DOSSIER : Lecture, écriture et informatique

# Décrire un texte et son écriture
# Présentation de premières analyses

Claire DOQUET Benoît FOUCAMBERT

On sait que le logiciel **Lecture méthodique** procède à une analyse formelle des textes qu'on lui soumet* et que la **Genèse du texte** ne se limite pas à la restitution chronologique de l'écriture mais récapitule et classe sous forme de tableaux et de graphiques les éléments susceptibles de caractériser les activités d'écriture. À l'aide de ces informations, Claire DOQUET et Benoît FOUCAMBERT ont procédé à une analyse statistique d'une vingtaine de textes afin d'en tenter une "description objective". Du fait de la nature et du nombre de textes étudiés, il s'agit plus d'une proposition méthodologique que d'une véritable présentation de résultats probants. On verra pourtant quelle aide ces outils peuvent apporter à ceux qui, sensibles à la difficulté d'apprécier une production écrite, cherchent par des moyens rigoureux une approche différente.

* cf. ***un logiciel de lecture méthodique***. Denis FOUCAMBERT. AL n°37, mars 92, p.28

L'axe statistique de la recherche Genèse du texte consiste en une investigation systématique de processus d'écriture et de textes produits selon des consignes précises. L'objectif à long terme est d'établir des corrélations entre types de texte et stratégies d'écriture, ces stratégies pouvant varier selon la tranche d'âge des scripteurs. Autrement dit :

• en quoi l'écriture d'une nouvelle diffère-t-elle, par exemple, de celle d'une lettre ou d'un article documentaire?

• en fonction de l'âge du scripteur, quelles stratégies sont mises en œuvre dans l'écriture d'un même type d'écrit ?

L'analyse exposée ici se préoccupe de la deuxième question : selon l'âge des scripteurs, qu'est-ce qui varie dans manière d'écrire un texte selon une même consigne, quels indices sont discriminants lorsqu'on observe les textes produits ?

16 processus d'écriture d'enfants scolarisés de 8 à 16 ans et 6 processus d'écriture d'adultes familiers de cette activité ont été recueillis** selon la consigne :
"*Regarde cette image : tu es le personnage derrière la fenêtre,-tu vois ce qui se passe dehors. Qu'est-ce que tu penses ? Qu'est-ce qui risque de se passer après ?*"

Disons tout de suite le vague de cette consigne et les inconvénients qu'il engendre pour l'analyse. Le destinataire n'étant pas précisé, le texte produit prend des allures de prétexte. D'autre part, certains scripteurs nous ont dit avoir hésité entre deux partis opposés : raconter en se mettant en scène ("*j'étais assis dans mon fauteuil quand j'entendis un bruit terrible...*"), ce qui rapproche l'écrit d'une rédaction classique, ou bien écrire les pensées qui traversent la tête du personnage ("*Mais qu'est-ce qui se passe ? Non mais je rêve !...*"), ce qui produit en général un texte plus original, très éloigné en tout cas des productions scolaires habituelles.

** Il a été demandé à une dizaine de personnes de bien vouloir "noter" (au sens le plus scolaire du terme) sur une échelle de 0 à 40 chaque texte présenté anonymement et la note à laquelle il est fait référence plus loin est donc la moyenne de ces opérations chiffrées.

Il faut encore signaler que les adultes scripteurs ont été mis dans la situation particulièrement difficile de devoir écrire un texte selon une consigne qui n'était pas faite pour eux. Tandis que certains ont tenté de jouer le jeu, d'autres ont simplement écrit *ce qu'ils attendraient de leurs élèves*, ce qui crée évidemment des décalages dans la manière d'envisager le texte, donc de l'écrire.

Ces réserves quant à la consigne de production ne disent que trop celles qu'il faudra émettre quant aux résultats de l'analyse, où l'incertitude s'accroît à cause du faible effectif de l'échantillon (22 individus). Toutefois, il nous a semblé intéressant de présenter les variables que nous avons recensées ainsi que les investigations statistiques effectuées. Plus qu'un ensemble de résultats sur un échantillon donné, cet article présente la genèse d'une méthodologie d'analyse.

## Nous avons observé...

L'analyse a été effectuée à partir des processus d'écriture des textes grâce au logiciel **Genèse du texte** et à partir de leur état final grâce au logiciel **Lecture méthodique.**

Dans le processus d'écriture des textes, nous avons recherché les variables suivantes :
- **Taux de reformulation**. En comparant le nombre d'ajouts effectués en cours d'écriture au nombre de mots final, on obtient un coefficient correspondant à la quantité de mots auxquels le scripteur a eu recours pour un mot finalement conservé.
- **Productivité**. C'est le temps moyen d'écriture d'un mot restant, mesuré en divisant le temps d'écriture du texte par le nombre de mots du texte final.
- **Rythme de production**. Il est mesuré par l'activité moyenne déployée par le scripteur en cours d'écriture. L'activité consiste en l'ajout et la suppression de mots, elle est exprimée en mots/heure.
- **Nature de l'activité à la fin de l'écriture**. On compare le nombre de mots que comporte le texte à la fin du deuxième tiers de son écriture par rapport au nombre de mots final. Il s'agit de savoir en quoi a consisté l'activité du dernier tiers de l'écriture : continuation du processus entamé (ou poursuite des ajouts selon un rythme constant) ou révision du texte, qui peut se concrétiser par une baisse sensible du rythme des ajouts ou même une chute du nombre de mots signifiant que l'auteur a surtout supprimé.
- **Tendance de la lisibilité**. Le texte a-t-il la même lisibilité en début et en fin d'écriture, ou bien cette lisibilité augmente-t-elle (simplification) ou au contraire se réduit-elle (complexification) ?
- **Gestion de la globalité du texte**. Elle est mesurée par la proportion d'opérations en lecture par rapport à l'ensemble des opérations qui montre à quel degré l'auteur travaille son texte en dehors de la phrase qu'il est en train d'écrire. En d'autres termes : y a-t-il gestion de la cohésion du texte dans son entier ou seulement de la cohérence de l'environnement proche ?
- **Expansion du texte**. La proportion d'ajouts en lecture par rapport aux ajouts en écriture permet d'observer si l'auteur "gonfle" son texte en revenant en arrière et en ajoutant ou au contraire s'il écrit "au fil de la plume".
- **Élagage du texte**. La proportion de suppressions en lecture par rapport aux suppressions en écriture montre à quel moment interviennent les remords : après la relecture d'un passage ou au contraire au cours de l'écriture.
- **Nature du travail en lecture**. La proportion d'ajouts en lecture par rapport aux suppressions en lecture indique si les opérations en lecture sont plutôt des ajouts ou plutôt des suppressions.
- **Temps d'attente moyen entre les opérations**. C'est un indice du temps de réflexion nécessaire à l'écriture. Il ne tient pas compte de la vitesse de frappe.
- **Temps d'attente en milieu de phrase**. On cherche à savoir si, dans le cadre d'une opération en écriture, les temps d'attente notables (supérieurs à la moyenne + écart-type des attentes) se situent en milieu de phrase ou en fin de phrase. Ceci correspondrait à deux comportements typiques :

l'attente en fin de phrase (donc en début de la phrase suivante) signifie que l'auteur forme presque complètement une phrase dans sa tête avant de l'écrire. L'attente en milieu de phrase est typique d'une personne qui, en entamant une phrase, ne sait pas exactement ce qu'elle va devenir et utilise l'écrit comme moyen de poursuivre sa pensée en la formulant.
- **Temps d'attente moyen avant opération en lecture**. Comparé au temps moyen avant toute opération, il permettra se savoir si l'auteur attend plus longtemps avant d'effectuer une opération en lecture qu'une opération en écriture.

Dans les textes produits, nous avons recherché les variables suivantes :
- **Lisibilité du texte final**. Cet indice est calculé automatiquement selon la formule de Rudolph FLESH, basée sur le nombre de mots par phrase et la longueur des mots employés.
- **Taille du texte**, exprimée en nombre de mots.
- **Longueur moyenne des phrases**. La longueur des phrases est témoin de leur complexité.
- **Composition du texte**. La taille moyenne des paragraphes par rapport à la taille du texte, exprimée en pourcentage, c'est un indice de la structuration du texte, en partant de l'idée que plus un texte est découpé en paragraphes, plus sa composition est claire.
- **Nombre des temps de verbes utilisés**. Ce chiffre brut est significatif de la maîtrise des temps des verbes ainsi que de la variété temporelle du texte. S'il est courant d'employer trois temps différents, ce seuil franchi il s'agit vraisemblablement d'un phénomène de gestion du temps dans le récit ou de recherche stylistique.
- **Densité verbale**. Il s'agit du nombre de verbes par rapport à l'ensemble des mots.
- **Proportion d'organisateurs temporels**. Les organisateurs temporels sont la marque de la gestion de la temporalité dans le texte. Il s'agit par exemple de locutions ou d'adverbes tels que : le lendemain, un peu plus tard, auparavant, etc.
- **Complexité du texte**. Il s'agit de la proportion de subordonnées par rapport au nombre de verbes conjugués que contient le texte. Cette proportion peut être comparée à la longueur moyenne des phrases, indicateur de la complexité du texte. En principe, ces deux indices devraient varier en même temps. Que dire d'un texte aux phrases très longues comportant peu de subordonnées, par exemple ?
- **TTR** (proportion de mots différents par rapport au nombre de mots total). On mesure ici la variété du vocabulaire employé.
- **Proportion de formats anormaux par rapport au nombre de mots**. Les formats anormaux sont l'emphase ("*c'est Marie qui...*"), les formes passives complètes, les formes impersonnelles et les sujets inversés. Ils sont révélateurs de la volonté de l'auteur d'employer une formulation inhabituelle.
- **Proportion de phrases déclaratives**. C'est la diversité de la ponctuation qui est mesurée ici. Selon le type de texte, on peut s'attendre à trouver plus ou moins de phrases interrogatives ou exclamatives.
- **Densité syntagmatique**. C'est le rapport entre nombre de noms-noyaux et le nombre de qualifiants. Un nom-noyau entretient avec le verbe un rapport casuel correspondant aux fonctions traditionnelles de sujet, complément d'objet, complément circonstanciel... Au contraire, les qualifiants sont des périphériques des noms-noyaux : les adjectifs, mais aussi les compléments de nom, par exemple.
Plus la densité syntagmatique est forte, plus les qualifiants sont nombreux par rapport aux noms-noyaux : on a affaire à des syntagmes plutôt longs et complexes, symptomatiques de la maîtrise de l'agencement des éléments dans la phrase.
- **Modalités d'énoncés par rapport au nombre de mots**. Les modalités d'énoncé sont des adverbes ou locutions qui donnent à l'énoncé une valeur de certitude, de probabilité ou de nécessité, par exemple des adverbes tels que certainement, absolument ou des locutions telles que : il est clair que, il est possible que, il me semble que, etc. Elles portent sur le propos qu'elles nuancent.

## Comment ces enfants, ces adolescents, ces adultes ont-ils répondu à la consigne ?

Les scripteurs se répartissent selon quatre niveaux :

- niveau 1 : école primaire (classes de cycle 3)
- niveau 2 : collège (classes de 6ème)
- niveau 3 : lycée (classes de 2nde et 1ère G)
- niveau 4 : adultes "experts", habitués à écrire.

Les collégiens se caractérisent par une importante proportion d'opérations en lecture par rapport à l'ensemble des opérations effectuées et d'ajouts en lecture par rapport à l'ensemble des ajouts. Leur écriture pourrait être qualifiée de récursive, avec de fréquents retours en arrière destinés à ajouter du texte. Il s'agit d'une sorte de gonflement périodique du déjà écrit. Ce processus ne donne pas lieu à une attente particulière : les collégiens n'ont pas besoin de passer du temps à relire leur texte avant de procéder à des ajouts.

Les écoliers et les adultes s'opposent du point de vue de l'activité, des temps d'attente, du temps moyen d'écriture d'un mot restant, de la taille du texte final et de la note attribuée à ce texte. À des adultes très actifs qui hésitent peu entre les opérations pour produire des textes longs et bien notés s'opposent des écoliers écrivant plus lentement (mais la vitesse de frappe peut être en cause), qui attendent longtemps avant de se décider à ajouter ou supprimer un mot, qui produisent des textes courts auxquels ont été attribuées les moins bonnes notes.

Les adultes se caractérisent par des attentes en écriture situées en fin de phrase plutôt qu'au milieu. Ce comportement est symptomatique de scripteurs qui préparent mentalement leurs phrases avant de les inscrire. Il peut être imputable au décalage entre la légère difficulté de la consigne et les savoir-faire de ces scripteurs, adultes très familiers de l'écriture, qui semblent se conformer à une tâche facile dans laquelle ils s'impliquent peu.

Les écoliers ont ceci de spécifique qu'ils effectuent beaucoup de suppressions en lecture par rapport aux suppressions en écriture. Ils se différencient ainsi des collégiens qui ont le même comportement pour les ajouts. En revanche, ils attendent longtemps avant d'effectuer une opération en lecture, ce qui n'est pas le cas des collégiens. Leur comportement n'est pas de "gonfler" le déjà écrit mais plutôt d'ajouter en fin de texte pour le relire périodiquement et effectuer des suppressions.

On peut s'interroger sur l'absence de caractérisation des lycéens. En effet, rien dans leur comportement ne les spécifie par rapport aux autres groupes. En observant les processus d'écriture de leurs textes, on s'aperçoit qu'ils écrivent relativement vite (mais sont certainement moins actifs que les adultes), qu'ils procèdent à un rythme moyen à des opérations en lecture (mais sûrement moins fréquemment que les collégiens), que les textes produits sont assez longs (mais plus courts que ceux des adultes) et notés autour de la moyenne. Ces scripteurs entraînés n'ont plus les comportements hésitants des écoliers, leur manque de pratique ne leur a pas donné l'aisance des adultes, les variables analysées ne sont pas caractéristiques de leur mode d'écriture.

Les corrélations montrent que les différences de niveau des scripteurs se manifestent dans la manière d'écrire mais non dans le résultat de l'écriture tel que le logiciel l'a analysé, hormis la variable concernant le nombre de mots du texte final. Les marques de surface repérées de façon automatique font apparaître des différences quantitatives mais non qualitatives, du moins celles-ci ne sont-elles pas significatives du niveau des scripteurs.

Ceci peut s'expliquer, dans le cas des scripteurs débutants, par leur manque de pratique et leur méconnaissance du fonctionnement de l'écrit. De fait, si leurs textes sont parfois corrects du point de vue syntaxique et orthographique, ils manquent souvent de cohésion et n'emploient que très peu ses instruments que peuvent être, par exemple, les organisateurs temporels ou un nombre important de temps de verbes. Les "effets de style" repérables grâce aux formats anormaux ou à une densité syntagmatique inhabituelle sont trop peu nombreux et trop irrégulièrement dispersés pour que l'on puisse les considérer comme significatifs d'un niveau ou d'un autre.

Dans le cas des adultes, qui sont tous, répétons-le, des scripteurs réguliers, on s'étonne de l'absence de ce type de marques. Elle peut cependant s'expliquer par la relative facilité de la consigne et la situation artificielle dans laquelle sont mis les adultes, que nous avons déjà évoquée plus haut. Conscients que leurs écrits ne sont là que pour servir de repères, les adultes ne mettent probablement pas en oeuvre leurs comportements habituels de scripteurs mais se contentent d'aligner quelques phrases selon la consigne imposée, en cherchant parfois l'originalité mais sans utiliser l'écriture comme outil de réflexion et de pensée. À deux ou trois exceptions près peut-être, leur point de vue sur l'image et les personnages n'évolue pas entre le début et la fin de l'écriture.

## Qui attend beaucoup écrit lentement...

L'analyse des corrélations des variables entre elles, si elle aboutit parfois à ce genre de lapalissade, permet heureusement des rapprochements plus intéressants. En croisant les variables analysées, on peut discerner trois groupes de corrélations :

- celles qui concernent la mesure quantitative de l'activité d'écriture : activité/attente,
- celles qui concernent la lisibilité du texte et ses caractéristiques lexicales : densité syntagmatique/densité verbale,
- celles qui concernent la taille du texte et la note qui lui a été attribuée.

**Le premier groupe de variables** décrit le mode de production des textes. Il oppose des scripteurs rapides qui produisent des textes longs et bien notés à des scripteurs plus lents, qui attendent longtemps avant d'ajouter ou de supprimer un mot, que ce soit en lecture ou en écriture.

Les scripteurs qui travaillent le plus leur texte ont déjà écrit, en deux tiers de leur temps d'écriture, la presque totalité de ce texte : ils passent vraisemblablement le troisième tiers de l'écriture à travailler le texte et les modifications qu'ils y apportent vont au moins autant dans le sens de la suppression que dans celui de l'ajout.

**Le deuxième groupe de variables** décrit la syntaxe employée. Les textes se répartissent en deux groupes :

- Des textes très lisibles, où l'on trouve beaucoup de verbes, des phrases courtes composées de propositions simplement structurées ; ces phrases sont préparées mentalement par le scripteur et écrites sans attente en milieu de phrase.
- Des textes moins lisibles, où la forte densité syntagmatique témoigne de la complexité des structures nominales avec, par exemple, un taux important d'adjectifs ou de relatives ; ces phrases ne sont pas écrites d'un jet, le scripteur a besoin de temps de réflexion pour les travailler.

**Le troisième groupe de variables** dégage les caractéristiques des textes en fonction de la note qui leur a été attribuée. C'est un indicateur des critères d'évaluation du texte.

Meilleure est la note, plus long est le texte et plus forte a été l'activité déployée lors de son écriture. Au contraire, les textes dont l'écriture a donné lieu à beaucoup d'attente et où la productivité est la plus faible sont généralement assez peu appréciés.

## Qu'est-ce qu'on note quand on note ?

Pour affiner les critères de notation, nous avons tenté, en mettant en relation les caractéristiques des textes produits et la note qui leur est attribuée, de recenser les variables discriminantes. Nous avons formé trois groupes de textes en fonction de la note (de 0 à 40) : ceux qui obtiennent une note comprise entre 0 et 16, ceux qui obtiennent une note comprise entre 17 et 27, ceux qui obtiennent une note comprise entre 28 et 40.

Le premier groupe se caractérise par une forte proportion de déclaratives et de modalités d'énoncés ainsi que par l'absence de formats anormaux et de paragraphes : ces textes peu structurés formellement sont aussi assez uniformes du point de vue du type de phrases, leur vocabulaire est peu varié.

La caractéristique principale du deuxième groupe est une forte densité verbale, qui s'accompagne de phrases courtes et peu complexes. Ces textes simples et très lisibles sont néanmoins structurés par une forte proportion d'organisateurs temporels.

Les textes du troisième groupe comportent des phrases longues et complexes. Ce sont aussi des textes longs où la forte densité syntagmatique révèle la complexité des groupes nominaux. Les temps des verbes employés et le vocabulaire sont assez variés.

En bref, les évaluateurs se sont laissé séduire par des textes répondant aux normes classiques de la richesse de l'expression : phrases longues et complexes, variété des mots employés.

## Proposition de traitement automatique

À partir des résultats concernant le processus de production des textes, nous avons cherché, à travers une analyse en composantes principales, à déterminer des types de comportement. Avant de livrer les résultats obtenus, précisons que ce traitement statistique est généralement appliqué à une population plus importante que celle que nous étudions ici : l'analyse en composantes principales a pour principe de regrouper des variables étroitement corrélées pour dégager des axes d'analyse et elle est d'autant plus fiable qu'un grand nombre d'individus permet de minorer les comportements atypiques. Ce n'est pas le cas ici, nous l'avons déjà précisé. Les axes dégagés, et les quatre types de comportement qui en dépendent, n'ont de valeur que l'exemple qu'ils donnent des possibilités d'analyse sur un échantillon plus large, donc plus représentatif de la réalité.

**Le premier axe** dégagé par l'analyse en composantes principales oppose l'attente à l'activité : c'est l'axe du rythme de production.
**Le deuxième axe** oppose une forte proportion d'opérations en lecture et des stratégies de gonflement du déjà écrit (proportion importante d'ajouts en lecture par rapport aux ajouts en écriture) à l'absence de ces opérations : il se fonde sur l'intensité des opérations en lecture.
**Le troisième axe** oppose un taux important de "gaspillage" (mots utilisés/mots restants) et des stratégies de relecture pour correction pendant le 3ème tiers de l'écriture (le texte comporte à peu près autant de mots à la fin du 2ème tiers qu'en fin d'écriture) à un taux important d'ajouts en lecture

par rapport aux suppressions en lecture. Comme le deuxième axe, il se fonde sur les opérations en lecture, considérées ici qualitativement : elles peuvent provoquer un élagage du texte ou au contraire son expansion.

À partir de ces trois axes ont été déterminées quatre classes regroupant des individus :

- **La première classe** rassemble des individus assez productifs qui effectuent peu de suppressions : ils écrivent d'un jet, pratiquement sans retoucher le texte.
- Les individus de **la deuxième classe** se caractérisent par un taux important de retouches de leur texte, qui semble être leur mode d'écriture habituel puisque les opérations en lecture ne nécessitent pas d'attente parti culière.
- **La troisième classe** se compose de scripteurs qui travaillent dans la proximité : ils corrigent beaucoup, toujours tout près de l'endroit où ils travaillent.
- **La quatrième classe** rassemble des scripteurs lents, qui corrigent relativement peu et plutôt pour supprimer que pour ajouter du texte.

En mettant en relation ces quatre classes de scripteurs avec les textes produits, nous avons cherché à savoir si l'on pouvait établir une corrélation entre comportement scriptural et produit final.

**L'analyse discriminante** des variables concernant les textes produits par les quatre classes donne les résultats suivants :

- Les scripteurs de la classe 1 (écriture d'un jet, sans retouche) ne sont pas caractérisés par le type de texte qu'ils produisent.
- Les scripteurs de la classe 2 (forte proportion de retouches) produisent des textes où le lexique est riche et varié mais le mode d'expression peu remarquable.
- Les scripteurs de la classe 3 (gestion de proximité) nuancent leur propos en faisant varier les temps des verbes et les types de phrases employées.
- Les scripteurs de la classe 4 (scripteurs lents) structurent leurs textes en mettant en valeurs certains éléments grâce à des constructions syntaxiques de type emphatique, par exemple.

Il est difficile de tirer plus d'informations de ces résultats qui ne valent, répétons-le, que pour cet échantillon de texte. L'ambition de découvrir un lien entre mode d'écriture et texte produit ne sera pas satisfaite ici, à cause de la probabilité d'incertitude des résultats obtenus, essentiellement due à la faible représentativité des trois axes de départ (80% de la variance totale).

Suivant la démarche présentée ici, nous étudierons dès que nous en disposerons, un échantillon plus large produit selon une consigne moins ambiguë que la présente. Ceci nous permettra d'affiner notre démarche en caractérisant plus précisément les textes et leurs processus d'écriture et de travailler sur le lien pouvant exister entre eux.

Claire DOQUET Benoît FOUCAMBERT