du Nord, et ils ont tous appris le français comme L2 dans des contextes variés. Selon eux, s'ils ont appris d'autres langues, les participants les maîtrisent moins que le français. Ils sont tous inscrits dans un cours de langue universitaire au niveau B2 respectant les critères du CECR (2001).

Les 8 participants finlandais sont âgés de 20 à 43 ans (moyenne d'âge 24,6), leur L1 est le finnois. Il est à noter que la plupart des apprenants en Finlande choisissent l'anglais comme leur première langue étrangère à l'école, à l'âge de 10 ans (à partir du printemps 2020, à l'âge de 7 ans) et que le suédois est la deuxième langue officielle en Finlande, mais que le niveau de compétences en suédois varie beaucoup d'un étudiant à l'autre. Tous les étudiants ont dû passer un concours d'entrée pour pouvoir commencer les études de ces langues au niveau universitaire ; le niveau des compétences linguistiques est au moins le niveau B1 selon les critères du CECR (2001).

Les 23 participants chinois sont âgés de 18 à 21 ans (moyenne d'âge 19,5), leur L1 est le mandarin et ils étudient le français comme matière principale en Licence 2. Tous ces étudiants apprennent l'anglais comme leur première langue étrangère dès l'âge de 10 ou 12 ans. Leur niveau de français devrait atteindre le niveau B1 selon les critères du CECR (2001).

2.2 Procédure

Lors d'une première étape, chaque étudiant a téléchargé et installé sur son ordinateur personnel le logiciel *GGXLog* à partir du site https://www.ggxlog.net/. Lors de la seconde étape, les étudiants ont pu prendre connaissance de la tâche de rédaction, par l'explication orale des consignes d'écriture (tableau 1).

Il a été signifié aux étudiants que le temps d'écriture est volontairement limité à 10-15 minutes (Canada, Finlande) ou à 40 minutes (Chine, à l'occasion d'un contrôle continu) entre le moment où l'enregistrement est déclenché et celui où l'enregistrement est arrêté. Les outils extérieurs tels que les dictionnaires n'étaient pas censés être utilisés.

Une fois l'enregistrement terminé, chaque étudiant a été chargé de faire parvenir son fichier (anonymisé par *GGXLog*) au chercheur responsable. L'anonymisation finale des données a été faite après l'analyse ; nous avons suivi les consignes éthiques de chaque université pour protéger les données personnelles des participants. Ces participants ont eu à signer un formulaire de consentement.

2.3 Matériel

Le corpus s'articule autour d'une seule et unique tâche narrative proposée dans les travaux de Garcia-Debanc et Bonnemaison (2014) repris dans la thèse de Zhong (2020).

Tableau 1. Consigne d'écriture.

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/s'est retourné en entendant ce grand bruit.*
- *Depuis cette aventure, les enfants ne sortent plus la nuit.*

Vous serez attentifs aux éléments présents dans ces trois phrases qui vous permettront d'organiser votre texte : personnages, lieu, moment.

Cette tâche, destinée initialement à des écritures sur papier-crayon, a été adaptée à des écritures numériques enregistrées à l'aide du programme *GGXLog*. De plus, étant donné le niveau de compétences linguistiques des participants, l'utilisation du passé simple s'est avérée être trop exigeante puisque ces scripteurs ne connaissent la forme verbale du passé simple que passivement et ne peuvent pas l'utiliser activement dans leur production. Nous avons donc introduit la possibilité que les participants utilisent aussi le passé composé.

3 Résultats et discussions

Les résultats sont présentés en trois sections. Après une rapide description des textes produits, nous présentons une analyse des statistiques issues des graphes générés par *GGXLog*. Cette partie nous permet de trouver quelques écritures représentatives de chacun des groupes de participants. Nous montrerons alors, dans une troisième section, ces graphes exemplaires.

3.1 Description sommaire des textes finaux

Le tableau 2 présente quelques éléments sur la longueur des textes produits par les différents participants en fonction de leur L1.

Tableau 2. Variables descriptives des textes produits en fonction de la L1 des scripteurs.

	L1	N	Moyenne	SD	Min.	Max.
Nombre de mots	Anglais	13	155,08	51,62	94	257
	Finnois	8	168,25	64,37	85	264
	Mandarin	23	191,74	46,68	127	319
Nombre de phrases	Anglais	13	11,77	4,88	5	19
	Finnois	8	13,25	4,10	8	20
	Mandarin	23	14,39	4,27	3	22
Nombre de paragraphes	Anglais	13	2,23	1,54	1	5
	Finnois	8	2,63	2,77	1	8
	Mandarin	23	2,00	1,04	1	4
Longueur des phrases	Anglais	13	14,03	3,28	7,95	18,8
	Finnois	8	12,56	2,21	10,40	16,5
	Mandarin	23	15,00	7,79	8,42	45,7

Plusieurs ANOVAS univariées indiquent qu'on n'observe de différences significatives entre les trois groupes de participants ni pour le nombre de mots [$F(2, 41) = 2{,}27$; $p < 0{,}2$], ni pour le nombre de phrases [$F(2, 41) = 1{,}46$; $p < 0{,}3$], ni pour la longueur moyenne des phrases [$F(2, 41) = 0{,}50$; $p < 0{,}65$], ni pour le nombre de paragraphes [$F(2, 41) = 0{,}46$; $p < 0{,}7$]. Ces différents résultats indiquent une relative similarité quant à la surface des textes finaux produits par les participants, quel que soit leur L1.

Cependant, étant donné la disparité dans les consignes d'écriture en ce qui concerne les durées d'écriture, nous retiendrons principalement dans les analyses suivantes les pourcentages moyens d'événements par écriture.

3.2 Analyses de statistiques basées sur les graphes

Complémentairement à la visualisation des graphes sur les processus d'écritures, *GGXLog* extrait un ensemble de variables (environ 30) qui synthétisent les éléments portés par les graphes et permettent une analyse comparative des différentes écritures d'un corpus. Pour tenter de répondre à nos objectifs, nous nous sommes servis d'un sous-ensemble de 14 variables (voir Annexe 1) sur les 44 écritures que nous avons pu traiter. La description exhaustive de ces variables et de leurs calculs dépasse largement le cadre de ce travail et peut être retrouvée très précisément dans l'article de Usoof *et al*. (2020). Notons tout de même qu'elles caractérisent les actions d'écriture et qu'elles peuvent se regrouper en grandes rubriques :

1. Une première (3 variables) qui caractérise la taille moyenne des unités sur lesquelles des opérations d'insertions, d'ajouts ou de suppressions ont eu lieu. Ce sont respectivement les variables 4, 5 et 6.
2. Une deuxième qui caractérise la ventilation des différents types d'opérations effectuées par la présentation des pourcentages d'occurrences de ces opérations d'insertions, d'ajouts, de suppressions ou de pauses. Ce sont les variables 7 à 14.
3. Une dernière rubrique de trois variables plus générales :
    a. La variable 1 montre le nombre de nœuds produits par minute.
    b. La variable 2 (*Final Text Edges to Chronology Edges Ratio)* donne un aperçu du nombre de nœuds dans le texte final par rapport à l'ensemble des événements de développement et d'édition du texte. La valeur maximale sera 1 lorsqu'il n'y a pas d'éditions ou de pauses dans la session d'écriture. Plus le nombre de révisions et de pauses est élevé, plus la valeur sera faible (Usoof *et al*., 2020).
    c. La variable 3 (*Final Text Edges to Total Edges Ratio*) donne un aperçu du nombre de nœuds dans le texte final produit par rapport à la complexité du processus d'écriture mesurée par le nombre total d'arêtes dans le graphe. La valeur maximale sera de 1 lorsqu'il n'y a pas d'édition ou de pause dans la session d'écriture. Plus le nombre d'éditions et de pauses est élevé, plus la valeur est faible ; plus le nombre est faible, plus le processus d'écriture est complexe. Le nombre d'effacements et de déplacements aura un impact plus important sur cette valeur (Usoof *et al*., 2020).

Pour traiter ce tableau de données relativement complexe, nous avons introduit simultanément l'ensemble de ces 14 variables dans une Analyse en Composantes Principales (dorénavant ACP), à l'aide du paquet FactoMineR (Le et al., 2008) intégré à Jamovi (The jamovi project, 2021). Les ACP font partie d'un des grands courants en statistique, l'analyse des données. Les principes reposent sur quelques points que Jean-Paul Benzecri (1982) a popularisés : statistique sans probabilité ni modèle a priori « le modèle doit suivre les données non l'inverse » (Benzecri & coll., 1973, p. 6), pour une analyse se donnant comme objectif de « traiter simultanément des informations concernant le plus

grand nombre de dimensions ». Les ACP, anciennes dans leur conception théorique (Hotelling, 1933; Pearson, 1901), permettent de décrire un ensemble d'informations effectuées sur plusieurs variables numériques en prenant en compte simultanément l'ensemble de la variance de la totalité des données (donc, toutes les mesures). Les analyses factorielles portent sur des « nuages de points » dont on cherche à trouver un certain nombre d'axes factoriels qui sont des combinaisons des variables initiales et qui ont comme caractéristiques principales d'être non corrélés entre-elles et de variance successivement maximale. L'objectif de l'ACP est de remplacer le tableau initial de données, difficile à lire et surtout impossible à représenter graphiquement, par des tableaux plus faciles à lire et par des représentations graphiques compréhensibles, tout en perdant le moins d'information possible.

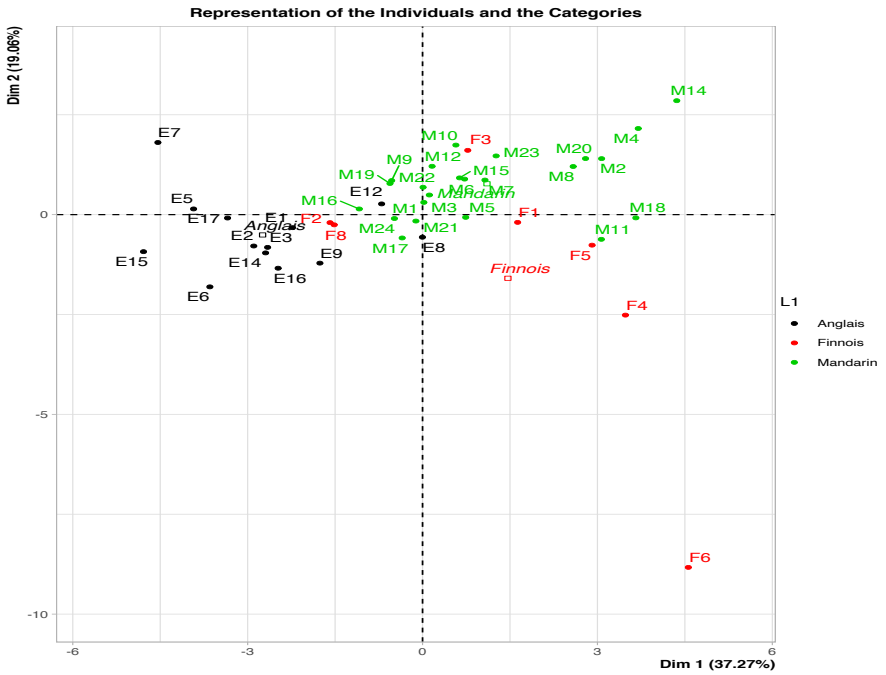


Fig 1. Projection des 44 écritures sur les deux premiers axes de l'ACP, en fonction des trois groupes de participants.

Pour commencer une ACP, il convient de vérifier l'adéquation de l'échantillonnage et la pertinence des données pour la réduction par, respectivement, le test Kaiser-Meyer-Olkin (KMO) et le test de sphéricité de Bartlett. Dans notre cas, le premier présente une valeur de 0.54 ce qui est considéré comme acceptable (Kaiser, 1958; Redmond et al., 2022), et le second s'avère significatif [$\chi^2 = 11{,}67$; df = 91; $p<0{,}001$]. Ces deux indices nous permettent de continuer les analyses (Redmond et al., 2022).

Il devient donc légitime d'observer la projection des différentes variables sur les axes factoriels (Annexe 2) et de tenter leur interprétation, en respectant la condition de l'indépendance totale des axes pris deux à deux. On ne retiendra ici que les deux premières dimensions (ou facteurs) qui représentent à eux seuls environ 60% de la variance totale (Axe 1 : valeur propre = 5,96; Axe 2 : valeur propre = 3.05). On observe à droite du premier facteur (horizontal) un ensemble de variables[i] qui dénotent un fort pourcentage d'insertion (*Insert.Nodes*) et de suppression (*Delete.Nodes ; Remove.Nodes*) de textes. À l'inverse, du côté gauche de ce premier axe, on voit des variables qui expriment un grand nombre d'ajouts de texte (*Append.Nodes*) et de pauses (*Pause.Nodes*). Deux types

d'écritures sont représentées par ce premier axe : du côté gauche, des textes écrits au fur et à mesure, sans retour en arrière, sans écriture au milieu de ce qui est déjà écrit. Très peu de suppressions de ce côté du facteur, mais, en revanche, beaucoup de circonspections et de réflexion dans l'écriture par un nombre important de pauses. À droite, des textes écrits avec des écritures au sein de ce qui existe déjà, donc en modifiant le déjà-écrit, voire en le faisant disparaître puisqu'on y observe également de nombreuses suppressions. Pour ce qui est de l'axe 2 (vertical), on observe, en haut du plan, deux variables qui expriment, d'une part, une écriture très peu complexe (*Final Text Edges to Total Edges Ratio*) avec très peu de suppressions et de déplacements (Usoof et al., 2020). Ces écritures ne présentent également que très peu d'éditions (de différentes sortes) ou de pauses (*Final Text Edges to Chronology Edges Ratio*). À l'inverse, la partie basse du plan est construite principalement par la variable exprimant le nombre de copies (*Percentage Copy Nodes*) et, de manière plus légère, par la variable exprimant les éléments provisoirement supprimés (*Percentage Removed Nodes*). Ces deux variables fonctionnent ici ensemble et montrent des écritures ayant un volume de déplacements important. Cet axe représente donc celui de la complexité dans l'acte d'écriture. L'ACP permet également de projeter dans le plan Axe1 x Axe2 les écritures observées et, par un jeu de couleurs, la L1 de chacun des individus ayant écrit (figure 1).

On observe une très grande homogénéité du groupe anglais_L1 (en noir) qui se situe à gauche de l'axe horizontal et légèrement en bas de l'axe vertical. On peut donc attribuer aux écritures faites par des individus anglais-L1 différentes caractéristiques propres à la gauche de cet axe (décrit plus haut) et sans caractéristiques spécifiques de la droite de l'axe. Sur l'axe vertical, les individus du groupe anglais-L1 ne sont pas du tout homogènes, les différences entre, par exemple, E7 et E6 pouvant témoigner de variations idiosyncrasiques. On peut également analyser la position des individus ayant le mandarin en L1 (en vert) et qui semblent également former un groupe plutôt cohérent. La position du centre de gravité (point carré vert) de ce groupe montre que ces écritures sont plutôt caractérisables par une activité sur le déjà-écrit tel que nous l'avons décrit plus haut dans la caractérisation de la droite de l'axe 1. Enfin, le groupe ayant le finnois comme L1 (en rouge) est beaucoup plus éclaté sur le plan factoriel. Ceci peut s'expliquer par deux éléments distincts : d'une part, la (trop) petite taille de ce groupe qui complique singulièrement le travail de l'ACP (on voit par exemple, le poids de 6 qui est d'autant plus important que le groupe est de petite taille) et, d'autre part, par la possible difficulté à bien caractériser le statut du français comme L2 ou L3 pour les étudiants finlandais (on peut voir que F2 et F8 se comportent comme le groupe des anglais-L1 alors que F4 ou F5 ont une projection sur l'axe horizontal proche de celle des mandarin-L1).

Notons également que des indices supplémentaires montrent que l'axe horizontal discrimine significativement le groupe anglais-L1 et mandarin-L1 ($p<0.001$; $R2 = 0,55$) et que l'axe vertical discrimine le groupe mandarin-L1 et finnois-L1 ($p<0.03$; $R2 = 0.26$), bien que la représentativité du centre de gravité du groupe finnois-L1 reste sujette à caution.

3.3. Extraction de sous-graphes

À partir des résultats de l'analyse précédente, nous pouvons ainsi extraire les écritures qui sont les plus proches des centres de gravité de chacun des groupes. Pour l'exercice, nous ne prendrons que des écritures représentant maximalement les groupes anglais-L1 (E14) et mandarins-L1 (M23).

Plus précisément, nous avons choisi de nous concentrer sur le tout début de deux écritures afin de rendre plus visible et plus lisible cette façon nouvelle de rendre compte

visuellement de l'écriture. Chacune des deux écritures a été limitée au même nombre d'événements (proches de 100) afin de les rendre comparables.

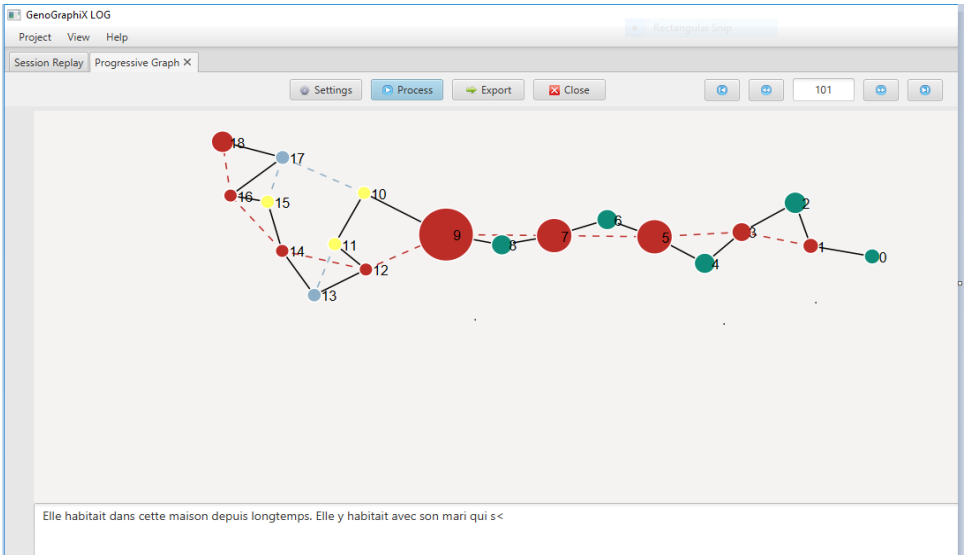


Fig 2. Sous-graphe d'un scripteur du groupe anglais L1 (E14)

La figure 2 montre le graphe qui visualise le tout début de l'écriture de E14, en insistant sur 3 points principalement : 1) Le *code couleur*, spécifié dans les paramètres, est adaptable (cf. Annexe 3) ; 2) les numéros sur les nœuds donnent l'ordonnancement, autrement dit la *successivité* dans laquelle ceux-ci ont été produits : ainsi le nœud noté 0 est celui qui amorce l'écriture pour arriver jusqu'au nœud noté 17 qui clôture cette séquence ; 3) le rayon de ces nœuds note le nombre de caractères : ainsi le nœud 9 comprend plus de caractères que le nœud 1.

Ainsi, cette écriture d'une très courte séquence est signée de la manière suivante :

- Une première phase avec une série d'ajouts successifs (nœuds 1, 3, 5, 7 et 9) mêlée à des pauses (nœuds 2, 4, 6 et 8).
- Une seconde phase d'écriture, où sont mêlés segments ajoutés et suppressions immédiates (nœuds 10, 11 et 15). La couleur jaune, associée par défaut, à un nœud supprimé rend compte de l'hésitation du scripteur à rédiger le mot *mari* : celui passe ainsi par plusieurs phases (*amr ->a -> ama -> mari*), ce qu'il serait facile de voir en déplaçant le curseur de la souris sur les nœuds. Cette couleur est uniquement présente dans cette deuxième phase.

Ces deux phases sont très distinctes : la première est construite sur des segments de plus en plus construits suivie d'une autre phase bien plus hésitante.

Voyons un autre exemple (figure 3) qui permet de mettre en contraste des *signatures* très différentes. Si le premier exemple montrait une signature en deux phases distinctes, cette seconde séquence (M23), présente des phases autrement construites.

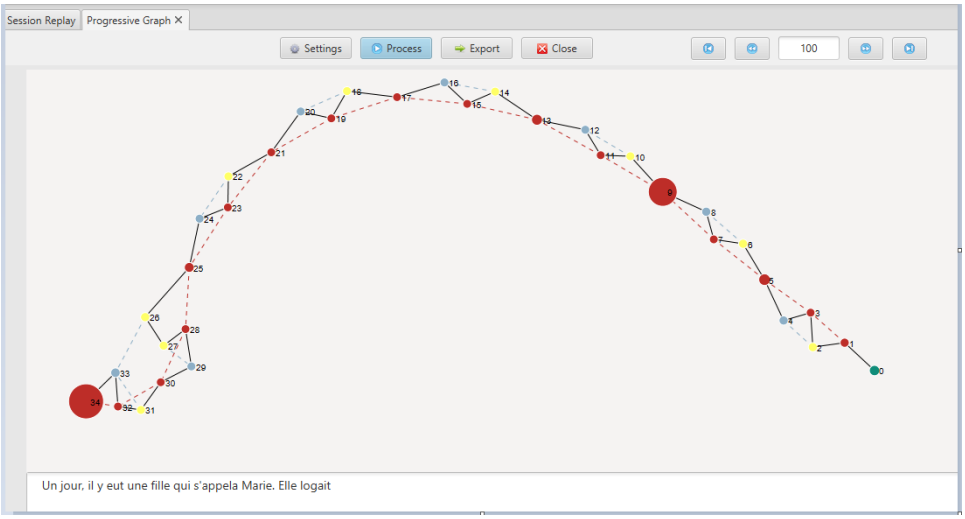


Fig 3. Sous-graphe d'un scripteur du groupe mandarin (M23)

Ainsi, cette écriture d'une courte séquence est signée de la manière suivante :

- Une première phase (nœuds 0 à 9) se termine par un ajout d'un segment représenté par le nœud 9. Cette phase est composée de pauses et de suppressions.
- Une seconde phase, depuis le nœud 10 jusqu'au nœud 34 lequel représente un ajout de segment. Cette phase est également composée de pauses et de suppressions.

Ces deux phases sont construites de manière identique, comme si, avant de produire un segment plus conséquent, plus construit, le scripteur avait besoin de ruminations (pause et suppression).

La brièveté de ces deux séquences nous empêche de tirer des conclusions trop hâtives : seule la suite des textes peut permettre de voir au sens propre si ces toutes premières descriptions confirment, ou pas, une vraie signature. Nous avons ainsi conscience que chacun des scripteurs a pu librement user de son habileté, ou non, à se servir des médiums disponibles (clavier, pavé tactile, flèches directionnelles, souris). Les choix dans les gestes d'écriture sont alors clairs : écrire des deux mains, ou de quelques doigts ne relève pas de la même compétence scripturale ; d'une manière identique, savoir se déplacer dans le texte déjà produit implique aussi des choix à faire entre la souris, le pavé tactile ou les flèches directionnelles, ce qui a des conséquences dans la description des opérations d'écriture. Il est, néanmoins, attendu que les scripteurs ont déjà eu à montrer, bien avant cet enregistrement, tant dans leurs scolarités achevées que dans leurs études universitaires, entamées ou achevées, une habileté minimale. Mais, cette habileté-là qui n'a pas été vérifiée a pu avoir un impact dans une épreuve en temps limité.

4 Conclusion

En premier lieu, et comme nous l'avons montré plus haut, ce travail permet de visualiser certaines différences entre les scripteurs de différentes L1, notamment entre le groupe de sinophones et celui d'anglophones. Dans le cadre des productions issues de la tâche que nous utilisons, les processus rédactionnels semblent très différents entre ces deux groupes. Si de nombreuses recherches doivent approfondir ce constat et essayer d'en comprendre les fondements, nous pouvons dès à présent émettre quelques pistes de travaux à entreprendre pour mieux cerner ces spécificités. Ces différences sont-elles simplement liées au niveau de compétences linguistiques en L2 des participants ou sont-elles le fait de la L1 en elle-

même, de son fonctionnement comme des différences purement linguistiques avec la L2 ? Peut-on envisager que le rôle des attitudes et des motivations différentes face à une L2 (Pernet-Liu, 2017) et à son utilisation ou son apprentissage (Rifai, 2010) induisent ces différences de comportements scripturaux ? Faut-il y voir une interaction de ces deux facteurs avec la consigne en elle-même et le type de texte qu'elle amène à produire ? La question de comment intégrer à ces observations les travaux classiques sur le développement de l'écriture (Alamargot & Fayol, 2009, Bereiter & Scardamalia, 1987) reste également ouverte et notre approche peut y conduire à de nouveaux éclairages. Enfin, confronter les processus d'écritures des mêmes sujets lorsqu'on leur demande d'écrire des textes d'un autre genre peut s'avérer fécond pour mieux comprendre ce que la consigne en elle-même contraint comme éventuelles spécificités dans les opérations d'écriture.

Du point de vue didactique, notre approche montre un intérêt pour les chercheurs mais aussi pour les enseignants. Nous avons présenté de manière plus détaillée les écritures qui étaient les plus proches des centres de gravité de chacun des groupes illustrant ainsi des spécificités de certains groupes de scripteurs ayant la même L1. Or, la visualisation des sous-graphes d'un scripteur permet à l'enseignant d'utiliser le logiciel en tant qu'outil pédagogique pour indiquer les points problématiques, les retours sur le texte ainsi que les parties des processus automatisés (cf. entre autres Mutta & Salminen, 2021). Cela permet aussi au scripteur de reconnaître sa signature, ou même, son propre profil d'écriture.

En second lieu, et de manière plus méthodologique, les deux analyses successives – issues de données provenant d'un même logiciel – que nous avons présentées sont complémentaires. La première analyse, l'ACP, permet de mesurer ce qui est spécifique entre les écritures des trois groupes de participants. La seconde, la visualisation des graphes exemplaires, permet d'approfondir qualitativement le regard porté sur un processus d'écriture particulier, repéré grâce à la première analyse factorielle.

Indépendamment de cette liaison entre les deux analyses, chacune des deux garde ses forces intrinsèques qui peuvent être mises à profit dans d'autres recherches portant sur un corpus important d'écriture. La proposition méthodologique que nous faisons, intégrant des aspects mathématiques dans leurs conceptions (analyses factorielles et théorie des graphes) et inductifs dans leurs utilisations (sans hypothèses *a priori*), permet d'aborder la complexité des processus d'écriture avec des outils aussi novateurs que féconds. En ce sens, et bien que sur un objet légèrement différent, nous partageons le plaidoyer d'Enever et Lindgren (2017) sur la nécessaire utilisation des méthodes mixtes dans l'analyse des objets complexes.

Le corpus utilisé ici nous a permis de mettre en place un certain nombre de procédures qui pourraient être reproduites et servir de base dans un corpus plus ambitieux. L'idée est de recueillir une quantité précise d'enregistrements permettant de mettre en place des visualisations adaptées. Et la recherche des « écritures exemplaires » peut se faire sur des regroupements de processus d'écriture différents de ceux que nous avons présentés pour permettre d'observer les différences (et les similarités) dans les processus selon, par exemple, le genre de textes à écrire ou selon différents niveaux d'habileté langagière que ce soit dans la L1 ou dans les langues secondes… La liste est longue des objets à mieux comprendre.

Références bibliographiques

Alamargot, D., & Fayol, M. (2009). Modelling the development of written composition. In R. Beard, D. Myhill, M. Nystrand, & J. Riley (Eds.), *Handbook of Writing Development* (pp. 23–47). Sage.

Anokhina, O., & Pétillon, S. (2009). De l'archive de la création aux processus cognitifs. In O. Anokhina & S. Pétillon, *Critique génétique: Concepts, méthodes, outils* (pp. 5–19). Éditions de l'IMEC.

Bécotte-Boutin, H.-S., Caporossi, G., Leblay, C., & Hertz, A. (2019). Writing and rewriting: Keystroke logging's colored numerical visualization. In E. Lindgren & K. P. H. Sullivan, *Observing writing: Logging handwriting and computer keystrokes* (pp. 96–124). Brill Academic Publishers.

Benzécri, J.-P. (1982). *Histoire et Préhistoire de l'Analyse des Données*. Bordas.

Benzecri, J.-P., & coll. (1973). *L'analyse des correspondances*. Dunod.

Bereiter, C., & Scardamalia, M. (1987). *The Psychology of Written Composition*. Lawrence Erlbaum Associates.

Breuer, E., Odilia. (2019). Fluency in L1 and FL writing : An analysis of planning, essay writing and final revision. In E. Lindgren & K. P. H. Sullivan, *Observing writing : Insights from keystroke logging and handwriting* (p. 190-211). Brill Academic Publishers.

CECR. (2001). *Cadre européen commun de référence pour les langues. Apprendre, enseigner, évaluer*. Didier.

Charolles, M. (1988). La gestion des risques de confusion entre personnages dans une tâche rédactionnelle. *Pratiques*, *60*, 75-97.

Cislaru, G., & Olive, T. (2018). *Le processus de textualisation*. De Boeck Supérieur. https://doi.org/10.3917/dbu.cisla.2018.01

Doquet-Lacoste, C. (2003). *Étude génétique de l'écriture sur traitement de texte d'élèves de cours moyen 2, année 1995-96*. Université Paris 3.

Enever, J., & Lindgren, E. (Éds.). (2017). *Early language learning : Complexity and mixed methods* (Vol. 1-1). Multilingual Matters; DiVA. https://doi.org/10.21832/ENEVER8316

Fabre, C. (1987). *Les activités métalinguistiques dans les écrits scolaires*. Descartes Paris V.

Garcia-Debanc, C., & Bonnemaison, K. (2014). La gestion de la cohésion textuelle par des élèves de 11-12 ans : Réussites et difficultés. *SHS Web of Conferences*, *8*, 961-976. https://doi.org/10.1051/shsconf/20140801349

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417-441. https://doi.org/10.1037/h0071325

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*, 187-200. https://doi.org/10.1007/BF02289233

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

Le, S., Josse, J., & Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*, *25*(1), 1-18. https://doi.org/10.18637/jss.v025.i01

Leblay, C., & Caporossi, G. (2015). A graph theory approach to online writing data visualization. In G. Cislaru, *Writing(s) at the Crossroads. The Process-Product Interface* (p. 171-181). John Benjamins Publishing Company.

Lebrave, J.-L. (2011). Computer forensics : La critique génétique et l'écriture numérique. *Genesis*, *33*, 137-147. https://doi.org/10.4000/genesis.633

Lindgren, E. (2005). *Writing and revising : Didactic and methodological implications of keystroke logging* [PhD thesis]. Umeå University.

Lindgren, E., Spelman Miller, K., & Sullivan, K. P. H. (2008). Development of Fluency and Revision in L1 and L2 Writing in Swedish High School Years Eight and Nine. In *ITL - International Journal of Applied Linguistics* (Vol. 156, Numéro 1, p. 133-151). John Benjamins.

Lindgren, E., & Sullivan, K. P. H. (2019). *Observing writing : Insights from key-strokelogging and handwriting*. Brill Academic Publishers.

Mutta, M. & Salminen S. (2021). Les séquences préfabriquées dans la production écrite dans le cas de scripteurs finnophones de français et suédois L2. *Synergies pays riverains de la Baltique 14/20*, 11-26

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559-572. https://doi.org/10.1080/14786440109462720

Pernet-Liu, A. (2017). "Cultural Anchors" of Chinese Students' University Writing: The Learning Culture and Academic Traditions. In S. Plane, B. Charles, F. Rondelli, C. Donahue, A. N. Applebee, & C. Boré (Eds.), *Research on Writing: Multiple Perspectives* (pp. 299–310). The WAC Clearinghouse; CREM. https://doi.org/10.37514/INT-B.2017.0919.2.16

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

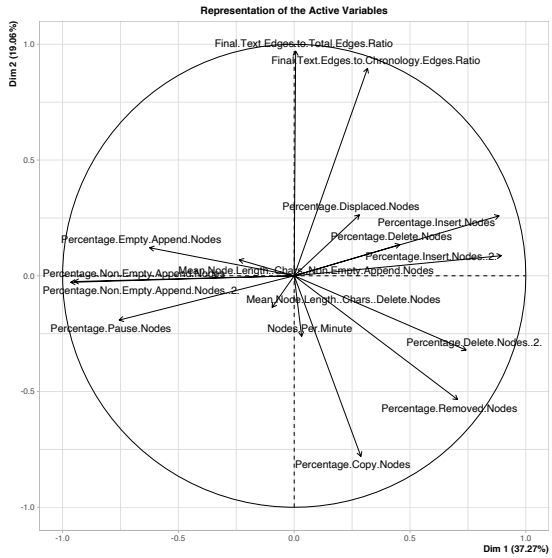
Redmond, L., Foucambert, D., & Libersan, L. (2022). Language corpora and factor analysis. In D.

Woolford, D. Kotsopoulos, & B. Samuels (Éds.), *Applied Data Sciences : Data Translators Across the Disciplines* (p. 30p).
Rifai, N. A. (2010). Attitude, motivation, and difficulties involved in learning the English language and factors that affect motivation in learning it. *Procedia - Social and Behavioral Sciences*, *2*(2), 5216–5227. https://doi.org/10.1016/j.sbspro.2010.03.849
Spelman Miller, K. (2000). Academic writers on-line : Investigating pausing in the production of text. *Language Teaching Research*, *4*(2), 123-148. https://doi.org/10.1177/136216880000400203
The jamovi project. (2021). *Jamovi* (2.0) [Computer software]. https://www.jamovi.org
Usoof, H., Leblay, C., & Caporossi, G. (2020). GenoGraphiX-Log version 2.0 user guide. *Les Cahiers Du GERAD*, *2020*(68), 1-63.
Zhong, M. (2020). *Produire un texte cohérent dans une langue étrangère : L'exemple d'étudiants chinois de niveau intermédiaire et avancé de FLE en France* [Thèse de doctorat en sciences du langage]. Paris 3.

Annexe 1

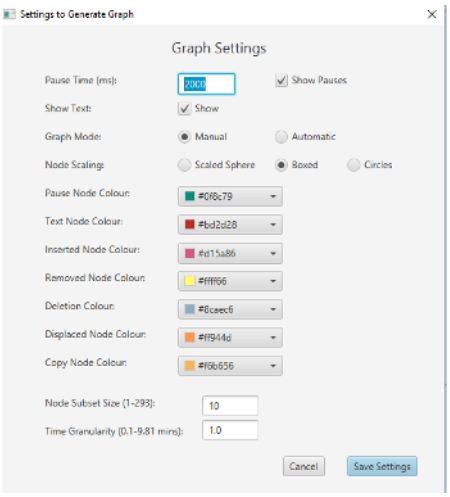
Num	Variables	Moyenne	Écart-type
1	Nodes Per Minute	17,44	7,18
2	Final Text Edges to Chronology Edges Ratio	0,52	0,06
3	Final Text Edges to Total Edges Ratio	0,30	0,03
4	Mean Node Length Non-Empty Append Nodes	19,13	35,29
5	Mean Node Length Delete Nodes	3,43	2,46
6	Mean Node Length Insert Nodes	3,20	1,79
7	Percentage Insert Nodes	23,55	17,06
8	Percentage Removed Nodes	22,12	6,68
9	Percentage Non-Empty Append Nodes	13,5	8,21
10	Percentage Non-Empty Append Nodes (2)	53,19	27,22
11	Percentage Insert Nodes (2)	27,05	22,55
12	Percentage Delete Nodes (2)	19,97	8,8
13	Percentage Pause Nodes	6,78	11,24
14	Percentage Empty Append Nodes	15,55	8,42

Annexe 2



Projection des 14 variables sur les deux premiers facteurs de l'ACP.

Annexe 3



Annexe 4

E14_Texte final: *Elle habitait dans cette maison depuis longtemps. Elle y habitait avec son mari qui s'appelle Thomas. Thomas et sa femme dînaient sur la nuit d'halloween. Le ciel était gris et il y avait une mauvaise tempête. Thomas est allé vers la salle de bain pendant le dîner. Il s'est retourné en entendant ce grand buit. << Quel bruit>> a dit par Thomas. Sa femme l'ai regardé avec la peur dans les yeux. Les enfants de Thomas et sa femme sont descendus les escaliers au grand bruit. Ils se couvraient le visage avec leurs couvertures. Puis, le courant s'arrêtait de travailler. Tout le monde a crié. Depuis cette aventure, les enfants ne sortent plus la nuit.*

Annexe 5

M23_Texte final: *Un jour, il y avait une fille qui s'appelait Marie. Elle avait dix ans. Elle habitait à la compagne avec ses parents. Sa famille avait une grande maison dans laquelle on plantait trop de légnmes et de fleurs. Chaque fois que le Printemps arrivait, les gens toujours sentaient des herbes et fleurs. Marie l'aimait.*
Elle habitait dans cette maison depuis longtemps. Elle faisait beaucoup d'amis avec d'autres enfants qui y logeait aussi. François, le bon ami de Marie, ils allaient souvent d'ailleurs sans prévenir leurs parents. Le soir, ils se sont amusés de plus. On allait bien. Mais quand ils rentrèrent chez leur, il était nuit. Ils passèrent un grand forêt. Soudain, le petit garçon, il se retourna en entendant ce grand bruit. Puis, Marie vit un loup! Les enfants courir en criant. En ce moment, leurs parents leur ont trouvé et amené. Ces deux enfants criaient tout le temps.
Depuis cette aventure, les enfants ne sortent plus la nuit.

[i] Différents critères permettent d'évaluer la contribution des variables aux axes. Dans le cadre de cet article, nous ne présenterons que la corrélation entre les variables et les axes : on ne retiendra ici que les variables avec une | corrélation axe x variable | > 0.5.