Mise en correspondance de données textométriques et comportementales : segments répétés et jets textuels

Georgeta Cislaru[1], Thierry Olive[2]

[1] Université Sorbonne nouvelle Paris 3, Clesthia

[2] CNRS & Université de Poitiers, Centre de Recherche sur la Cognition et l'Apprentissage

Abstract

This study aims at connecting the analysis of the text as a finished product and the dynamics of text production. It thus compares two types of linguistic units: sequences that are produced spontaneously during the writing process (bursts of writing) and sequences that are recurrent in the final text (repeated segments). From a methodological point of view, this articulation questions the nature of analyzable textual units, and offers new insights on concepts like segmentation and frequency. From a theoretical point of view, it questions the place of linguistic prefabs in discourse production.

Key words: bursts of writing, repeated segments, dynamics of writing, routine, textometry.

Résumé

La présente étude se propose d'articuler des approches du texte fini et des dynamiques textuelles pendant le processus de production, en mettant en regard deux séries d'unités langagières : des séquences textuelles telles qu'elles sont produites spontanément dans le processus d'écriture (les jets textuels) et des séquences textuelles qui sont récurrentes dans le texte finalisé (les segments répétés). D'un point de vue méthodologique, cette articulation interroge la nature des unités textuelles analysables, en ouvrant de nouvelles perspectives sur les notions de segmentation et de fréquence. D'un point de vue théorique, elle questionne les stratégies de production discursive et le rôle joué par les préfabriqués langagiers.

Mots-clés : jets textuels, segments répétés, dynamique d'écriture, routine, textométrie.

1. Introduction

Les objets de l'Analyse des données textuelles (ADT) évoluent – généralement en se complexifiant – avec les conceptions théoriques du texte. La linéarité fait ainsi de la place à la réticularité et à la topographie (Mayaffre 2007), en prenant en compte les articulations et les enchaînements qui font l'unité du texte. Mais où commence le texte ? Pour les linguistes qui s'intéressent à la genèse textuelle, il est présent dès les premiers brouillons (Mahrer et Nicollier Saraillon 2014). Or, bien que devenues accessibles grâce au développement d'outils de suivi de rédaction qui enregistrent le processus d'écriture en temps réel, ces données (avant-)textuelles échappent le plus souvent aux analyses linguistiques ou textométriques. Le texte finalisé est ainsi complètement détaché de ses conditions de production linguistique, même si on s'intéresse au contexte situationnel de ces dernières. Il s'agit dès lors de se donner les moyens d'interroger la continuité du processus de textualisation, en tentant un rapprochement des deux.

L'objectif de ce travail est d'amorcer une approche longitudinale du texte, en articulant les données processuelles aux données statiques du texte finalisé. Nous mettrons en regard les séquences textuelles produites spontanément lors du processus de rédaction (les jets textuels)

et les séquences textuelles récurrentes dans le texte finalisé (les segments répétés). Plusieurs niveaux d'analyse seront pris en compte : la continuité formelle des séquences textuelles, la fréquence, les patrons morphosyntaxiques et les possibilités d'étiquetage automatique.

2. Le traitement automatique de la variation textuelle et les types de segmentation des formes

Les approches ADT prennent habituellement pour objet les textes finalisés tels qu'ils sont diffusés par différentes sources éditoriales (médias, livres, internet, discours politiques ou institutionnels, etc.). Ces textes sont sélectionnés et mis en série (par ordre chronologique, générique, thématique…) selon des questions portant sur le genre textuel, le lexique, la sémantique, les enjeux discursifs, etc.

Toute analyse d'un texte brut implique cependant une segmentation en amont (cf. Lebart et Salem, 1988 ; Mayaffre, 2014). Dans le cadre de ces approches, les données textuelles sont d'abord segmentées de manière formelle sur la base de délimiteurs graphiques, et ensuite en fonction des variations opposant des cibles à une source. On obtient ainsi i) des unités graphiques (chaînes de caractères) telles qu'elles sont répertoriées dans les dictionnaires de fréquences ; ii) des segments répétés identifiés en tant que séquences invariantes et récurrentes ; iii) des séquences équivalentes associant une langue cible à une langue source ; iv) des séquences invariantes ou, au contraire, non homogènes distinguant une version cible d'une version source ; v) des séquences isolées autour de mots-pivots ; vi) des unités catégorisées par des étiqueteurs, etc. L'annotation syntaxique ou l'annotation sémantique de haut niveau permettent de projeter sur le texte brut des catégories morphosyntaxiques, sémantiques ou discursives et d'obtenir ainsi des séquences graphiques enrichies en informations. Cette façon de procéder intègre à la segmentation des contraintes analytiques et oriente nécessairement le point de vue de l'analyste.

Lorsque la textométrie touche à d'autres champs disciplinaires, tels que la philologie, la traduction ou la critique génétique, elle gagne un nouveau terrain, qui est celui du texte au travers ses versions de traduction ou de rédaction. On assiste ainsi ces dernières années au déploiement de nouvelles méthodes d'analyse suite au développement de nouveaux outils et à l'identification de nouveaux objets textuels.

D'une part, on a vu émerger les approches contrastives multilingues, avec le développement d'outils d'alignement textuel à l'instar de MkAlign. L'angle de vue sur les textes change donc, car on s'intéresse à la manière dont des équivalences sémantiques et syntaxiques – et leurs variations éventuelles lorsqu'on a affaire à plusieurs traductions du même texte – se mettent en place dans une perspective bi- ou multilingue. La segmentation du texte s'opère à partir de principes qui mettent le repérage des formes au service de l'analyse sémantique (Fleury et Zimina, 2008).

D'autre part, on a entrepris d'observer la dynamique d'un texte au fil de sa genèse, en alignant ses versions et en identifiant les séquences homogènes et non homogènes à l'aide de logiciels tels EDITE MEDITE (Fenoglio et Ganascia, 2007) ou ALLONGOS (Lardilleux et al., 2013). Chaque nouvelle version de texte est comparée à la précédente (une colonne correspond à une version, Fig. 1 ci-après) et un code couleur permet d'identifier les ajouts (séquences nouvellement produites), les suppressions, les substitutions (une séquence vient remplacer une autre) et les déplacements.

Conclusion			
Conclusion	Conclusion	Conclusion	1 insertion
Le placement familial continue d'être	Le placement familial continue d'être	Le placement familial continue d'être	1 insertion
la *solution*	la réponse	la réponse	1 insertion 1 substitution
la mieux adaptée à la situation	la mieux adaptée à la situation	la mieux adaptée à la situation	1 insertion
de Nathalie.	de Nathalie et à ses besoins.	de Nathalie et à ses besoins.	2 insertions
Le changement de famille d'accueil intervenu	Le changement de famille d'accueil intervenu	Le changement de famille d'accueil intervenu	1 insertion
en *urgences*	en urgence	en urgence	1 insertion 1 substitution
n'a pas trop perturbé Nathalie	n'a pas trop perturbé Nathalie	n'a pas trop perturbé Nathalie	1 insertion
qui *continue pour le moment à réaliser* son projet	qui se situe toujours dans la réalisation de son projet	qui se situe toujours dans la réalisation de son projet	1 insertion 1 substitution
professionnel.	professionnel.	professionnel.	1 insertion
Le rythme	Le rythme	Le rythme	1 insertion
des accueils	de ses accueils en famille	de ses accueils en famille	2 insertions 1 substitution
en concertation	en concertation	en concertation	1 insertion
avec *la famille*	avec chacun de ses parents	avec chacun de ses parents	1 insertion 1 substitution
nous parait	nous parait	nous parait	1 insertion
être *la meilleure solution*.	être la solution la plus adaptée aux demandes de Nathalie, de Monsieur POURTOIS et de Madame	être la solution la plus adaptée aux demandes de Nathalie, de Monsieur POURTOIS et de Madame	2 insertions 1 déplacement 1 substitution
	COURCELLES..	COURCELLES.	1 insertion 1 suppression
42 insertions	13 insertions 6 déplacements 37 substitutions 4 suppressions	1 substitution 1 suppression	77 insertions 6 déplacements 38 substitutions 5 suppressions

Figure 1. Exemples d'alignement avec Allongos[1].

Le traitement automatique de la dynamique textuelle a ainsi bénéficié dernièrement d'avancées importantes (Bourdaillet et al., 2008). La critique génétique peut dès lors s'appuyer sur le calcul des insertions, remplacements et suppressions pour obtenir le tableau d'ensemble des modifications subies par un texte en devenir. La même approche a été adaptée aux versions publiées d'un texte (rééditions, cf. Mahrer et al., 2015).

Ces nouvelles approches laissent cependant de côté le temps réel de l'écriture, et la part des spéculations « après-coup » reste importante : que se passe-t-il réellement lors du processus de rédaction ? Quelles formes la production langagière prend-elle ? Quelle continuité entre les brouillons, ratures, reformulations, etc. et la version finale du texte ?

Il existe actuellement plusieurs logiciels et méthodes d'analyse du processus de configuration du texte qui donnent accès aux stratégies de production écrite en temps réel (Caporossi et Leblay, 2011 ; Wengelin, Torrance, Holmqvist, Simpson, Galbraith, Johansson, et Johansson, 2009). L'écriture enregistrée, ou écriture *online*, devient ainsi un objet d'analyse possible. Mais ces outils sont le plus souvent exploités pour répondre à des problématiques psycholinguistiques concernant la temporalité de l'écriture et ne prennent que peu en compte la dimension linguistique à proprement parler. Les derniers développements du logiciel de suivi d'écriture en temps réel Inputlog (Leijten, Van Horenbeeck et Van Waes, 2015) intègrent des outils de segmentation et d'annotation linguistique (lemmatisation, n-grams) pour l'anglais et le flamand, le module français étant en cours de développement. Il est donc pour l'instant impossible d'avoir une représentation complète et structurée des processus linguistiques propres au temps réel de l'écriture. Dans l'état actuel des choses, nous nous tournons vers les données qui sont livrées directement par les logiciels de suivi de rédaction, les jets textuels.

[1] Le texte finalisé et ses versions liminaires produites à la fin de chaque session de travail ont été anonymés.

3. Une segmentation chronologique spontanée : les jets textuels en production écrite

Dans ce travail, nous choisissons de prendre pour point de départ l'amont de la dynamique textuelle, en nous appuyant sur des données textuelles segmentées de manière spontanée au cours de la production écrite d'un texte, par des pauses signifiant des arrêts ponctuels dans la rédaction. En effet, d'un point de vue comportemental, la rédaction d'un texte peut se décrire de la façon suivante : les scripteurs alternent des moments de pauses, sans écriture, avec des périodes de transcription continue du texte[2].

Les pauses surviennent parce que les rédacteurs n'ont plus d'information pour continuer leur texte, ou parce qu'ils doivent l'évaluer ou le relire[3]. D'un point de vue cognitif, lors des pauses, les scripteurs peuvent mettre en œuvre des processus de planification (pour préparer le contenu du texte), de mise en texte (pour préparer les configurations grammaticales) ou de révision du texte déjà produit. La durée d'une pause est considérée comme reflétant la durée mais aussi la complexité du (ou des) processus cognitif(s) engagé(s) durant cette pause.

Les périodes de transcription correspondent quant à elles aux moments pendant lesquels le rédacteur transcrit son texte de façon ininterrompue. Lors des périodes de production, les scripteurs produisent des séquences textuelles, que nous appellerons « jets textuels » (le terme original anglais est "burst", Chenoweth et Hayes, 2001). Ces jets textuels peuvent prendre la forme d'une lettre, d'un mot, ou d'une séquence de mots. Par exemple, l'énoncé *une cousine qui peut venir partager du temps avec elle pendant le week-end* peut être produit sous la forme suivante :

> [pause] *une cousine qui* [pause] *peut venir partager du temps avec elle pendant* [pause] *le* [pause] *w* [pause] *eek* [pause] – [pause] *end*. [pause]

Les pauses segmentent ainsi des séquences textuelles qui constituent, d'un point de vue linguistique, des données empiriques attestées. Une durée de pause minimale de 2 secondes a été retenue pour identifier les jets textuels dans cette étude. Ce seuil étant utilisé régulièrement en psycholinguistique pour identifier les jets textuels (voir par exemple Alves et al., 2007 ; Baaijen et al., 2012 ; Chenoweth et Hayes, 2001). Les jets textuels sont donc des séquences textuelles qui rendent compte de la dynamique textuelle et qui résultent d'une segmentation spontanée, guidée cognitivement et déterminée par les propriétés fonctionnelles des processus rédactionnels de chaque rédacteur.

4. Statut cognitif et linguistique des jets textuels : état des lieux

Plusieurs travaux ont montré que la longueur et la durée des jets textuels varient selon la compétence d'écriture des scripteurs (pour une revue, voir Olive, 2014).). Par exemple, les dactylographes plus expérimentés produisent en moyenne des jets plus longs. Il en est de même pour les enfants les plus habiles à transcrire leur texte. De même, les rédacteurs ayant une plus grande expérience linguistique rédigent leurs textes avec des jets textuels plus longs. La durée et la longueur des jets textuels traduiraient alors, chez les rédacteurs les moins expérimentés, le coût de la transcription du texte et, chez les rédacteurs plus expérimentés,

[2] En moyenne, un scripteur passe la moitié de son temps de rédaction à transcrire et l'autre moitié à préparer mentalement son texte (Alves et al., 2007).

[3] Au-delà des interruptions purement mécaniques pour produire, par exemple, le point d'un « i » ou passer à un nouveau mot.

leur aptitude à simultanément préparer des segments de textes et à en transcrire d'autres et ainsi à maintenir une production fluide, adaptée au rythme de la pensée du rédacteur

Alors que la psycholinguistique s'attache à quantifier la longueur des jets textuels, ou encore le nombre d'unités actualisées en moyenne, rien ou si peu a été dit quant à la nature linguistique de ces séquences. A notre connaissance, deux recherches seulement se penchent sur la problématique linguistique des jets textuels. Dans un travail datant de 1986, Kaufer, Hayes et Flower constatent que les jets textuels ont tendance à prendre la forme des propositions, en s'arrêtant aux frontières de celles-ci, plutôt qu'aux frontières des syntagmes ou expressions. Les auteurs interprètent ces données comme une preuve que les scripteurs choisissent d'abord un thème (un topic) pour ensuite produire des parties de la proposition tout en évaluant leur cohérence grammaticale vis-à-vis des parties déjà produites. Le second travail, conduit par Olive et Cislaru (2015), reconnaît à certains jets textuels un statut d'automatismes d'écriture, et entreprend de les comparer à des éléments de routines discursives identifiés sur les textes finaux, les segments de discours répétés. Les segments répétés (SR) sont définis comme des suites d'occurrences non séparées par un délimiteur de séquence[4] dont la fréquence est supérieure ou égale à 2, dans un texte finalisé ou un corpus de textes (cf. Salem 1986 ; Lebart et Salem, 1988). Les auteurs constatent que peu de jets et de segments répétés partagent les mêmes contenus (forme et sens). Ces données soulèvent la question du statut des routines, et mettent en exergue l'intérêt que la description linguistique des jets textuels pourrait présenter pour une nouvelle heuristique.

5. Corpus d'analyse

Nous avons enregistré, à l'aide du logiciel de suivi de rédaction Inputlog (cf. Leijten et al., 2015), les sessions d'écriture d'éducateurs spécialisés rédigeant des rapports de suivi d'enfants en risque de danger. Inputlog a été installé sur des ordinateurs portables fournis à l'association SAFE de Caen. Ces ordinateurs étaient utilisés par les éducateurs spécialisés en charge de plusieurs enfants en placement familial ou en foyer afin de rédiger les rapports de synthèse ou de suivi de leur situation. Les rapports sont destinés au juge pour enfants qui s'appuie sur leur contenu afin de prendre une décision vis-à-vis de la poursuite de la mesure de protection sociale. Ils peuvent cependant être consultés par les familles depuis la loi de 2002. Ces rapports, d'une longueur allant de trois à dix pages, articulent des séquences descriptives, narratives, évaluatives et argumentatives, en suivant une trame prédéfinie. Ainsi, les rapports s'organisent en un certain nombre de sections qui portent des intitulés semblables (situation familiale, historique du placement, scolarité, santé, loisirs, relations avec les parents/la fratrie, conclusions). Il s'agit donc de textes relevant d'un genre routinier et ayant une visée performative marquée (la prise de décision). Ces deux dimensions nous semblent susceptibles de déterminer les stratégies de textualisation. Par ailleurs, le fait que les rapports sociaux ont un double destinataire (juge et famille), manifestant potentiellement des attentes opposées, complexifie leurs conditions de production et laisse supposer que le processus de rédaction peut s'inscrire dans une tension visant à ménager les deux attentes. Ces caractéristiques du corpus laissent penser que les données enregistrées présentent un intérêt particulier pour l'étude des dynamiques d'écriture. On notera également l'intérêt de travailler sur une situation de production réelle, où il est possible de contextualiser les données relevées.

[4] Un délimiteur de séquence correspond à des signes de ponctuation du type : virgule, point-virgule, deux points, tiret, guillemets, point, parenthèses…).

Dans l'idéal, les scripteurs ouvraient une session d'écriture sur l'interface Inputlog dès qu'ils intervenaient sur le texte des rapports. Au final, une fois la version définitive du rapport rédigé et le dossier « clos », les travailleurs sociaux nous livraient le dossier complet avec l'ensemble des données en temps réel : durée des sessions, chronologie du processus, révisions du texte, ensemble des séquences textuelles produites et les pauses intercalaires, etc.

Dossier	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Nbre de sessions	13	5	13	8	9	8	4	9	9	2
Nbre de mots dans rapports finaux	3223	7400	3144	3356	2841	2878	892	1239	2776	1372
Nbre de SR	796	628	1432	934	983	1060	113	110	754	436
Nbre de jets textuels	1092	441	373	949	868	563	493	394	474	62

Tableau 1. Caractéristiques générales du corpus de rapports éducatifs enregistrés en temps réel.

Neufs dossiers de rédaction ont pu être recueillis à ce stade, produits lors de 84 sessions d'écriture, pour un total de 29121 mots dans les versions finales des rapports. A l'intérieur de ces dossiers d'écriture, Inputlog a dressé une liste de 5709 jets textuels, d'une durée moyenne de 7,9 secondes. Le corpus, qui représente un volume textuel réduit[5], compte 7246 segments répétés.

6. Jets textuels et segments répétés, quelques pistes d'analyse

6.1. Pourquoi comparer des jets textuels et des segments répétés ?

Le choix de comparaison se justifie par des hypothèses d'enracinement cognitif et de reproduction idiomatique du langage, issues respectivement de la linguistique cognitive, plus particulièrement la grammaire des constructions (Langacker, 1987 ; Schmid, 2010), et des courants de la linguistique de corpus (Sinclair, 1991 ; Wray, 2002) et de la linguistique des usages (Bybee, 2010). Ces études et théories considèrent que la réitération des formes contribue à leur stabilisation dans l'usage et à l'association entre sens et forme, les rendant ainsi saillantes et accessibles pour les locuteurs. Or, comme le signalent Legallois et Tutin (2013), le domaine de la phraséologie est relativement vaste et multiforme, et beaucoup de séquences idiomatiques ne relèvent pas de « la langue », mais d'habitus générique ou thématique. Les segments répétés constituent une entrée privilégiée (même si elle n'est pas la seule) dans ce genre d'habitus, recensant les récurrences propres aux corpus étudiés, les stéréotypies discursives qui sont sous-jacentes aux textes (cf. Mayaffre, 2007 ; Née et al., 2014).

[5] Compte tenu des conditions de recueil des données d'écriture en temps réel (un processus long, qui implique, au-delà d'une relation de confiance avec les scripteurs, un suivi rapproché du fonctionnement du logiciel et un temps long d'attente pour obtenir le dossier complet des sessions d'écriture) et du fait que les analyses sont faites manuellement (cf. le détail plus bas, en 6.3 et 6.4), le corpus pouvait difficilement atteindre les volumes exploités actuellement en ADT.

Au-delà des compétences personnelles, des études récentes mettent en avant la vitesse de traitement en réception plus élevée des séquences pré-fabriquées (Conklyn et Schmitt, 2008), qui sont mémorisées et automatisées grâce à leur récurrence d'usage notamment. On peut donc se demander si les jets textuels correspondent à des unités préfabriquées de type idiomatique. Les jets textuels pourraient être affiliés à des automatismes d'écriture (cf. section 4 plus haut), indice de compétences élevées d'écriture facilitant le traitement des contenus et des formes à produire à l'écrit.

6.2. Méthodologie : mise en correspondance des listes de jets textuels et segments répétés

Nous avons comparé les jets textuels produits pendant la rédaction d'un texte et les segments répétés de discours, tels qu'identifiés dans la version finale du même texte. Les listes de jets textuels et segments répétés issus du corpus ont été alignées à l'aide d'un outil développé par A. Lardilleux. Il s'agit d'un script Python, permettant d'extraire les jets textuels et les SR à partir d'un ensemble de sessions de rédaction enregistrées avec le logiciel Inputlog, et de les comparer en les alignant dans un fichier HTML. L'alignement des jets textuels et des segments répétés permet de rapprocher les séquences des deux bords qui sont graphiquement proches (à 85%) ; ce rapprochement est intéressant pour le regroupement des données et le repérage des unités lexicales qui intègrent plus facilement les jets ou les séquences.

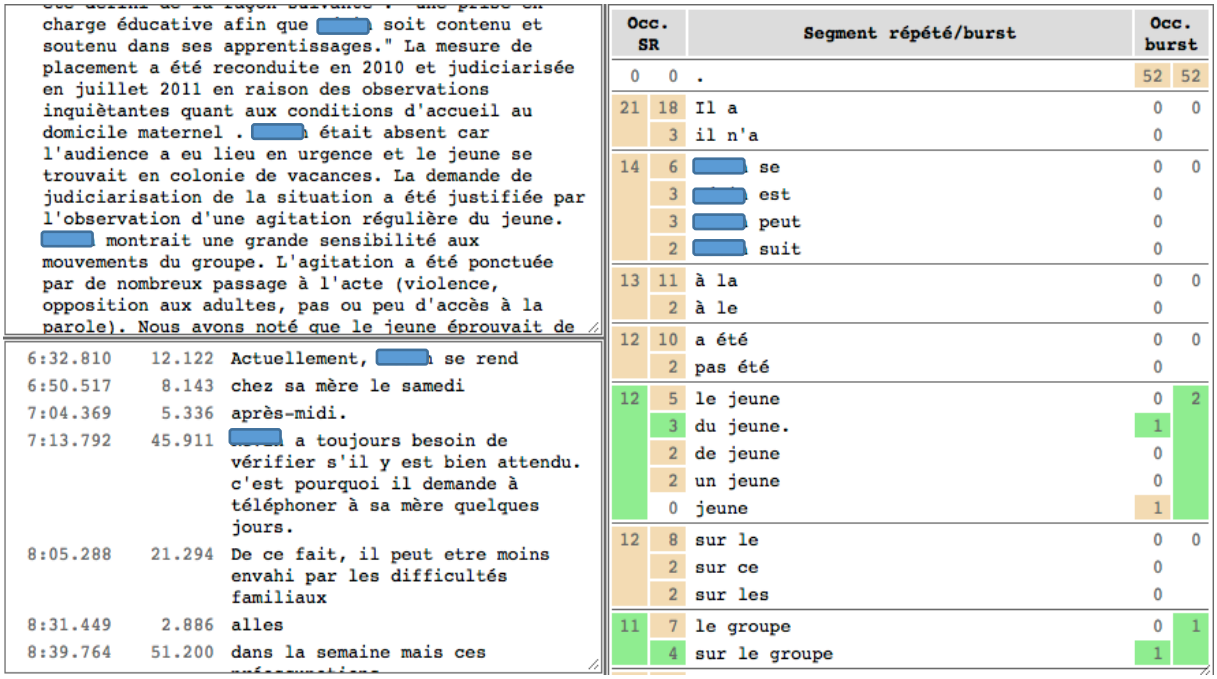

Figure 2. Alignement des jets textuels et des segments répétés[6].

La Figure 2 illustre un alignement des jets textuels et des segments répétés pour un rapport. Dans la colonne de droite, les nombres sur le bord gauche indiquent la fréquence des segments répétés, tandis que les chiffres sur le bord droit indiquent la fréquence des jets textuels. L'outil a comparé les deux listes et mis en évidence un pourcentage très limité de

[6] Les données recueillies en temps réel (i.e., tous les événements textuels et chronologiques qui ont lieu au cours d'une session Inputlog) ne peuvent pas être anonymées. Nous cachons systématiquement les données personnelles dans les documents rendus publics.

correspondances : seuls 5% des jets textuels et segments répétés partagent la même forme et contenu (en vert dans la colonne de droite). Une navigation entre la colonne de droite et des fenêtres de gauche rend possible la contextualisation des segments répétés (le texte se déroule dans la fenêtre du haut à gauche) et des jets textuels (la liste chronologique se déroule dans la fenêtre en bas à gauche).

6.3. Fréquence et nature des séquences : jets textuels vs segments répétés

Nous avons procédé, dans un premier temps, à une analyse des séquences qui sont partagées aussi bien par les jets textuels que par les segments répétés. Au-delà du taux de correspondance très bas, il nous a semblé que ce terrain de rencontre entre les deux types d'unités segmentales pouvait, d'une part, nous renseigner quant à leurs rapports et, d'autre part, nous aider à caractériser les jets textuels. La fréquence, critère définitoire des segments répétés et mécanisme sous-jacent aux automatismes du langage (et, donc, potentiellement, aux jets textuels) s'est imposée comme question liminaire.

Toutefois, bien qu'il soit possible d'évoquer la fréquence d'usage comme un préalable potentiel à l'émergence des jets textuels, la répétition n'est pas un critère pertinent pour les caractériser. Ainsi, les jets textuels qui comptent plus d'une occurrence dans le corpus correspondent le plus souvent à des graphèmes (lettres ou morphèmes du féminin ou du pluriel), à des mots-outils (*de, le, dans,* etc.) ou à des noms propres (prénom de l'enfant suivi), ce qui rend impossible tout parallèle statistique avec les segments répétés.

Les rares jets textuels polyformes qui apparaissent deux ou trois fois dans un dossier d'écriture (chez les mêmes scripteurs donc) n'ont pas toujours de segment répété correspondant, comme cela est indiqué dans le Tableau 2. La série A de jets textuels récurrents sans correspondant SR contient un délimiteur textométrique (point ou virgule), ce qui élimine d'emblée les segments répétés. La deuxième série (B) contient trois constructions préfabriquées qui renvoient aux pratiques sociales (*en milieu ouvert, les droits de visite et d'hébergement*) et aux critères d'évaluation de la situation (*son manque de travail*), ainsi qu'un jet textuel marquant la prise en charge énonciative (*nous observons*). Si chacune des catégories peut relever d'une routine justifiant l'automatisme en production, il est plus surprenant de constater l'absence de SR répété dans le texte final ; ces jets textuels ont visiblement fait l'objet de réécritures les modifiant au fil des versions. L'examen de la deuxième colonne met au jour plusieurs catégories, comme les connecteurs (*en effet, en permanence*…), les désignateurs (*Madame / Monsieur X*), les binômes *[Sujet +] Verbe être/avoir*, les séquences qui correspondent à des informations mémorisées en lien avec la situation évaluée (*cinq enfants, 12 ans*) et un segment relevant du jargon professionnel (*du groupe*, cf. Cislaru et al., 2013). Les séries 1-2 correspondent à des séquences préformées en langue (sauf lorsque le sujet est un prénom, on parlerait alors de patron préformé), tandis que les séries 3-4 correspondent à des séquences « apprises » par la répétition au cours de la mesure éducative (mais ayant des patrons morphosyntaxiques préformés : *N ans ; N enfants*).

Sans correspondant SR	Avec correspondant SR	
Série A	*Série 1*	*Série 3*
, autant	. De plus,	cinq enfants
, etc.	En effet,	(12 ans)
. Aussi	à chaque fois	
. Puis	à ce sujet	*Série 4*
. toutefois	au quotidien	Madame ChXXX
d'emploi	en permanence	Monsieur GXXXX
« , un antipsychotique »		
	Série 2	*Série 5*
Série B	il est	du groupe
Nous observons	FXXX a	de KXXX
Son manque de travail	a été	
En milieu ouvert		
Les droits de visite et d'hébergement		

Tableau 2. Jets textuels récurrents et correspondances avec les SR.

Au niveau plus global des patrons (ou des motifs, selon la terminologie de Longrée et Mellet 2013), on observe davantage de proximité entre les jets textuels et les SR. Il en est ainsi pour les constructions coordinatives, prenant les formes *X et Y, et X, X et*, où les unités coordonnées peuvent avoir des natures diverses (groupes nominaux, verbaux, prépositionnels, etc.). Mais une analyse détaillée de ces constructions, et des patrons plus généralement, nécessite un étiquetage morphosyntaxique du corpus, qui n'est pas sans poser quelques problèmes.

6.4. Saturation et traitement morphosyntaxique

L'annotation d'un corpus d'écriture enregistrée en temps réel soulève deux difficultés. Ainsi, lorsqu'il s'agit de segmenter morpho-syntaxiquement un texte bien construit, le seuil d'erreurs est plutôt bas – surtout après un entraînement sur corpus – pour des outils comme Cordial, TreeTaggeur, etc. Mais l'accès au contexte immédiat (cf. Mayaffre 2014) et à son potentiel de désambigüisation devient quasi-impossible dans le cas des jets textuels, la rédaction d'un texte n'étant pas parfaitement linéaire, mais impliquant des allers-retours à gauche et des projections-anticipations à droite. Par exemple, le jet textuel *qu'elle ne* peut être lu comme *Conj. + Pron. Pers. + NEG* ou bien comme *Pron Rel. + Pron. Pers. + NEG*, et seule une recherche « manuelle » permet de sélectionner le patron adéquat. La deuxième difficulté dépasse le cadre de l'annotation automatique et soulève la question des critères de classification morphosyntaxique des jets textuels. En effet, un nombre important de séquences relevées, qu'il s'agisse de jets textuels ou de segments répétés, sont non saturées syntaxiquement. Si la répétition permet de contourner ce problème pour les SR (voir toutefois Lebart et Salem 1988), en concentrant l'analyse sur leur fréquence, il en est autrement pour les jets textuels, qui émergent le plus souvent sous forme d'occurrences uniques. Or, 57% des jets textuels dans notre corpus sont non saturés syntaxiquement. Un décompte détaillé des types de séquences non saturées a été effectué sur la moitié des dossiers (Figure 3). A ce stade, on peut émettre l'hypothèse selon laquelle la non-saturation est un phénomène langagier à considérer, car elle jette les bases de schémas sémantiques à compléter en discours.

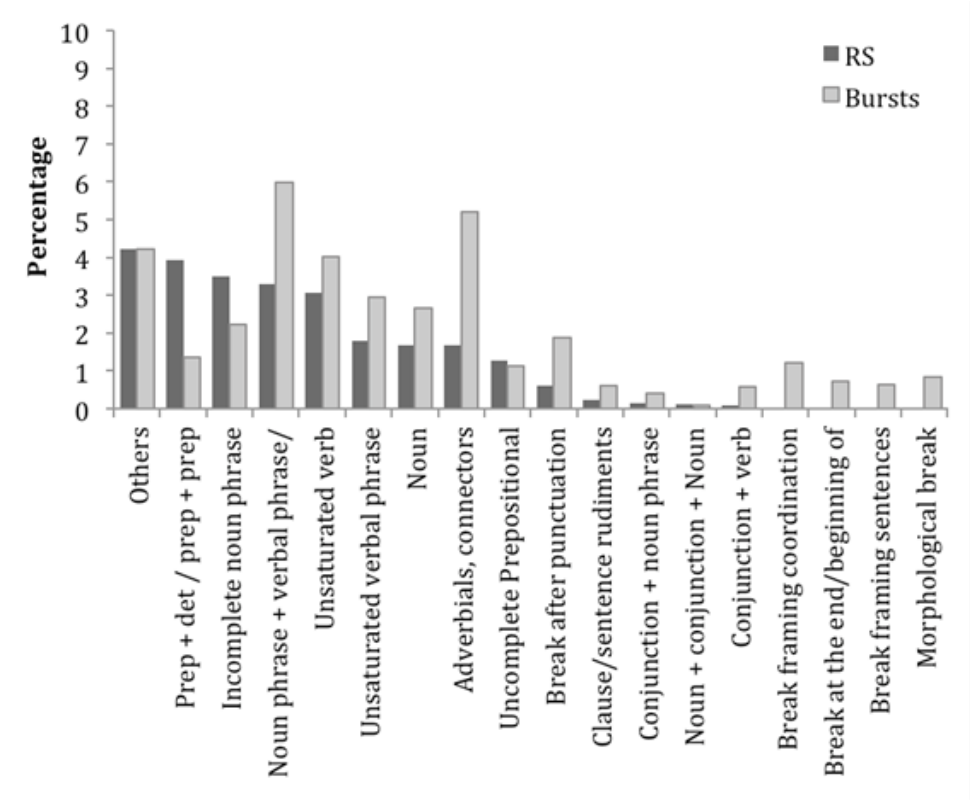

Figure 3. Pourcentage des jets textuels (bursts) et segments répétés (RS) non saturés selon les patrons morphosyntaxiques impliqués. Source : Olive et Cislaru (2015).

Lebart et Salem (1988 : 148) soulèvent la question de la saturation syntaxique des segments répétés, en signalant que des segments ne constituant pas une unité « en langue » sont mis sur le même plan que des segments correspondant à des syntagmes. Si leur regard est plutôt critique, les données obtenues par Biber (2009) sur les paquets lexicaux à l'écrit et à l'oral montrent que cette non saturation peut recevoir des justifications sémantico-discursives, permettant de compléter par des formules ad hoc, adaptées au contexte spécifique, des patrons à portée générale (cf. *peur de ne pas [pouvoir] ; les gens ne veulent pas*, Lebart et Salem, 1988 : 161-162). Mais, à la différence des paquets lexicaux identifiés par Biber (2009), dont la borne droite est non saturée, les jets textuels peuvent avoir les deux bornes non saturées. En effet, si la segmentation par la pause de 2 secondes découpe un syntagme ou une proposition en plein milieu, le premier segment aura la borne droite non saturée, tandis que le second segment aura la borde gauche non saturée, et ainsi de suite : [pause] *une cousine qui* [pause] *peut venir partager du temps avec elle pendant* [pause]. Dans ces conditions, le repérage des unités « tête » est plus complexe. Si nous retenons toujours la possibilité des patrons à ouverture à droite, nous recherchons, à l'intérieur des jets textuels non saturés syntaxiquement des unités de liage, qui pourraient constituer un noyau sémantique pertinent. A ce stade de l'analyse, nous avons pu repérer des patrons coordinatifs en ET et MAIS (*, les conflits et les dysfonctionnements fa / de la rencontrer mais nous n'avons*), des patrons structurés autour d'un connecteur (*Le travail est relativement désinsvestit. Malgré la volonté irrugièl*), ainsi que des jets organisés autour de signes de ponctuation (point, virgule : *à entretenir sa chambre. Ce manque de*). D'autres classifications sont en cours d'élaboration.

7. Conclusions

Notre approche prend en compte les jets textuels, qui constituent de nouveaux objets empiriques dans le temps réel de l'écriture, issus d'une segmentation longitudinale spontanée, et les met en regard avec les segments répétés. En cela, elle répond à deux objectifs : l'un méthodologique, visant à intégrer de nouvelles perspectives sur les données textuelles, l'autre théorique, interrogeant les stratégies de configuration textuelle. Au-delà des difficultés de mesure et de catégorisation syntaxique des jets textuels, leur analyse apparaît intéressante pour envisager la notion de fréquence et son rôle sous plusieurs angles (Ellis 2012) et montre que la dimension préfabriquée de la production linguistique se manifeste davantage au niveau des patrons structurels qu'au niveau de séquences figées spécifiques. Les patrons eux-mêmes sont non saturés syntaxiquement dans plus de la moitié des cas, laissant ainsi la place à l'actualisation d'unités contextuellement adaptées. Finalement, spontanéité et routinisation tendent à converger, les contraintes (génériques, cognitives, etc.) se manifestant au cœur de la dynamique de production textuelle.

Références

Alves R.A., Castro S.L., Sousa L. et Strömqvist S. (2007). Typing skill and pause-execution cycles in written composition. In Torrance M., Van Waes L. and Galbraith D. (éds), *Writing and cognition, research and applications*. Dordrecht: Elsevier Sciences Publishers, pages 55-65.

Baaijen V.M., Galbraith D. et de Glopper K. (2012). Keystroke Analysis: Reflections on Procedures and Measures. *Written Communication,* 29 (3): 246–277.

Biber D. (2009). A corpus-driven approach to formulaic language in English. Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics,* 14 (3): 275–311.

Bourdaillet J., Ganascia J.-G., et Fenoglio I. (2007). Machine Assisted Study of Writers' Rewriting Processes. *4th International Workshop on Natural Language Processing and Cognitive Science (NLPCS)*.

Bybee J. (2010). *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Caporossi G. et Leblay C. (2011). Online writing data representation: a graph theory approach. In Gama J. Bradley E., et Hollmén J. (éds.), *Advances in Intelligent Data Analysis X – 10th International Symposium, IDA 2011, Porto*, *Lecture Note In Computer Science,* 7014: 80-89.

Chenoweth A.N. et Hayes J.R. (2001). Fluency in writing. *Written Communication,* 18: 80–98.

Cislaru G., Sitri F. et Pugnière-Saavedra F. (2013). Figement et configuration textuelle : les segments de discours répétés dans les rapports éducatifs. In Bolly C. et Degand L. (éds), *Across the Line of Speech and Writing Variation. Corpora and Language in Use – Proceedings 2*. Louvain-la-Neuve: Presses universitaires de Louvain, pages 165-183.

Conklin K. et Schmitt N. (2008). Formulaic sequences: are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics,* 29: 72-89.

Ellis, N.C. 2012. What can we count in language, and what counts in language acquisition, cognition, and use? In *Frequency Effects in Language Learning and Processing*, ed. par Stefan Th. Gries et Dagmar Divjak. Berlin – Boston: De Gruyter Mouton. p. 7-33.

Fenoglio I. et Ganascia J-G. (2007). MEDITE : un logiciel pour l'approche comparative de documents de genèse. *Genesis* 27 : 166-168.

Fleury S. et Zimina M. (2008). Utilisations de mkAlign pour la traduction philologique. In *Actes JADT 2008, Lexicometrica* : 483-493.

Kaufer D. S., Hayes J.R. et Flower L. (1986). Composing written sentences. *Research in the Teaching of English,* 20: 121–140.

Langacker R. (1987). *Foundations of Cognitive Grammar*. Stanford: Stanford University Press.

Lardilleux A., Fleury S. et Cislaru G. (2013). Allongos: Longitudinal Alignment for the Genetic Study of Writers' Drafts. *Computational Linguistics and Intelligent Text Processing*, Springer LNCS 7817: 537-548.

Lebart L. et Salem A. (1988). *Analyse statistique des données textuelles. Questions ouvertes et lexicométrie*. Paris : Dunod.

Legallois D. et Tutin A. (éds). (2013). Vers une extension du domaine de la phraséologie. *Langages* 189.

Leijten M., Van Horenbeeck E. et Van Waes L. (2015). Analyzing writing process data: A linguistic perspective. In Cislaru G. (éd.), *Writing(s) at the crossroads: the process-product interface*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pages 277-302.

Longrée, D., Mellet, S. (2013). Le motif : une unité phraséologique englobante ? Etendre le champ de la phraséologie de la langue au discours. *Langages* 189 : 65-79.

Mahrer R., De Angelis R., Del Lungo A., Grésillon A., Lebrave J.-L., Nicollier Saraillon V., Poibeau T., Mélanie-Becquet F. et Vauthier B. (2015). Edotorial genesis. From comparing texts (product) to interpreting rewritings (process). In Cislaru G. (éd.), *Writing(s) at the Crossroads. The Process-Product Interface*. Amsterdam – Philadelphia: John Benjamins, pages 151-170.

Mahrer R. et Nicollier Saraillon V. (2014). Les brouillons font-ils texte ? Le cas des plans pré-rédactionnels de C. F. Ramuz. In Adam J.-M. (éd.), *Faire Texte. Frontières textuelles et opérations de textualisation*. Besançon: Annales littéraires de l´Université de Franche-Comté, pages 223–305.

Mayaffre D. (2007). L'analyse de données textuelles aujourd'hui : du corpus comme une urne au corpus comme un plan. Retour sur les travaux actuels de topographie/topologie textuelle (partie I). *Lexicometrica* 9. http://lexicometrica.univ-paris3.fr/numspeciaux/special9/mayaffre.pdf.

Mayaffre D. (2014). Plaidoyer en faveur de l'Analyse de Données co (n) Textuelles. Parcours cooccurrentiels dans le discours présidentiel français (1958-2014). *Lexicometrica, Actes JADT 2014* : 15-32.

Née E., Sitri F. et Veniard M. (2014). Pour une approche des routines discursives dans les écrits professionnels, *Actes du CMLF 2014*, http://www.linguistiquefrancaise.org/articles/shsconf/pdf/2014/05/shsconf_cmlf14_01195.pdf

Olive T. (2014). Toward an incremental and cascading model of writing: A review of research on writing processes coordination. *Journal of Writing Research,* 6: 173-194.

Olive T. et Cislaru G. (2015). Linguistic forms at the process-product interface: Analysing the linguistic content of bursts of production. In Cislaru G. (éd.), *Writing(s) at the crossroads: the process/product interface*. Amsterdam: John Benjamins, pages 99-123.

Salem A. (1986). Segments répétés et analyse statistique des données textuelles. *Histoire & Mesure,* 1(2): 5-28.

Schmid H.-J. (2010). Does frequency in text instantiate entrenchment in the cognitive system? In Glynn D. et Fischer K. (éds), *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches*. Berlin: Mouton de Gruyter, pages 101-133.

Sinclair J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Wengelin A., Torrance M., Holmqvist K., Simpson S., Galbraith D., Johansson V. et Johansson R. (2009). Combined eyetracking and keystroke-logging methods for studying cognitive processes in text production. *Behavior Research Methods,* 41: 337–351.

Wray A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.