Sujet clitique et dynamique de l'écrit : un éclairage par les jets textuels

Quentin Feltgen, Florence Lefeuvre, Dominique Legallois

Sujet clitique et dynamique de l'écrit : un éclairage par les jets textuels

Quentin Feltgen

Université de Gand (Belgique)

Florence Lefeuvre

Sorbonne Nouvelle, Clesthia

Dominique Legallois

Sorbonne Nouvelle, Lattice

Résumé (French abstract)

La dynamique de la production écrite peut être étudiée au moyen de logiciels d'enregistrement lors de la frappe, qui encodent le temps associé à chaque événement et permettent de restituer les séquences supprimées. De telles données sont commodément analysées par une segmentation de la séquence textuelle en unités de performance appelées *jets*, délimitées par des pauses de production excédant un seuil donné. Cependant, ces jets sont marqués par leur caractère très hétérogène. Pour saisir les enjeux linguistiques qu'ils révèlent, nous nous concentrons ici sur le rôle des pronoms clitiques sujets. En particulier, nous montrons, à l'aide d'une méthode statistique d'inspiration Monte-Carlo, que ces pronoms clitiques se retrouvent de manière marquée en début de jet. Un travail d'annotation nous permet de remarquer que ces occurrences en début de jet sont souvent précédées par une séquence de révision, ce qui suggère que les pronoms clitiques facilitent la reprise de la production. Nous soulignons en outre à travers une étude qualitative des jets textuels que l'écriture s'appuie principalement sur le couple grammatical sujet clitique-verbe pour développer des stratégies discursives qui suivent l'ordre thème-rhème habituel de la phrase en français. Ces observations concourent à montrer que les jets textuels, loin d'être aléatoires, reflètent le rôle et la fonction des unités linguistiques qu'ils articulent.

Mots-clés (French keywords): sujet clitique, dynamique de l'écrit, production textuelle, jet textuel, méthode Monte-Carlo, révisions, remplacements

Abstract

The dynamics of writing can be studied by resorting to keystroke logging softwares that record the timestamp of each event in the writing process, and keep track of the revised sequences. These data are usually framed into a sequence of *bursts* that segment the flow of events into pauses that exceed a given threshold. These bursts are, however, extremely heterogenous. To shed light on their linguistic contents and motivations, we focus on subject clitics. We show, using a Monte-Carlo-type statistical method , that these clitics are markedly found in the front position of the production bursts. By annotating these

clitics in front position, we observe that they often follow a revision event, which hints at the clitics' role in the resumption of the textual production. We also highlight based on a qualitative review of a few selected bursts that the writing hinges upon the grammatical pairing between subject clitic and verb to elaborate discourse strategies that follow the usual theme/rheme order in French sentences. These findings concur to show that the textual bursts are far from random, but reflect the function of the linguistic units that they construe into an utterance within the writing production flow.

Keywords: subject clitic, writing dynamics, written production, burst, Monte-Carlo statistics, revision, replacements

1. Introduction

Les travaux en syntaxe portent fondamentalement soit sur des structures idéalisées, c'est-à-dire des structures types dont il convient d'expliquer l'agencement, soit sur des occurrences étudiées dans des textes oraux ou écrits ; dans ce deuxième cas, l'analyse porte sur des unités produites (l'écrit) – elle ne peut pas prendre en compte le processus même de production (l'écriture) qui relève de la dynamique énonciative[1]. Certains dispositifs permettent cependant d'étudier en partie cette dynamique : c'est sur ces dispositifs que repose la présente étude, qui s'intéresse à l'écriture sous l'angle de la syntaxe.

On peut schématiquement dire que, quel que soit l'âge et la compétence des scripteurs, l'écriture n'est pas un processus fluide et homogène, mais se déroule à travers une série de "jets" successifs (appelés *bursts* en anglais[2]) entrecoupés de pauses plus ou moins longues. Ce phénomène est en partie comparable à la dysfluence à l'oral, mais s'en distingue notamment par les « retours » en arrière possibles et les révisions que permet le canal de l'écrit. L'articulation entre cette dynamique de la production et le contenu linguistique reste encore très mal comprise, même si Cislaru et Olive (2018) ont consacré un travail substantiel à cette problématique qui renouvelle assurément le regard que la linguistique peut porter sur la production écrite en temps réel.

Loin de proposer un traitement général des jets textuels, nous examinerons un seul phénomène, limité à la relation du pronom sujet clitique avec les jets. Plus précisément, nous chercherons à décrire et à caractériser la présence du sujet clitique dans les jets ou à proximité, en nous appuyant sur la significativité statistique de cette présence. Dans une première partie, nous présenterons le corpus, ainsi que la notion de *jets textuels*. La deuxième partie se concentrera sur la mise en oeuvre d'une méthode statistique pour caractériser l'emploi des pronoms clitiques dans le processus d'écriture : d'inspiration Monte-Carlo, cette méthode montre que les sujets clitiques se rencontrent avec une proportion en début de jet (c'est-à-dire immédiatement après une pause) plus importante que ce que le hasard produirait.

---

1 À part bien sûr la génétique textuelle.

2 On ne confondra pas cependant cet emploi de burst avec celui désignant une « rafale » lexicale, c'est-à-dire la concentration localisée dans un texte d'un ensemble de lexèmes (Lafon, 1981). Ni avec l'emploi plus fréquent encore désignant l'explosion lexicale dans l'acquisition d'une langue.

Pour expliquer cette affinité, nous proposons dans une troisième partie une annotation des occurrences en début de jet. Celle-ci montre qu'une part importante de ces sujets (20 %) est produite immédiatement après une révision. Nous avancerons alors l'hypothèse que les sujets servent de repères dans le processus de production textuelle, ce que nous évaluerons à partir d'un test statistique simple. Enfin, pour aller au-delà de cette vue essentiellement quantitative du rôle du sujet clitique dans la production textuelle, nous nous pencherons dans une dernière partie sur un petit nombre d'exemples choisis où s'illustrent les logiques de la production et la manière dont les sujets clitiques y interviennent[3].

2. Présentation du corpus et de la méthode statistique

Après avoir rappelé ce qui est entendu par *jet textuel*, nous exposerons les corpus utilisés pour cette étude.

2.1. Notion de jets textuels

Un jet textuel est une séquence linguistique produite entre deux pauses[4] dans le processus d'écriture en temps réel[5]. Il s'agit donc d'une unité de performance définie par les temps de pause, qui correspond rarement à une unité grammaticale (mot, syntagme, proposition, etc.). On considère généralement que l'article de Kaufer, Hayes et Flower (1986) a été le premier à aborder ce phénomène, en analysant les écrits de rédacteurs experts : ceux-ci produisaient des segments d'environ neuf mots en moyenne, séparés par des pauses de plus de deux secondes. De plus, ces scripteurs professionnels écrivaient plus de mots par jet qu'un panel d'étudiants. Comme le rappellent Alves et Limpo (2015), la plupart des modèles psycholinguistiques actuels semblent s'accorder sur le fait que quatre processus cognitifs sont à la base du processus d'écriture :

– le processus de planification qui détermine l'organisation des idées ;

– le processus de « traduction » qui convertit les idées en formes linguistiques ;

– le processus de « transcription » qui s'appuie sur l'orthographe et l'écriture manuscrite (ou dactylographique) pour produire l'expression sous la forme d'un texte écrit ;

– le processus de révision qui contrôle, évalue et modifie le texte jusqu'à l'aboutissement au texte définitif.

C'est par exemple grâce à un processus de « traduction » plus développé que les experts produisent plus de mots que les rédacteurs novices. Par ailleurs, l'efficience du processus de planification donne du crédit à l'hypothèse que certains jets répondent plus ou moins à une « logique » ou stratégie de textualisation, qui relève d'une compétence linguistique et rédactionnelle.

Dans la mesure où ils sont interprétables (ce qui n'est pas toujours le cas), les jets textuels sont donc des processus d'écriture investis dans la production dynamique des textes.

---

[3] Ce travail s'inscrit dans le projet ANR Pro-TEXT (N° ANR-18-CE23-0024-01).

[4] C'est-à-dire deux interruptions du processus de production considérées comme signifiantes ; voir *infra* (§2.2) pour la définition opératoire utilisée.

[5] Définition de Cislaru et Olive (2018).

2.2. Corpus utilisé

Notre étude repose sur un corpus de jets textuels collecté dans un contexte expérimental auprès de 83 étudiants de licence de psychologie[6]. Chaque participant a produit un texte au clavier, lors d'une session d'enregistrement de 15 minutes. Le texte consiste en un court essai sur un thème donné, supposé familier des scripteurs (place du tabac à l'université, contraception, sécurité routière, etc.). Les données associées à la production tapuscrite (temps de frappe, caractères produits, clics, etc.) ont été récupérées à l'aide du logiciel Inputlog (Leijten, Van Waes, 2013). Ces données consistent en une liste d'événements (touche produite, clic, délétion, etc.), et des temps associés (début et fin de l'événement ; pour une touche, il s'agit ainsi du temps de presse et du temps de relâche). Chaque texte :

– comprend en moyenne 3 600 événements, avec un minimum de 1 600 pour le plus court et de 7 000 pour le plus long ;

– est composé de 157 à 919 formes lexicales, selon les cas, pour une moyenne de 460 formes ;

– est structuré en 79 jets en moyenne, avec un minimum de 38 et un maximum de 155.

Il convient de souligner que le texte enregistré, puisqu'il garde la trace de l'ensemble du processus de production, et notamment les nombreuses suppressions et révisions, ne correspond pas au texte final. C'est toute la richesse – et toute la difficulté – de ce type de données. Cela suppose de marquer explicitement ces événements de suppression, ce que nous faisons au moyen du caractère ⌫. Comprendre ce qui a été effacé, et ce qui a remplacé le texte supprimé, suppose alors une reconstruction en suivant l'ordre linéaire de la production, ce qu'illustre l'exemple suivant (la production d'une espace est également explicitement marquée au moyen du caractère ␣) :

[1] Substance␣cis⌫⌫onsidéré␣⌫e␣illicite,␣elle␣est⌫⌫⌫⌫⌫⌫
⌫⌫⌫⌫␣n⌫dans␣notre␣pays,␣elle␣est␣pourtant␣acceptée␣ailleurs.

On peut lire, d'abord, que le « is » suivant le « c » de « cis » a été effacé (deux événements de délétion), puis que « onsidéré » a été tapé pour écrire « considéré ». L'espace suivant le mot a ensuite été effacée pour ajouter immédiatement le « e » voulu par l'accord du participe avec « Substance », initialement omis. Plus loin, dix caractères ont été effacés, ce qui correspond à la séquence « ,␣elle␣est », puis « dans notre pays » a été produit. Le texte final serait donc: « Substance considérée illicite dans notre pays, elle est pourtant acceptée ailleurs. »

Pour structurer ensuite le processus de production en jets textuels, nous avons considéré l'intervalle temporel entre deux événements consécutifs, souvent appelé IKI (pour Inter-Key Interval) dans la littérature (Conijn et al., 2019). La définition des jets suppose qu'au-delà d'un certain seuil, cet intervalle peut être considéré comme significatif et représenter alors une pause dans le processus de production. Bien souvent, ce seuil est défini à 2 secondes, pour des raisons de reproductibilité

[6] Pour une présentation de ce corpus, voir (Bouriga, 2020) et (Olive, Bouriga, 2022).

(Wengelin, 2006) ; ici, pour s'adapter aux possibles variations interindividuelles, nous avons conservé cette référence de 2 secondes, et calculé le quantile correspondant dans l'ensemble des données produites[7]. En reportant ce quantile dans les distributions individuelles, nous avons ainsi pu déterminer un seuil correspondant, individualisé. Cette procédure se justifie d'autant mieux que la population des participants s'avère relativement homogène (niveaux académiques et âges comparables).

L'exemple ci-dessus se compose par exemple de deux jets complets, une pause intervenant à la suite de « pourtant␣ » et avant « acceptée␣ailleurs.␣ », de même qu'avant « Substance » et après « ailleurs.␣ », ce que l'on peut rendre à l'aide du symbole « // » :

[1b] //Substance␣cis⌫⌫onsidéré␣⌫e␣illicite,␣elle␣est⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫␣n⌫dans␣notre␣pays,␣elle␣est␣pourtant␣//acceptée␣ailleurs.␣//

Le déséquilibre entre ces deux jets, en termes de contenu et de taille, est évident, ce qui montre bien le caractère disparate et hétérogène de ces objets. Par ailleurs, l'exemple est précédé et se termine par une pause ; on voit par là que les frontières linguistiques traditionnelles, correspondant à une segmentation en phrases, peuvent se retrouver dans la segmentation de la production en jets textuels.

La nature linguistique de ces jets textuels pose cependant question. En effet, leur caractère éminemment divers pourrait laisser penser qu'ils relèvent de trop de variables pour refléter le processus cognitif de la formulation linguistique. Pour sonder la sensibilité de ces jets au contenu linguistique qu'ils permettent de produire, une possibilité consiste à examiner la position relative de certaines unités linguistiques choisies vis-à-vis des jets – en particulier, si ces unités se retrouvent plutôt intégrées au jet, ou se révèlent au contraire liminaires, apparaissant préférentiellement en fin ou en début de jet. L'idée est alors que, si certaines unités se comportent d'une manière spécifique vis-à-vis des jets, alors ces jets sont effectivement motivés, au moins partiellement, par des considérations d'ordre linguistique. Dans cette étude, nous avons fait le choix (pour des raisons qui seront explicitées *infra* en 3.1) de nous intéresser à un type d'unité linguistique bien particulier, à savoir les sujets clitiques.

3. Répartition des occurrences de sujets clitiques à travers les jets

Après avoir rappelé les propriétés essentielles des sujets clitiques, nous présenterons les motivations théoriques qui justifient notre intérêt pour la position relative de ces pronoms par rapport aux jets textuels. Pour estimer la significativité des observations, il est nécessaire de mettre en œuvre une méthode statistique

[7] La méthode présentée ici a été proposée pour le traitement des jets par Olive et Bouriga (2022).

spécifique, d'inspiration Monte-Carlo. Les résultats de cette comparaison sont ensuite présentés et discutés.

3.1. Les sujets clitiques

Le terme de clitique renvoie à son origine à une notion phonétique, utilisée dans l'étude des langues tels que le latin ou le grec pour des mots inaccentués qui se trouvaient rattachés au mot précédent ou au mot suivant. Pour le français, la notion de clitique correspond à une notion syntaxique liée à la place occupée par le pronom vis-à-vis du verbe conjugué. D'autres expressions surviennent dans la littérature avec des sens voisins de celui de clitique : forme atone (Le Goffic, 1993 : 110), conjointe, faible (Foulet, 1990 : 107), non autonome (Léard, 1992 : 219), pourvue d'une prédicativité faible (Moignet, 1981). Ces pronoms clitiques (en ce qui concerne cet article les pronoms personnels *je, tu, il, ils*) s'opposent aux pronoms non clitiques *moi, toi, lui, eux. Elle, elles, nous, vous,* ne varient pas dans leurs formes en tant que clitiques ou non clitiques. *On*, pronom indéfini ou personnel, n'a pas de forme forte disjointe correspondante. Dans le corpus, composé de textes expositifs, la troisième personne domine très largement. Nous considérons ici les formes *il* et *ils* (sans nous limiter à ces pronoms par la suite) pour tester leurs propriétés, en nous basant sur ces exemples :

[2a] il␣est␣au⌫⌫également␣pratique␣car␣il␣

[3a] ils␣permettent␣une␣mobilité␣au␣sein␣des␣villes␣et␣⌫⌫⌫

qui donnent dans le texte final :

[2b] il est également pratique

[3b] ils permettent une mobilité au sein des villes

Les pronoms personnels clitiques se caractérisent par une contrainte forte qui est d'être joints à un verbe conjugué. Cette contrainte se décline en plusieurs propriétés (cf. Lefeuvre (2006) pour ce récapitulatif et, plus récemment, la GGF en IX-3 ainsi qu'Encyclogram). Les pronoms clitiques ne peuvent pas apparaître de façon isolée :

[4a] *il !

[5a] *ils !

ni composer un segment détachable :

[2c] *il, est également pratique

[3c] *ils, permettent une mobilité au sein des villes

contrairement aux pronoms personnels non clitiques :

[4b] Lui !

[5b] Eux !

[2d] Lui, est également pratique

[3d] Eux, permettent une mobilité au sein des villes

Dans ces deux derniers exemples, les pronoms personnels non clitiques apparaissent détachés notamment lorsqu'ils se trouvent en situation de contraste.

Les pronoms personnels clitiques sujets ne peuvent pas être élargis par un modifieur :

[2e] *il-même est également pratique

[3e] *ils-mêmes permettent une mobilité au sein des villes

[2f] Lui-même est également pratique

[3f] Eux-mêmes permettent une mobilité au sein des villes

ni par une coordination :

[2g] *il et les autres sont également pratiques

[3g] *ils et les autres permettent une mobilité au sein des villes

[2h] Lui-même et les autres sont également pratiques

[3h] Eux-mêmes et les autres permettent une mobilité au sein des villes

Ils ne peuvent pas non plus intégrer entre eux et le verbe conjugué un adverbe ou groupe prépositionnel d'énonciation :

[2i] *il aussi est également pratique

[3i] *ils aussi permettent une mobilité au sein des villes

[2j] Lui aussi est également pratique

[3j] Eux aussi permettent une mobilité au sein des villes

Les clitiques n'occupent forcément la même position dans la chaîne, par exemple avec les SN sujets postposés :

[6a] A gauche se dressait un haut buffet.

[6b] *A gauche se dressait il.

L'impersonnel *il* ainsi que le pronom *on* (pronom personnel selon Riegel et *al*. (2009), pronom indéfini selon la GGF en IX-7.1) partagent les propriétés de ces pronoms personnels clitiques sujets.

Ainsi on note, sur le plan grammatical, une forte cohésion entre le pronom sujet clitique (pronoms personnels et *on* en ce qui concerne cet article) et le verbe conjugué. L'étude des jets montrera si leur composition dépend de propriétés linguistiques, comme celles que nous venons de mettre en évidence.

À ces considérations grammaticales qui montrent l'intérêt de choisir pour notre étude les pronoms clitiques, s'ajoute le fait que ces formes présentent l'avantage pratique de pouvoir être extraites automatiquement à partir du texte brut. Elles sont en effet facilement identifiables sur la seule base de leur forme de surface, par contraste avec le reste des sujets possibles dont la détection devrait s'appuyer sur une annotation préalable du texte – annotation rendue difficile de par le caractère hétérogène des segments produits lors de la production. L'extraction automatique des pronoms clitiques n'est pas entièrement exempte de faux positifs (notamment dans le cas où leur forme coïncide avec leur équivalent non-clitique), mais ceux-ci demeurent marginaux et peuvent être écartés manuellement.

Voyons à présent comment s'établit l'articulation entre les jets textuels et les pronoms clitiques.

3.2. Articulation entre jets textuels et pronoms clitiques : la position relative des occurrences

Dans cette section nous abordons les enjeux théoriques concerant la position d'une unité linguistique par rapport à un jet textuel et nous présentons les cinq positions possibles.

3.2.1. Enjeux théoriques

Notre hypothèse de travail consiste à considérer la position relative des pronoms clitiques ou plus généralement des unités linguistiques vis-à-vis des jets comme un marqueur de la relation entre le processus de production, caractérisé par ses alternances de pauses et de jets textuels, et les propriétés linguistiques des unités produites. L'idée est ici que, si les unités linguistiques ne sont pas réparties au hasard à travers les jets, mais qu'au contraire cette répartition présente des déviations significatives par rapport à une simple distribution aléatoire, alors les jets sont sensibles aux propriétés des unités qu'ils contiennent – et le processus de production du texte est sensible à la teneur linguistique de ce qui est écrit. Cette méthode a notamment été employée (Cislaru, Olive, 2018) pour mieux saisir le rôle de certaines unités dans le processus de production, en particulier les connecteurs (*en effet, toutefois*, etc.), les conjonctions (en particulier *et*), les possessifs, et les démonstratifs.

Cette approche, centrée sur des unités linguistiques, contraste avec une seconde approche, qui consiste à partir des jets et à tenter d'en décrire le contenu (Cislaru, Olive, 2018 ; Gilquin, 2020). Ces deux approches sont bien entendu complémentaires, et répondent à des questions de recherche différentes ; la première approche s'intéresse au rôle joué par certaines unités linguistiques choisies dans le processus de production, les jets étant conçus comme une fenêtre épistémologique sur ce dernier, la seconde informe quant au contenu des jets textuels, dont on présuppose alors la pertinence.

Par rapport à cette première approche, que nous choisissons dans le cadre de ce travail, un autre point doit être précisé : notre hypothèse ne s'appuie pas sur un cadre théorique d'interprétation nettement défini. De manière générale, les unités linguistiques peuvent manifester une préférence (ou une réticence) à occuper les positions en début de jet, en fin de jet, ou en milieu de jet (notions que nous définirons formellement ci-dessous). Supposons que l'on observe une préférence pour la position "début". Deux interprétations peuvent alors s'opposer : d'une part, on peut considérer que l'unité linguistique favorise, après la pause, la reprise de la production. D'autre part, la pause peut traduire une difficulté à produire l'unité linguistique qu'elle précède. C'est donc l'ambiguïté théorique de la pause elle-même qui suscite les difficultés de l'interprétation : considérée comme un « indicateur équivoque » (Foulin, 1995) dont la fonction est « mal spécifiée » (Alves, Castro & Olive 2008), elle a été considérée comme pouvant aussi bien réfléter la saturation de la capacité cognitive ou l'expression d'un engorgement (Just & Carpenter 1992, Torrance & Galbraith 2006), que la difficulté de retrouver un contenu lexical donné (Kircher et al. 2004).

Quoique ces deux interprétations de la préférence pour le début des jets textuels paraissent s'opposer, elles sont probablement l'une et l'autre valides, la pause pouvant tout à fait manifester une diversité hétérogène de processus cognitifs relatifs à la

production langagière. Ainsi, une pause avant un terme spécifique est probablement liée à la difficulté de se le remémorer, comme dans l'exemple suivant, où les pauses (et donc, les frontières des jets) sont indiquées par le symbole //:

[7] //⌫⌫,␣ce␣qui␣//découragerait␣certains␣consommateurs␣à␣fumer␣ou␣bien␣à␣réduire␣considérablement␣//leur//␣habitude␣//fumatoire.␣//

La difficulté de rendre lexicalement le concept voulu est rendue manifeste par l'emploi de fumatoire, qui constitue, sinon un néologisme, du moins un terme peu fréquent ; la pause paraît donc bien suscitée par le contenu du jet en aval, dont l'articulation linguistique fait problème.

Par contraste, la conjonction *et* apparaît souvent en début de jet, et semble alors jouer le rôle de béquille pour la reprise de la production. C'est notamment le cas ici :

[8] //C'est␣une␣décision␣arbitraire␣qui␣touche␣directement␣les␣droits␣//des␣e⌫élèves//⌫⌫⌫⌫souhaite//␣et␣qui␣compore⌫te␣un␣soucis␣éthique//

Ici, l'interprétation du *et* comme reprise est d'autant plus claire qu'en réalité, la conjonction suit immédiatement "des élèves" dans la linéarité du texte. Il n'y donc a pas ici une simple pause, mais un retour en arrière dans le texte, une occurrence de *veut* antérieurement produite étant supprimée et remplacée par souhaite : entre les deux termes reliés par la conjonction s'intercale donc un contenu qui n'est pas lié au contexte immédiat de la production. Le jet commençant par *et* doit donc assurer la reprise de la production après une révision sur un segment antérieur qui ne s'articule pas directement avec le contexte immédiat.

En conclusion, l'interprétation de la position relative des occurrences vis-à-vis des jets n'est pas univoque et nécessite un examen plus approfondi du contexte. Quoiqu'il en soit, du moment que la répartition observée dévie de ce qui serait explicable par une distribution aléatoire, cela signifie que d'autres facteurs entrent en compte, en particulier les propriétés linguistiques des éléments constituant les jets textuels : cette déviation devient ainsi le signe du jeu qui se noue entre le processus de production et les unités linguistiques. Dès lors, elle cautionne l'attention portée au phénomène ainsi mis en évidence.

3.2.2. Définition des positions prises en compte

Une occurrence peut occuper cinq positions relativement au jet de production dans lequel elle s'inscrit : elle peut se trouver en début de jet, en milieu de jet, en fin de jet, mais aussi constituer le seul contenu du jet (occurrence dite "seule"), ou encore, se retrouver à cheval sur plusieurs jets (occurrence dite "scindée"). Formellement, ces cinq positions se définissent comme suit:

1) position "début" : si les premiers caractères alphabétiques du jet textuel coïncident avec les caractères composant l'occurrence. Celle-ci peut alors être précédée par des caractères non-alphabétiques (espace, ponctuation, symboles de révision).

2) position “fin” : si les derniers caractères alphabétiques du jet coïncident avec les caractères composant l’occurrence.

3) position “seule” : si les seuls caractères alphabétiques du jet sont les caractères composant l’occurrence.

4) position “scindée” : si les caractères alphabétiques composant l’occurrence n’appartiennent pas tous au même jet.

5) position “milieu” : tous les autres cas.

Ainsi, dans l’exemple 7, *leur*, *habitude*, *fumatoire* sont des occurrences seules, *ce* et *découragerait* sont des occurrences en début de jet, *qui* et *considérablement* sont des occurrences en fin de jet, et tous les autres mots sont en milieu de jet. Si l’on considère toutes les occurrences de syntagme nominal, alors l’occurrence du syntagme *leur habitude fumatoire* est une occurrence scindée (elle s’étale sur trois jets successifs).

Nous abordons à présent la méthode statistique qui nous permettra d’évaluer dans quelle mesure le comportement des occurrences enregistrées est un indicateur possible de la sensibilité linguistique des jets.

3.3. Méthode statistique

Pour évaluer la significativité statistique de nos observations, il convient de comparer la valeur de l’observation à une distribution aléatoire sur les valeurs possibles de l’observable. Pour ce faire, nous avons opté pour une méthode d’inspiration Monte-Carlo consistant à générer un nombre important de segmentations aléatoires et alternatives du texte, l’observable étant mesuré pour chacune des segmentations afin de construire la distribution correspondante. Cette méthode a été exposée en détail lors d’un travail antérieur (Feltgen et al., 2022), auquel nous renvoyons le lecteur pour toutes les précisions techniques.

3.3.1. Principe général

Considérons un texte donné (texte non pas au sens du produit final, mais au sens de la succession, linéaire dans le temps, de caractères produits, caractérisé par une segmentation en jets. Les occurrences d’intérêt (dans notre cas, les sujets clitiques), se distribuent dans le texte. Pour chacune de ces occurrences, on peut alors déterminer sa position dans le jet, en fonction des définitions fournies en 3.2.2. L’observable d’intérêt, pour ce texte, est la proportion d’occurrences dans chacune des cinq positions définies.

Le principe de la méthode est alors de considérer, pour le même texte, une segmentation alternative, générée aléatoirement. Le texte (c’est-à-dire la séquence de caractères produits) reste inchangé, mais les frontières des jets se répartissent différemment. Dès lors, les positions des occurrences relativement aux jets sont appelées à changer. Nous générons ainsi 20 000 segmentations alternatives et aléatoires du texte, suivant une probabilité génératrice dont nous détaillerons la structure dans la section suivante. L’observable est alors mesuré sur chacune de ces segmentations afin de construire la distribution attendue pour ces valeurs.

Pour évaluer la significativité de l’observation effectuée sur la segmentation réelle du texte, il suffit alors de comparer la valeur mesurée pour la segmentation réelle à la distribution. On peut alors obtenir la valeur *p*, c’est-à-dire la probabilité d’observer

une valeur plus extrême. Le seuil de significativité est fixé à 0,01 (0,05/5, car nous effectuons 5 observations en parallèle, ce qui augmente la probabilité d'un faux positif) Si la valeur *p* est inférieure à ce seuil, alors l'observation dévie suffisamment des valeurs attendues suivant le hasard pour pouvoir être regardée comme pertinente, c'est-à-dire, dans notre cas, associée à des logiques afférentes à la fonction linguistique de l'unité considérée.

3.3.2. Probabilité génératrice des jets

Notre méthode repose essentiellement sur la génération de segmentations alternatives aléatoires du texte. Pour générer de telles segmentations, il est possible de redistribuer aléatoirement les frontières des jets. Cependant, une telle approche n'est pas sans poser problème. Ainsi, les frontières des jets ne se retrouvent qu'exceptionnellement à l'intérieur des mots, alors que cela constitue la majorité des positions possibles. Par ailleurs, la segmentation textuelle (phrases, paragraphes), dessinée par la ponctuation, impacte la segmentation en jets de production, dans le sens où la fin d'une phrase coïncide fréquemment avec la fin d'un jet. Cet aspect est tout particulièrement crucial dans le cas des sujets, qui figurent généralement en tête des phrases. Si l'on ne tient pas compte de l'impact du découpage en phrases sur le rythme de la production, alors l'observation que les sujets se retrouvent en tête de jet se révélera inévitablement significative – alors même qu'elle pourrait n'être qu'un corollaire de l'association fin de phrase/fin de jet.

La probabilité génératrice des jets doit donc prendre en compte un certain nombre de facteurs. Nous en avons défini 5 : un facteur "baseline", correspondant à une probabilité constante de s'arrêter après chaque événement de frappe ; un facteur "entre deux mots" ; un facteur "après une ponctuation faible (virgule, parenthèse, tiret)", un facteur "après une ponctuation forte (point, point d'exclamation, point d'interrogation)", et enfin un facteur "avant révision". En effet, les révisions sont souvent précédées d'une étape de relecture d'un segment antérieurement produit, ce qui peut induire une pause dans la production de la séquence des événements clavier.

L'impact de ces cinq facteurs est estimé à partir d'une régression multinomiale effectuée sur la segmentation réelle du texte. Formellement, la variable modélisée est un vecteur binomial Y dont la taille est égale au nombre d'événements constituant le texte, et qui encode 1 si l'événement est suivi d'une pause, 0 sinon. Cette variable est modélisée par la transformation logistique du produit βX, β encodant le poids de chacun des cinq facteurs, X étant une matrice qui, pour chaque facteur, pour chaque événement, encode 1 si le facteur s'applique (ex. si l'événement est une ponctuation faible), 0 sinon. Cette matrice X est définie par le texte et les poids β sont obtenus par régression de la variable Y originale. Pour générer une variable Y alternative et aléatoire, il suffit alors d'effectuer un tirage aléatoire de la probabilité donnée par la transformation logistique du produit βX, qui définit une probabilité locale d'occurrence de pause après chaque événement de la séquence produite.

3.3.3. Distribution au niveau du groupe

Jusqu'alors, nous avons défini l'observable au niveau du texte, cet observable se définissant comme la proportion d'occurrences de sujets clitiques trouvées pour chacune des cinq positions définies. Il nous faut maintenant considérer ce résultat au

niveau du groupe. Pour cela, un changement de la définition de l'observable est nécessaire : il s'agit désormais de la moyenne de ces ratios sur l'ensemble des participants. Pour pouvoir construire la distribution correspondante, il faut donc générer aléatoirement des corpus alternatifs et calculer la moyenne sur chacun, ce qui ne présente pas de difficulté technique additionnelle particulière.

3.4. Résultats

L'ensemble du corpus comprend 848 occurrences de sujets clitiques, et leur répartition à travers les jets est récapitulée dans le Tableau 1. La répartition observée de ces occurrences montre une concentration des sujets en position "milieu" (70%, soit 594 occurrences), ainsi qu'une présence marquée en position "début" (24%, soit 193 occurrences), alors que la position "fin" n'est pratiquement pas occupée (3%, 25 occurrences). Les deux autres positions étant marginalement représentées, nous les laisserons de côté dans la suite. Cette observation est cependant à nuancer : en effet, il y a beaucoup plus de chances de figurer en milieu de jet qu'en bordure de jet, puisque les jets comprennent typiquement plusieurs mots. De fait, le ratio de 70% correspond à ce que l'on pourrait attendre si les jets étaient aléatoirement structurés suivant l'hypothèse nulle décrite plus haut : même si cette position reste majoritaire, elle n'est pas anormalement représentée compte tenu des attentes.

Autrement plus remarquable, on note une forte préférence des sujets clitiques pour la position "début". Certes, une nouvelle phrase s'accompagne souvent d'un nouveau jet, et les sujets clitiques, étant fréquemment en début de phrase, sont donc plus attendus en cette position. C'est ce que reflète la distribution aléatoire, qui tient compte de ce facteur explicatif : on attend en moyenne de 14 à 20% de sujets en début de jet (soit entre 119 et 170 occurrences), contre seulement de 6 à 10 % en fin de jet. Cependant, le ratio réellement observé se révèle bien supérieur, se situant au-delà de cinq écart-types au-dessus de la moyenne attendue. Inversement, la proportion d'occurrences en fin de jet est nettement plus faible qu'attendu, ce qui témoigne de la cohésion très nette entre le clitique et le verbe qu'il introduit.

Si la faible proportion de clitiques en fin de jet s'explique facilement compte tenu des propriétés lexico-grammaticales de ces unités, leur prévalence en début de jet pose question et ne permet pas d'interprétation directe. C'est la raison pour laquelle nous nous sommes penchés sur ces occurrences, afin de dégager quelques clefs de compréhension sur le rôle de ces unités dans le processus de production.

Position	Observation	Intervalle de confiance	Valeur *p*
début	24%	[14% - 20%]	0.0001
milieu	70%	[68% - 75%]	0.35
fin	3%	[6% - 10%]	0.0001
seule	2%	[0% - 3%]	0.43
scindée	0.01 %	[0.03% - 3%]	0.0001

Tableau 1 : statistiques Monte-Carlo de la répartition des occurrences des sujets clitiques à travers le jet ; pour chaque position, nous donnons la valeur observée pour

le corpus, l'intervalle de confiance associé à la distribution de cette valeur sur l'ensemble de nos corpus alternatifs et aléatoires, et la significativité de l'observation vis-à-vis de cette distribution.

4. Annotation manuelle des occurrences en début de jet textuel

Les pronoms clitiques sujets manifestent une nette préférence pour la position "en début de jet", que leur positionnement fréquent en tête de phrase ne permet pas seul d'expliquer. Pour comprendre ce comportement, nous avons donc annoté les 193 occurrences de pronoms clitiques sujets figurant en tête de jet pour les répartir entre différentes catégories et comprendre la logique de ces occurrences.

4.1. Catégories utilisées

Nous avons annoté plusieurs cas de figure. La première possibilité concerne une révision : le pronom personnel suit des marques de révision :

> [9] //Rese⌫te␣a␣voir␣si␣ces␣brouilleurs␣ont␣aussi␣un␣effet␣sur␣la␣connexion␣sal⌫tellite␣des␣téléphones,␣//si␣les␣⌫⌫⌫⌫//**ils**␣ot⌫nt␣aussi␣la␣capacité␣d'interdé⌫⌫férer␣avec␣//⌫⌫'//les␣envois␣e⌫de␣messages␣ou␣appels␣téléphi⌫oniques␣des␣étudiants.//

Notons ici que l'élément précédant le clitique après révision (ici, il s'agit de *si*, puisque *les*␣ est supprimé) pourrait amener à catégoriser le clitique d'après son rôle dans *Reste à voir s'**ils** ont la capacité d'interférer avec les envois de messages ou appels téléphoniques des étudiants*, c'est-à-dire comme intervenant immédiatement après un subordonnant. Ceci est néanmoins contraire à notre approche, où nous exploitons les jets de production pour approcher de l'opération cognitive qui précède la production du sujet clitique et introduit la pause. Dans la linéarité de ce processus cognitif, et contrairement à la linéarité textuelle du produit final, *si* ne précède pas *ils*, et c'est pourquoi nous distinguons les événements de révision.

Une deuxième possibilité concerne un jet précédent clos par un point ou la juxtaposition d'une proposition :

> [10] Cepn⌫endant,␣nous␣constatons␣un␣⌫e␣augmentation␣permanent␣⌫e␣de⌫␣⌫u␣nombre␣de␣fumeurs␣//**en␣France.␣**//**Nous**␣constatons␣également␣que␣l'âge␣auquel//␣//les␣gens␣commencent␣à␣fumr⌫er␣est␣de␣plus␣en␣plus␣précoce.
>
> [11] //N'étant␣pas␣a⌫"accro",␣ils␣peuvent␣être␣dégou⌫ûtés␣par␣les␣images␣chocs//,␣**ils**␣peuvent␣⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫␣vont␣y␣être␣plus␣sensible,

Si ces deux cas de figure paraissent équivalents sur le plan linguistique, nous noterons ici que, dans le premier cas, ces occurrences sont bien expliquées par l'hypothèse nulle, qui prévoit une frontière de jet probable à l'issue d'une phrase. Pour les secondes, en revanche, la frontière entre propositions juxtaposées n'est pas matérialisée par un signe détectable de manière univoque ; celles-ci peuvent donc contribuer au surplus de sujets clitiques en début de jets.

Une troisième possibilité prend en compte différentes catégories syntaxiques se trouvant devant le pronom sujet clitique. Peuvent se trouver dans cette position des connecteurs :

[12] **//Par␣contre**␣//**on**␣peut␣aussi␣pebnse⌫⌫⌫⌫nser␣que␣cette␣mesure␣est␣vaine␣étant␣donné␣//qu'on␣trouvera␣toujours␣un␣moyen,␣⌫⌫␣de␣se␣distraire␣où␣//de␣se␣connecter␣//

des cadratifs :

[13] **De␣nos␣jours**,␣//**il**␣est␣impor⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫nous␣avons␣accés␣à␣de␣nombreux␣transport␣⌫s␣//en␣communs.//

des cadratifs qui prennent la forme de propositions subordonnées :

[14] //En␣effet,␣**lorsque␣vous␣serez␣en␣cours**//,␣**il**␣n'y␣aura␣aucun␣mio⌫⌫oyen␣//de␣pouvoir␣se␣"divertir".␣Exposons␣les␣pours␣et⌫t␣les␣ovn⌫⌫⌫contres//

Dans une quatrième possibilité, le pronom clitique survient après un subordonnant :

[15] cette␣question␣divise␣l'opinion␣publique//:␣certains␣pesne⌫⌫⌫ne⌫se␣qu'elle␣st⌫⌫est␣nécessaire␣et␣d'autres␣pensent␣au␣contraire␣**qu'**//**elle**␣n'est␣pas␣utile.//

Nous avons également relevé les cas de figure correspondant à des cas de rupture de la linéarité de la production (le jet reprend à un endroit du texte qui n'est pas contigu au jet précédemment produit) que nous avons notés par la lettre X, et que nous avons exclus de l'analyse.

4.2. Répartition statistique

La répartition des occurrences suivant les catégories listées plus haut est affichée Figure 1. Comme prévu, un nombre important d'occurrences sujets en début de jet s'explique par le passage d'une phrase à une autre (40%, soit 78 occurrences). Les occurrences en début de jet dans un contexte de juxtaposition de deux propositions indépendantes sont au nombre de 14 ; elles ne sont pas prises en compte en tant que telles dans le cadre de l'hypothèse nulle et peuvent donc expliquer partiellement le surplus d'occurrences de jets, mais leur nombre reste cependant trop limité pour que ce facteur puisse suffire à rendre compte du ratio observé.

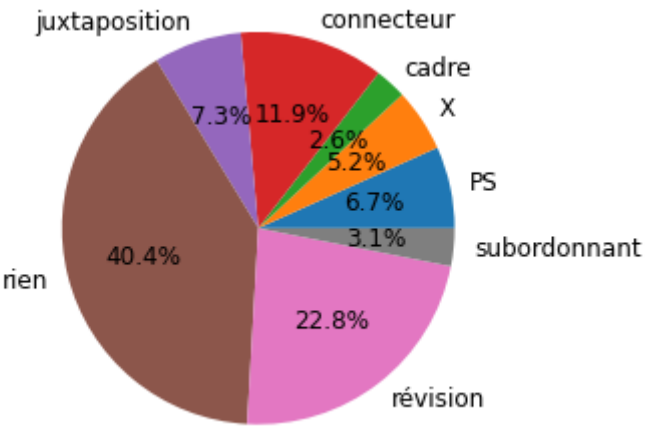

Figure 1 : Catégorisation des occurrences de pronoms clitiques en début de jet d'après une annotation manuelle

On remarquera en outre qu'une large fraction d'occurrences est précédée par un élément de contextualisation : connecteur, cadre, ou proposition subordonnée (21 % soit 41 occurrences au total). En ce cas, le sujet clitique initie encore, à travers un nouveau jet de production, la proposition proprement dite. Les sujets clitiques peuvent également être précédés par un subordonnant, mais il s'agit d'une situation peu fréquente.

Enfin, une fraction importante d'occurrences (23%, soit 48 occurrences) est immédiatement précédée d'événements de révision : c'est-à-dire qu'un contenu linguistique, produit mais non retenu, s'interpose entre le sujet clitique et ce qui le précède dans le texte final. Cette forte proportion d'occurrences immédiatement précédées par une révision est à souligner ; elle suggère que le sujet clitique (et sans doute le sujet en général) sert de pivot dans la production : lorsqu'il y a révision du texte, on en revient plus souvent à une position sujet, et c'est à partir de ce repère rédactionnel que la production peut reprendre. C'est l'hypothèse que nous testerons dans la section suivante.

4.3. Rôle des révisions

Notre annotation des occurrences nous a permis de remarquer une proportion importante de révisions précédant immédiatement la production d'un sujet clitique. Pour savoir s'il s'agit d'une spécificité de cette unité linguistique, nous allons d'abord évaluer la significativité de la présence du sujet suite à une révision. De là, nous allons interroger son rôle dans les séquences remplacées, partant de l'hypothèse que les remplacements se font de manière privilégiée en remontant jusqu'au clitique. Enfin, nous proposerons une analyse qualitative de ces séquences.

4.3.1. Le sujet comme faisant immédiatement suite à une révision

Pour s'assurer de la tendance des sujets à participer d'un phénomène de reprise de la production après une révision, nous avons cherché à évaluer statistiquement l'association entre sujets clitiques et événements de révision. Par événement de

révision, nous entendons ici toute séquence continue d'utilisation du caractère d'effacement, ⌫. Par exemple, la délétion de « il␣est␣impor » (cf. exemple 13), comprenant 12 caractères, implique une séquence de 12 caractères d'effacement consécutifs, qui correspondent donc à un événement de révision unique. Nous considérons ensuite le premier mot après un événement de révision, adoptant ici une définition particulièrement lâche d'un mot comme une suite consécutive de caractères alphabétiques, jusqu'à interruption par un événement ne correspondant pas à un caractère alphabétique (caractère d'espacement, de ponctuation, clic de souris, pression des touches directionnelles, utilisation de la touche Control, etc.). Le corpus comprend 39 489 de ces mots dont 892 sont des sujets clitiques ; 6 644 interviennent après un événement de révision (avec éventuellement un caractère d'espacement interposé), dont 135 sujets clitiques. Ceci permet de mesurer la sur-représentation ou non des sujets clitiques après révision via un test exact de Fisher. Il s'avère que ce résultat n'est pas significatif ($p = 0.19$).

En effet, nombre d'événements de révision sont des événements à caractère orthographique (p. ex. suppression d'une lettre en cours de mot pour corriger une faute de frappe). Pour remédier à ce biais, nous avons donc considéré les événements de révision correspondant à des séquences comprenant plusieurs délétions consécutives. Le nombre d'événements décroît en conséquence, ainsi que le nombre de sujets clitiques trouvés après de tels événements de révision ; mais la significativité augmente, ainsi qu'il est visible d'après le Tableau 2.

Longueur des séquences	≥ 1	≥ 2	≥ 3	≥ 4	≥ 5	≥ 6	≥ 7	≥ 8
Nombre de révisions	6644	3753	2473	1815	1452	1189	1002	877
Révisions avant sujet	135	97	76	62	50	42	41	39
p (Test de Fisher)	0.19	0.17	0.006	0.002	0.003	0.005	0.0003	7e-5

Tableau 2 : Nombre de révisions correspondant à des séquences supérieures ou égales à une certaine longueur, nombre de révisions total et avant sujet clitique parmi ces séquences, et significativité de l'association entre sujets et révisions d'après le test exact de Fisher.

Pour des séquences de longueur au moins 4, la présence de sujets clitiques est tout à fait significative ($p = 0.002$). Or, si l'on considère la longueur médiane des mots tels que nous les avons définis, il apparaît que cette médiane est de 4 caractères ; les événements de révision d'une longueur d'au moins 4 caractères correspondent donc approximativement à la délétion d'un mot ou plus, et reflètent bien le phénomène qui nous intéresse.

Le rôle privilégié des sujets clitiques comme unité produite après une révision (au niveau morphosyntaxique plutôt qu'au niveau orthographique) est donc bien confirmé.

4.3.2. Le sujet comme repère lors du processus de révision

Une autre question, proche de la précédente, consiste à s'interroger sur le rôle de balise que peut jouer le sujet lors d'un processus de révision ; puisque le processus de production reprend de manière privilégiée par un sujet après une révision, il est légitime de s'attendre à ce que, lors d'une révision, on efface la séquence précédemment produite jusqu'à atteindre le sujet. Pour confirmer cette hypothèse, nous avons donc considéré toutes les séquences de mots effacées par une séquence de révision d'une longueur d'au moins 10 caractères, et retenu, pour chacune de ces séquences, le dernier mot effacé (chronologiquement le premier mot qui avait été produit). On trouve 449 de ces séquences dans notre corpus ; parmi elles, 22 ont pour premier mot un sujet clitique. L'association entre les deux est nettement significative ($p = 0.001$). Le sujet clitique (à défaut de pouvoir tester cette hypothèse pour tout type de sujet) correspond donc bien à un repère pertinent dans le processus d'écriture.

Ces résultats se révèlent remarquablement similaires aux observations effectuées sur les phénomènes de rectification à l'oral, lesquelles impliquent généralement une redite du syntagme dans son entier (Blanche-Benveniste, 2003). Dans ce contexte, le rôle du pronom clitique comme "borne initiale" de l'unité verbale, reprise donc lors d'une rectification du syntagme, avait notamment été souligné (Blanche-Benveniste, 2010 : 79).

4.3.3. Révision et reformulation du sujet

Il est intéressant de considérer, pour ces 22 occurrences correspondant à l'effacement d'une succession de mots jusqu'à un pronom clitique, si la nouvelle formule retenue présente ou non un pronom clitique, et si ce n'est pas le cas, par quoi ce pronom clitique se trouve remplacé. Considérant la taille réduite de notre échantillon, une étude quantitative est bien sûr impossible, aussi proposons-nous un aperçu qualitatif de ces exemples.

Dans trois cas, l'enchaînement des révisions et des ré-écritures rend le contexte global trop obscur pour démêler par quoi le segment amorcé par le pronom clitique est remplacé ; dans le seul exemple qui reste déchiffrable, la séquence « Nous pouvons aussi tr », commencée par un sujet clitique, est effacée, puis remplacée par « Et sans com », séquence à son tour effacée et remplacée par « Cela pourrait aussi poser des problèmes de sécurité. »

Dans un seul cas, on observe une explicitation du référent (dans tous les exemples qui suivent, la séquence supprimée par la révision est indiquée en gras) :

[16] //**il ␣ circule**// ⌫ ⌫ ⌫ ⌫ ⌫ ⌫ ⌫ ⌫ ⌫ ⌫ le ␣ produit ␣ circule ␣ beaucoup␣dans␣le␣pays.␣//

Dans trois cas, la nouvelle structure reste identique à la structure remplacée, mais le clitique est modifié, voire rectifié ; par exemple :

[17] [La prise de stupéfiants débuterait sûrement plus jeunes] ,␣ comme␣la␣prise␣de␣cigarettes,␣puisqu'**elle␣pourrait␣**//⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫ils␣pourraient␣être␣achee⌫tés␣par␣tous.␣ //

L'hésitation ici reflète la tension entre *stupéfiants*, le référent de *ils*, qui est bien le sujet de la phrase *pourraient être achetés par tous*, et *la prise*, sujet de *débuterait sûrement plus jeunes* et donc référent attendu du pronom clitique produit immédiatement après.

Dans un autre cas, le clitique est repris pratiquement tel quel, mais il y a changement de connecteur, *donc* étant remplacé par *pour finir* :

[18] //**On␣peut␣donc␣en␣**⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫ Pour ␣ finir, ␣ on ␣ peut ␣ conclure ␣ que ␣ le ␣ //cannabis ␣ et ␣ sa ␣ consommation␣posent␣un␣large␣problème␣de␣santé␣publique␣

On observe également un cas où le clitique est repris, le référent reste le même, mais le propos exprimé est sensiblement différent :

[19] le␣cannabis␣à␣⌫⌫a//␣un␣usage␣ré//créatif,␣**il␣est␣utilisé**// ⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫il␣est␣donc␣relativement␣méconnu␣ quant␣aux␣maus⌫x␣qu'il␣induit␣

Dans les treize autres cas (pour sept d'entre eux associés à un emploi impersonnel du pronom clitique), l'effacement de la séquence s'accompagne d'un changement de stratégie discursive qui peut :

– conserver l'essentiel des éléments produits en prenant comme sujet un terme nominal :

[20] //**il␣a␣été␣décidé␣de**⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫ ⌫cette␣mesure␣a␣été␣décidé

– reprendre un élément introduit dans la séquence effacée, avec ici la substitution d'un marqueur de discours par un autre, formellement et fonctionnellement très similaire (Fagard, Charolles, 2018) :

[21] **Il␣est␣d'ailleurs**//⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫Par ␣ailleurs,␣le␣s⌫⌫s␣étudiants␣boursiers␣peuvent␣demander␣une␣ remboursement␣des␣frais␣d'inscrption␣dans␣certains␣établissemnt.␣

– faire intervenir un sujet plus étoffé :

[22] Pour␣cela,␣**il␣faudrait␣**//⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫le␣ premier␣argument␣en␣faveur␣//

– ou aboutir, dans un seul cas, à une délition du sujet dans la nouvelle séquence :

[23] [Si l’on rendait ce type de transports gratuits, seraient-ils d’aussi bonne qualité?]//Le␣p//⌫⌫⌫⌫**Il␣sembler**⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫ Non,␣pro//bablement␣pas,␣car␣plus␣de␣personnes␣fréquenteraient␣le␣réseau␣de␣bus,␣

On le voit, il n’y a donc pas de correspondance directe entre le fait, après une révision, de reprendre la production avec un pronom clitique (tendance dont nous avons vu qu’elle était significativement marquée), et le fait, lors d’une révision, de remonter jusqu’au pronom sujet clitique (tendance là encore significativement marquée) puisque, dans la plupart des cas, la nouvelle séquence n’a que peu à voir, en termes de structure, avec la séquence supprimée. La diversité des stratégies discursives qui font suite à une telle délétion, et que nous n’avons pu détailler, constitue par ailleurs un marqueur du rôle pivot que joue le sujet dans la production : le propos s’articule bien à partir de là, et c’est à partir de lui que la proposition prend forme.

5. Quelques phénomènes connexes associés au sujet clitique

L’analyse de la position relative des occurrences de sujets clitiques dans les jets textuels nous a permis d’expliquer le recours régulier en début de jet de ces unités linguistiques, celles-ci permettant d’amorcer la production, et servant de balises dans le processus : les séquences de révision remontent préférentiellement jusqu’au clitique pour changer de stratégie discursive, et la production après révision reprend de manière significativement fréquente avec un pronom clitique.

Nous souhaitons compléter maintenant cette analyse par deux ensembles d’observations qui nous permettront d’éclairer le rôle du pronom clitique. Le premier ensemble a pour objet le couple sujet clitique-verbe en tant qu’il permet de déployer des stratégies discursives pour présenter un constituant rhématique. Nous examinons notamment ci-dessous comment le sujet clitique peut être reformulé par un sujet rhématique ; indépendamment de la segmentation en jets textuels, cela permet de mieux cerner la fonction du sujet clitique dans la séquence énonciative. Quant au second ensemble d’observations, il a pour but de dresser un aperçu général des occurrences de sujets clitiques en milieu de jet : même si nous avons vu que leur présence en cette position n’était pas significativement remarquable, celle-ci rend compte néanmoins de la majorité des occurrences, aussi apparaissait-il nécessaire d’en faire cas et d’expliciter les comportements observés de manière récurrente.

5.1. Sujet clitique et constituant rhématique

Peut-on donner une raison autre que grammaticale à la fréquence importante du couple sujet clitique-verbe ? Plusieurs stratégies se mettent en place pour présenter un constituant rhématique. À partir du pronom sujet clitique, peuvent se déployer des stratégies informationnelles. Ainsi, le sujet impersonnel *il* (*il existe*) permet de présenter un rhème dans une séquence, constituée par un GN indéfini ou un infinitif qui pourraient jouer le rôle d’un sujet grammatical (*de constater que* P) :

[24] afin␣de␣voir␣s' **il␣existe␣**// rék⌫ellement␣des␣effets␣positifs␣sur␣les␣élèves

[25] //L⌫**Il␣est␣norla⌫⌫mal␣de␣constater␣que**␣les␣étud⌫⌫⌫⌫individus␣non-fule=⌫⌫⌫meurs//␣sont␣dérangés␣par␣la␣fuméé⌫e␣des␣cigarettes//,␣et␣que␣cela␣puise⌫se␣nuire␣à␣leur␣santé␣de␣manière␣indirecte⌫e.

comme dans :

[24b] Des effets positifs sur les élèves existent réellement (GN sujet)

[25b] De constater que les individus non-fumeurs sont dérangés par la fumée des cigarettes est normal. (Infinitif sujet)

Le sujet clitique apparaît comme une espèce d'indice de départ à partir duquel s'échelonnent les informations données dans la phrase. La séquence est parfois ressentie comme le véritable sujet, ce qui peut expliquer des phénomènes d'accord :

[26] //⌫à␣et␣ **il␣existent⌫⌫⌫e**␣des␣boites␣pouir⌫⌫r␣cahc⌫⌫cher␣ceux-ci.//

Ici corrigé (⌫⌫⌫), *existent* est réécrit en *existe*.

Le pronom personnel sujet *nous* qui dénote l'énonciateur permet de mettre en avant, dans une complétive, un constituant rhématique qui correspond ici à l'idée défendue dans le texte :

[27] //Alor⌫⌫⌫⌫Pour␣conclure,␣**nous␣pouvons␣dire␣que**␣comme␣chaque␣choses␣⌫⌫␣il␣y␣a␣des␣bons␣et␣des␣mauvais␣points,␣mais␣je␣ope⌫⌫⌫pense//␣que␣chacuns␣peut␣y␣trouver␣son␣compte␣al⌫vec␣l'id⌫⌫⌫es␣fumoirs,␣//

On, qui renvoie à un énonciateur plus vague, permet également d'introduire un constituant rhématique :

[28] En␣effet,␣**on␣peut␣voir**//␣//des␣publicités␣a⌫à␣la␣télévision␣qui␣sensibilisent␣la␣population␣à␣//ne␣pas␣prendre␣la␣voiture␣en␣ayant␣bu␣de␣l'aco⌫⌫lcool␣//

Néanmoins, d'autres structures concurrentielles peuvent se mettre en place au détriment des pronoms personnels clitiques. Il existe ainsi des GN indéfinis qui occupent la fonction sujet (*Des campagnes de publicité*) dans un énoncé qui pourrait être analysé comme thétique (cf. Kuroda 1973) :

[29] On␣peut␣aussi␣voir␣un␣autre␣risque␣qui␣est␣le␣téléphone␣au␣volant,␣en␣effet␣celà⌫a␣est␣très␣dang⌫gereux␣car␣le␣

conducteur ␣ n'est ␣ plus ␣ concentré ␣ sur ␣ a ⌫ la ␣ route. ␣ // ␣ //**Des ␣ campagnes␣de␣pubilci⌫⌫⌫⌫licités␣**//sont␣toujours␣présenten⌫s ␣pour␣sensibiliser␣les␣gens␣à␣ne␣pas␣utiliser␣leur␣téléphone␣au␣ volant.

Le sujet clitique se trouve parfois reformulé à l'aide d'un GN sujet postposé. La présence d'un cadratif peut amener à une recomposition de l'ordre argumental de la phrase, en conduisant à l'inversion du sujet, comme dans cet exemple :

[30] //⌫⌫Dans␣un␣pri⌫emier␣temps,␣//sur␣les␣paquets␣de␣ cigarret⌫⌫⌫ettes␣**nous␣retrouv⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫sont␣ visibles␣des␣e⌫messages␣//tel␣que␣"fumer␣tue⌫e"**//

Le sujet clitique laisse la place à un GN sujet rhématique : « des messages tel que "fumer tue" ».

On voit ainsi que l'écriture s'appuie principalement sur le couple grammatical sujet clitique-verbe pour développer des stratégies discursives qui suivent l'ordre thème-rhème habituel de la phrase en français (Le Goffic, 1993), le sujet clitique représentant un thème minimal. D'autres stratégies discursives peuvent toutefois intervenir.

5.2. Occurrences de sujets clitiques à l'intérieur des jets

Nous avons pu voir que le sujet clitique manifestait une préférence pour la position en début de jet, avec presque un quart des occurrences dans cette position, une proportion qui s'éloigne très clairement de ce qui serait attendu à partir d'une segmentation aléatoire, et nous nous sommes efforcés d'expliquer cette préférence marquée par le rôle que peut jouer le sujet (ici, uniquement clitique) à l'égard du processus de production. Cependant, il n'en reste pas moins vrai que 70% des occurrences de sujet clitique se trouvent en milieu de jet, ce qui constitue une large majorité ; notre étude serait donc incomplète sans considérer, ne serait-ce que brièvement et de façon qualitative, les principaux cas de figure rencontrés pour cette position en milieu de jet.

Bien souvent, le pronom clitique situé à l'intérieur du sujet n'est en fait précédé par rien d'autre qu'un élément de contextualisation (cadratif, etc.) :

[31] //Dans␣le␣cadre␣d'une␣expr⌫érience␣**on**␣ma␣demandé␣de␣ choisir␣un␣theme␣//et␣de␣f⌫⌫⌫⌫⌫⌫⌫//afin␣de␣debatr⌫tre␣ des␣argument␣spou␣⌫r␣⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫ents␣pour␣et␣contre ␣a␣son␣sujet.

Cet exemple a par ailleurs ceci d'intéressant qu'il s'agit d'une présentation réflexive de la tâche par l'un des participants.

Dans de nombreux cas où le clitique est sujet d'une subordonnée, l'élément précédent dans le jet se résume au seul subordonnant :

[32] cette ␣ mesure ␣ est ␣ vaine ␣ étant ␣ donné ␣ //**qu'on** ␣ trouvera ␣ toujours␣un␣moyen,␣⌧⌧␣de␣se␣distraire␣où␣//de␣se␣connecter␣//

On remarquera que ces deux cas de figure sont très proches de certains exemples en début de jet annotés : il ressort en effet que les cadratifs et les subordonnants peuvent se retrouver indifféremment de l'un ou l'autre côté de la frontière du jet. De même, on retrouve des sujets clitiques faisant immédiatement suite à un événement de révision pouvant s'inscrire dans les limites du jet :

[33] //Pourtant␣⌧,␣o⌧comme␣tou⌧⌧⌧⌧⌧⌧⌧⌧⌧⌧**il**␣ne␣faut ␣pas␣oublier␣que␣cette␣plante␣est␣aussi␣ue⌧ne␣drogue␣

Dans tous ces cas de figure, le pronom clitique reste un élément liminaire, même si d'autres éléments peuvent s'intercaler entre l'occurrence et la frontière gauche du jet. En revanche, on trouve plusieurs exemples où l'intégration du pronom clitique à l'intérieur du jet textuel est plus complète, généralement lorsque le pronom clitique intervient dans une séquence formulaire (Wray, Perkins, 2000) ou préfabriquée (Bolinger, 1976) tels que *nous allons voir que*, *il est vrai que*, etc. :

[34] //␣il␣est␣donc␣céces⌧⌧⌧⌧⌧nécessaier⌧⌧e⌧re␣de␣ prevenir␣des␣risuq⌧⌧ques␣que␣le␣tabac␣provoque␣su␣⌧r␣la␣ santé␣ds␣⌧⌧es␣personnes␣mais␣**il␣ne␣faut␣pa␣n⌧⌧s**␣non␣plus␣ le␣diaboliser.//

En outre, le clitique intégré au jet fait souvent partie d'une subordonnée :

[35] //Ce␣qui␣fait␣**qu'il**␣est␣donc␣difficile,␣//même␣ap//⌧⌧avec ␣//de␣bonnes␣campagnes␣anti//-tabas//,␣//de␣réussir␣à␣arrêter␣//et␣ stopper␣cela.␣//

On notera, dans ce dernier exemple, le caractère haché de la production après le premier jet, ce qui permet par contraste de faire ressortir le caractère cohésif du premier jet. Et là encore, l'occurrence participe d'un fabriqué (*il est difficile de*), et illustre alors la fonction des séquences formulaires qui assurent la continuité de la production (Wood, 2002) et constituent ainsi une part conséquente de la production langagière (Erman, Warren, 2000). Tous les exemples ne relèvent toutefois pas de ce cas et certaines attestations de subordonnées en milieu de jet ne relèvent pas du registre formulaire :

[36] //L⌧Un␣aur⌧tre␣problme␣⌧⌧⌧eme␣du␣preservatif␣est␣ **qu'il**␣est␣fabriqué␣en␣latex,␣ce␣qui␣le␣rend␣unulit⌧si⌧⌧⌧⌧⌧ ⌧⌧inutilisable␣pour␣les␣personnes␣allergiques␣a⌧à␣cette␣man⌧ tiere.␣//

Enfin, certaines occurrences de sujet clitique, qui introduisent une proposition principale et qui ne constituent pas des formules préfabriquées, peuvent se retrouver

en milieu de jet. Celles-ci sont rares (on en compte tout au plus une dizaine parmi les quelques 600 occurrences en milieu de jet), mais elles existent néanmoins. Dans un cas, il s'agit d'une reprise anaphorique dont on notera par ailleurs l'impressionnante continuité au sein d'un unique jet de production (qui se poursuit encore ensuite) :

[37] //vous ␣ avez ␣ fait ␣ des ␣ études ␣ supérieures, ␣ **vous** ␣ n'êtes ␣ surement␣pas␣passé␣à␣côté␣du␣nombre␣plus␣que␣conséquent␣de␣ fumeurs␣que␣l'on␣peut␣appercevoir␣sur␣les␣marc⌫⌫⌫⌫⌫⌫⌫⌫ ⌫⌫⌫⌫devant␣les␣marches␣des␣universités

Dans quelques cas exceptionnels, plusieurs propositions peuvent s'enchaîner dans le même jet :

[38] //les␣études␣ne␣montrent␣pour␣l'heure␣aucun␣⌫⌫⌫⌫⌫⌫ que␣peu␣de␣risques␣⌫,␣mais␣comme␣pour␣le␣tabac,␣s'il␣doit␣y␣ avoir␣de␣⌫s␣dégats␣causés␣par␣le␣"vapotage",␣**ils**␣ne␣seront␣peut ␣être␣visibles␣que␣dans␣//quelques␣années

Enfin il est certains exemples pour lesquels l'interprétation nous fait défaut, par exemple lorsqu'un jet chevauche deux phrases sans en recouvrir aucune, comme dans l'exemple suivant (là encore, la proposition initiée par le clitique présente un caractère formulaire) :

[39] le␣gouvernement␣n'e⌫hes⌫⌫ésite␣pas␣à␣monter␣un␣peu␣ plus␣chaque␣année␣le␣prix␣du␣tb⌫abac␣pour␣ainsi␣insiter␣les␣ gens␣//à␣arréter⌫â//␣leur␣consommation␣abusives␣du␣tabac.␣Mais␣ **nous**␣pouvons␣voir␣que␣cal␣⌫⌫⌫ela␣n'empêche␣âs␣⌫⌫⌫pas␣ au␣consommateurs␣de␣o⌫ci⌫ontinuer␣//à␣fumer.␣//

Cela illustre bien que les quelques considérations que nous avons développées ici n'épuisent pas la diversité des phénomènes manifestés par les jets textuels.

La position d'un sujet clitique en milieu de jet correspond néanmoins, pour une large majorité des occurrences, à l'une de ces trois situations que nous avons relevées : le sujet est immédiatement précédé par un cadratif, lequel ouvre le jet ; il peut être intégré dans une proposition subordonnée ; il peut être constitutif d'un élément préfabriqué (p. ex. *il faut que*). Cela suggère en creux que, lorsque le pronom clitique est référentiel et sujet de la proposition principale, sa position privilégiée est bel et bien en tête des jets (ne se laissant précéder pour l'essentiel que par un cadratif éventuel).

6. Conclusion

Après avoir présenté le corpus et la notion de jets textuels, nous avons expliqué la mise en œuvre d'une méthode statistique pour caractériser l'emploi des pronoms clitiques vis-à-vis du processus de production : d'inspiration Monte-Carlo, cette

méthode met en évidence que les sujets clitiques surviennent avec une proportion en début de jet plus importante que ce que le hasard produirait. En nous appuyant sur une annotation des occurrences en début de jet, nous avons montré qu'une part importante de ces sujets clitiques (20 %) est produite immédiatement après une révision : ceux-ci constituent ainsi des repères dans le processus de production textuelle, et jouent le rôle de balises à partir desquelles peut s'opérer un changement de stratégie énonciative. Quelques exemples choisis ont illustré plus avant les logiques de la production et la manière dont les sujets clitiques y interviennent. Deux cas ont été présentés : l'apport informationnel du couple sujet clitique-verbe qui permet d'introduire à sa suite un élément rhématique, et les occurrences en milieu de jet, dont le caractère souvent formulaire permet l'intégration dans le flux de production.

La question se pose désormais de savoir dans quelle mesure les phénomènes observés pour les sujets clitiques relèvent de leur fonction de sujet, ou s'avèrent spécifiques de leur qualité de clitique. On peut s'attendre à des différences ; notamment, la cohésion entre le sujet non clitique et le verbe est moins marquée sur le plan formel, et la sélection lexicale peut notamment s'effectuer en deux temps (on devrait donc observer plus d'occurrences en fin de jet) ; en outre, les référents explicites se prêtent moins au registre formulaire (ce qui devrait être associé à un nombre plus faible d'occurrences en milieu de jet). En revanche, on peut avancer l'hypothèse que les sujets gardent un rôle privilégié dans l'alternance des phases de production et de révision, comme on a pu l'entrevoir d'ailleurs en observant les séquences réécrites après effacement des sujets clitiques, et qui impliquaient souvent un sujet plus explicite. Cette dernière remarque laisse entrevoir la possibilité d'un coût cognitif moindre pour les pronoms clitiques, qui, par l'économie qu'ils représentent, peuvent alors jouer un rôle favorable au déroulé de la production, et en faciliter la fluidité. Ainsi, les jets contenant un sujet clitique sont plus longs qu'attendus (ils se retrouvent en moyenne dans le 77ème quantile, contre un intervalle de confiance situé entre les quantiles 70 et 74 d'après notre méthode Monte-Carlo).

Il est par ailleurs possible d'envisager que le rôle de repère joué par les pronoms clitiques lors des processus de révision soit en réalité lié à leur cohésion au verbe, lequel serait alors l'unité pertinente dans les phénomènes de réécriture. Cependant, les remplacements que nous avons observés peuvent substituer un sujet plein à un clitique, laissant entendre que le jeu paradigmatique de révision opère bien sur cette position syntaxique.

L'étude des sujets clitiques montre surtout que les jets textuels, malgré leur caractère hétérogène et à première vue difficile à appréhender, présentent une sensibilité marquée au contenu linguistique ; en cela, les jets textuels offrent bien une perspective empirique sur les processus d'articulation linguistique du contenu informationnel dont ils constituent la traduction et la trace. Par ailleurs, les phénomènes qu'ils donnent à voir constituent le pendant, à l'écrit, des phénomènes de disfluence, de répétition et de reformulation déjà observés et étudiés pour l'oral, et habituellement occultés dans le produit final du processus d'écriture. Ces similitudes suggèrent un ensemble de processus cognitifs communs à la production de la langue écrite et de la langue orale.

Bibliographie

ABEILLÉ A. et GODARD D. (en coll. avec DELAVEAU A. et GAUTIER A.) (éd.) 2021. *Grande Grammaire du Français.* Paris: Actes Sud / Imprimerie Nationale. 2628 p.

ALVES, R. A., CASTRO, S. L. et OLIVE, T. 2008. Execution and pauses in writing narratives: Processing time, cognitive effort and typing skill. *International journal of psychology*, 43(6) : 969-979.

ALVES, R. A. et LIMPO, T. 2015. Progress in written language bursts, pauses, transcription, and written composition across schooling. *Scientific Studies of Reading*, 19(5) : 374-391.

BLANCHE-BENVENISTE, C. 2003. La naissance des syntagmes dans les hésitations et répétitions du parler. In J.-L. AROUI (éd.), *Le sens et la mesure. De la pragmatique à la métrique, Hommages à Benoît de Cornulier*. Paris : Honoré Champion : 153-169.

BLANCHE-BENVENISTE, C. 2010. Lexique et grammaire dans les reformulations. In M. CANDEA et R. MIR-SAMII (éds.), *La rectification à l'oral et à l'écrit. Hommage à Marie-Annick Morel*. Paris : Ophrys : 77-89.

BOLINGER, D. (1976). Meaning and memory. *Forum Linguisticum* 1(1) : 1-14.

BOURIGA, S. (2020). Papier-crayon vs. écran-clavier : Effets sur le coût cognitif et sur la dynamique de la production de textes. Sous la direction de Thierry Olive. Université de Poitiers.

CISLARU, G. et OLIVE, T. 2018. *Le processus de textualisation: Analyse des unités linguistiques de performance écrites.* Bruxelles : De Boeck Supérieur.

CHOI-JONIN, I. & LAGAE, V. 2015. Les pronoms personnels clitiques, in *Encyclopédie grammaticale du français*, en ligne: encyclogram.fr

CONIJN, R., ROESER, J., et VAN ZAANEN, M. 2019. Understanding the keystroke log: the effect of writing task on keystroke features. *Reading and Writing*, 32(9) : 2353-2374.

ERMAN, B. et WARREN, B. 2000. The idiom principle and the open choice principle. *Text & Talk*, 20(1) : 29-62.

FAGARD, B. et CHAROLLES, C. 2018. Ailleurs, d'ailleurs, par ailleurs : De l'espace à l'humain, de l'humain au discours. *Journal of French Language Studies*, 28(3) : 351-375.

FELTGEN, Q, CISLARU, G. et BENZITOUN, C. 2022. Étude linguistique et statistique des unités de performance écrite : le cas de *et*. *SHS Web of Conferences*, 138, 10001.

FOULET, L. 1990. *Petite syntaxe de l'ancien français*. Paris: Champion.

FOULIN, J. N. 1995. Pauses et débits: les indicateurs temporels de la production écrite. *L'année psychologique*, 95(3) : 483-504.

GILQUIN, G. 2020. In search of constructions in writing process data. *Belgian Journal of Linguistics*, 34(1) : 99-109.

JUST, M. A. et CARPENTER, P. A. 1992. A capacity theory of comprehension: individual differences in working memory. *Psychological review*, 99(1) : 122-149.

KAUFER, D. S., HAYES, J. R. et FLOWER L. 1986. Composing written sentences. *Research in the Teaching of English* 20(2) : 121-140.

KIRCHER, T. T., BRAMMER, M. J., LEVELT, W., BARTELS, M. et MCGUIRE, P. K. 2004. Pausing for thought: engagement of left temporal cortex during pauses in speech. *NeuroImage*, 21(1) : 84-90.

KURODA, S.Y. Le jugement catégorique et le jugement thétique, Exemples tirés de la syntaxe japonaise. Langages, 30 : 81-110.

LAFON, P. 1981. Statistiques des localisations des formes d'un texte. *Mots* 2 : 157-188.

LE GOFFIC, P. 1993. *Grammaire de la Phrase Française*. Paris: Hachette Supérieur.

LÉARD, J.-M. 1992. *Les gallicismes : Etude syntaxique et sémantique*. Paris – Louvain-la-Neuve : Duculot.

LEFEUVRE, F. 2006. *Quoi de neuf sur quoi ? Etude morphosyntaxique du mot quoi*. Rennes: Presses Universitaires de Rennes.

LEIJTEN, M. et VAN WAES, L. 2013. Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communications* 30(3) : 358-392.

MOIGNET, G. 1981. *Systématique de la langue française*. Paris: Klincksieck.

OLIVE, T. et BOURIGA, S. 2022. *Effects of cognitive demands of planning on bursts when writing with a pen or with a computer*. Communication à la conférence SIG-Writing 2022, Umeå, Suède.

RIEGEL, M., PELLAT, J.-C. et RIOUL, R. 2009. *Grammaire méthodique du français.* Paris : Presses Universitaires de France.

TORRANCE, M. and GALBRAITH, D. 2006. The Processing Demands of Writing. In C. MACARTHUR, S. GRAHAM et J. FITZGERALD (éds.), *Handbook of Writing Research*. New York : Guilford Publications : 675-698.

WENGELIN, Å. 2006. Examining Pauses in Writing: Theory, Methods and Empirical Data. In K. P. H. SULLIVAN and E. LINDGREN (éds.), *Computer key-stroke logging and writing: methods and applications*. Oxford : Elsevier : 107-130.

WOOD, D. 2002. Formulaic language acquisition and production: Implications for teaching. *TESL Canada Journal*, 20(1) : 1-15.

WRAY, A. et PERKINS, M. R. 2000. The functions of formulaic language: An integrated model. *Language & Communication*, 20(1) : 1-28.