Etude linguistique et statistique des unités de performance écrite

Quentin Feltgen, Georgeta Cislaru, Christophe Benzitoun

Étude linguistique et statistique des unités de performance écrite : le cas de *et*

Quentin Feltgen[1,*], *Georgeta* Cislaru[2], et *Christophe* Benzitoun[3]

[1]Université Sorbonne Nouvelle, CLESTHIA, 75005 Paris, France
[2]Université Paris Nanterre, MoDyCo, 92001 Nanterre, France
[3]Université de Lorraine, ATILF, 54063 Nancy Cedex, France

Résumé. Cet article aborde la question de la segmentation par des pauses du flux de production écrite enregistré en temps réel et de la motivation linguistique et statistique de l'emplacement des pauses. En effet, les pauses segmentant des séquences textuelles linguistiquement analysables, il est crucial de comprendre si des contraintes régulières en fixent les frontières. Nous avons choisi de nous pencher sur le cas de la conjonction *et*, en vertu de la diversité sémantique et morphosyntaxique des relations qu'elle sémiotise. Après avoir mis en perspective les résultats d'une analyse de corpus antérieure, nous procédons à une annotation manuelle des occurrences en départageant les emplois extra- et intraphrastiques de *et* dans un corpus de textes courts produits par des adultes (étudiants). Une méthode d'analyse statistique est ensuite appliquée aux données annotées pour tester les attentes statistiques en termes d'emplacement des pauses. Cette analyse permet de faire ressortir des différences de segmentation en fonction du type d'emploi de *et*.

Abstract. Linguistic and statistical study of written performance units: the case of *et* in French. This paper addresses the issue of segmentation by pauses of the stream of written production recorded in real time and the linguistic and statistical relevance of the location of these pauses. Indeed, as pauses segment linguistically analyzable textual sequences, it is crucial to understand whether regular constraints set their boundaries. We have chosen to focus on the case of the conjunction *and* (*et*) due to the semantic and morphosyntactic diversity of the relations it semiotizes. Once put into perspective the results of a previous corpus analysis, we proceed to a manual annotation of the occurrences by separating the extra- and intraphrastic uses of *and* in a corpus of short texts produced by adults (students). A statistical analysis method is then applied to the annotated data to test statistical expectations in terms of break location. This analysis reveals differences in segmentation based on the type of use of *et*.

* Corresponding author : quentin.feltgen@gmail.com

1 Introduction

L'étude du processus d'écriture enregistré en temps réel grâce à des logiciels de suivi des frappes soulève de nombreuses questions méthodologiques. En effet, ce champ implique un changement de perspective sur la production langagière, en introduisant une dimension temporelle et dynamique (linéarité du processus) ainsi qu'une dimension comportementale (alternance entre pauses et production, gestes de réécriture). Dans cette optique, se pose la question de l'identification et de la définition des objets d'analyse linguistique. Dans la littérature s'intéressant au processus d'écriture, on distingue habituellement les événements processuels ayant une incidence sur la production langagière et les unités de performance écrite. Les événements processuels comprennent l'ensemble des opérations d'écriture et de réécriture, tels les ajouts, les suppressions, les déplacements, les substitutions ou les insertions – ces objets sont définis dans le cadre de la génétique textuelle (Grésillon 2016 [1994], Lebrave 1987, Doquet & Leblay 2014). Les unités de performance écrite correspondent aux séquences linguistiques produites entre deux pauses et/ou entre deux événements de réécriture – ces objets sont initialement identifiés en creux par les études psycholinguistiques s'intéressant au rôle cognitif des pauses à l'écrit (Foulin 1995, Matsuhashi 1981, Schilperoord 1996, Spellman-Miller 2006, Alves *et al*. 2007), pour ensuite être étudiés dans le cadre d'approches linguistiques notamment sous le nom de *jets textuels* (Cislaru & Olive 2018). Ces derniers sont définis comme des séquences langagières segmentées par deux pauses d'une durée suffisamment longue pour pouvoir leur conférer un rôle cognitif (planification, sélection, sémiotisation) en vertu d'hypothèses comportementales formulées par la psycholinguistique de l'écrit. Mais leur traitement linguistique soulève plusieurs difficultés, dans la mesure où moins de la moitié des jets textuels que nous avons observés correspondent formellement aux unités habituellement segmentées en linguistique de l'écrit, tels les lexèmes, les groupes, les phrases (Cislaru & Olive 2018 : 90-94). Quelle motivation linguistique peut-on associer à la segmentation pausale ?

Dans cet article, nous nous pencherons sur les principes de segmentation des jets textuels et discuterons de la possibilité d'établir des relations linguistiques ou des corrélations statistiques entre l'occurrence d'une pause et la structure interne des jets textuels. Nous présenterons tout d'abord les conditions de recueil des données processuelles et nos choix terminologiques et méthodologiques. Nous proposerons ensuite une étude de cas en prenant pour objet les jets textuels contenant la conjonction *et* ; cette étude comportera un volet descriptif, basé principalement sur l'annotation morphosyntaxique des jets textuels, et un volet statistique, calculant la probabilité d'occurrence de la conjonction en début, fin ou milieu de jet textuel.

2 Les jets textuels comme observables linguistiques

L'étude du processus d'écriture étant un domaine relativement nouveau en linguistique, nous baliserons ici le périmètre de nos observables et les outils mobilisés.

2.1 Nature des données processuelles

Le processus d'écriture peut être enregistré en temps réel grâce à des logiciels de suivi des frappes. Ces logiciels captent la chronologie du processus à travers l'ensemble des événements au clavier et les mouvements de souris d'une part, et l'ensemble des formes écrites d'autre part. Les formes écrites correspondent généralement à des séquences linguistiques constituant le produit textuel, même si tout enregistrement peut comporter des résidus graphiques inclassables linguistiquement. Dans ce cadre, les pauses au cours de l'activité d'écriture occupent une place importante.

En effet, la psycholinguistique de l'écrit fournit des hypothèses robustes justifiant le caractère cognitivement motivé des pauses (cf. Olive 2012) comme, par exemple, la détermination de leur longueur par le degré de complexité du segment qui les suit – plus la complexité et le niveau hiérarchique sont élevés, plus la durée attendue est importante (van Hell et al. 2008, Immonen & Mäkisalo 2010, Medimorec & Risko 2017). La segmentation pausale livre ainsi des séquences linguistiques – les jets textuels – dont la pertinence est justifiée d'un point de vue cognitif. La question se pose de savoir si la motivation cognitive est marquée linguistiquement par des régularités au niveau des lieux de segmentation d'une part et de la cohérence interne des jets textuels d'autre part. Dit autrement, nous posons que la distribution des pauses au sein de la chaîne écrite n'est pas le résultat du hasard mais qu'elle est partie-prenante des dynamiques de *chunking*, défini comme la capacité d'un individu à regrouper des informations en utilisant ses connaissances afin de traiter l'information cible (voir Gobet *et al*. 2001) et d'assurer la fluidité de la production en permettant des sauts qualitatifs du simple au complexe (Christiansen et Chater 2016).

Notre analyse s'effectue sur un corpus de tels jets textuels (i.e. séquences délimitées par des pauses dépassant un seuil), enregistré en contexte expérimental. 83 participants en Licence 2 de Psychologie ont chacun rédigé un texte court sur un thème imposé et variable, mais supposé familier des scripteurs (consommation de cannabis, usage du préservatif, des transports en commun, etc.), en un temps imparti (15 minutes par session). Lors de la rédaction au clavier, le processus de production était recueilli via le logiciel Inputlog (Leijten & van Waes 2013), qui enregistre, pour chaque événement de frappe, le caractère saisi, l'instant où la touche a été pressée, et l'instant où la touche a été relâchée. Chaque texte correspond en moyenne à 3 600 événements, avec un minimum de 1 600 pour le plus court et de 7 000 pour le plus long ; en termes de mots, entre 157 et 919 mots, avec une moyenne de 460.

Le délai entre deux événements se définit comme l'intervalle de temps entre la pression des deux touches concernées. Lorsque ce délai dépasse le seuil de 2 secondes, on considère qu'il constitue une pause dans le processus de production. Celui-ci se trouve ainsi segmenté en une succession de jets textuels, lesquels présentent une linéarité temporelle, mais non spatiale (le scripteur peut réviser ce qu'il vient d'écrire, revenir en amont du texte pour ajouter ou amender une partie, etc.). Si ce seuil de 2 secondes a souvent été retenu par la littérature (Wengelin 2006), la détermination la mieux appropriée du seuil reste sujette à débat (Dragsted 2005) : plus le seuil est faible, moins il est sélectif (on enregistre alors comme pauses des heurts dans la traduction mécanique au clavier de la production langagière) ; plus il est élevé, moins il y a de jets, ce qui laisse potentiellement de côté de nombreux phénomènes pertinents. Cependant, cette difficulté n'est pas critique, des résultats statistiquement comparables pouvant être obtenus pour une fourchette allant de 1 à 3 secondes typiquement. C'est la raison pour laquelle nous avons retenu ici ce seuil de 2 secondes.

2.2 Du corpus aux observables

D'un point de vue technique, les jets textuels sont des chaînes graphiques séparées par des pauses d'une longueur suffisamment importante pour être considérées comme cognitivement pertinentes (ici, 2 secondes). En voici un exemple :

Exemple 1
[pause] . De plus, l'augmentation du prix des paquets [pause] profitent aux r [pause] égions [pause] régions proches de la frontières [pause], car les [pause] t [pause] [1]

Une fois le seuil de pause pertinent fixé, il est convenu que les séquences graphiques ainsi segmentées puissent constituer des observables linguistiques plausibles : en effet,

d'une part, on reconnaît dans ces séquences des formes linguistiques analysables (morphèmes, lexèmes, groupes syntaxiques, ponctuation) et, d'autre part, l'application d'un seuil homogène garantit les mêmes conditions de segmentation et permet de définir les séquences ainsi identifiées comme des unités de performance écrite.

Dans les faits, la segmentation pausale, qui trouve bien une justification comportementale cognitive, donne lieu à des unités de performance non homogènes, pouvant aller de la lettre (notamment dans le cas des révisions-corrections) à la phrase voire au-delà. Même si, dans l'ensemble, les jets textuels correspondent plutôt à des enchaînements de quelques mots comprenant parfois des signes de ponctuation, il est impossible de structurer ces données en utilisant strictement les catégories lexico-grammaticales habituelles. Ce constat suffirait-il à disqualifier les unités de performance écrite en tant qu'objets d'analyse linguistique ? Il nous est permis de répondre par la négative, dans la mesure où une approche émergentiste comme la nôtre est d'emblée orientée vers le repérage des occurrences et non des types. Bien au contraire, en sachant que, d'une part, un texte ne peut pas être produit d'une seule traite et que le processus de production est constitué d'une alternance de pauses et de productions linguistiques et, d'autre part, un texte ne peut pas être constitué de séquences linguistiques identiques ou trop répétitives (en vertu des principes de cohérence et de progression), les unités de performance écrite ne peuvent qu'être diversifiées.

Chaque jet textuel étant unique, sinon dans sa nature, du moins vis-à-vis de son environnement de production, qui est à chaque fois différent, la segmentation pausale est susceptible d'être conditionnée par un complexe de facteurs linguistiques, cognitifs et situationnels. De ce fait, les principes de segmentation des jets textuels et la motivation linguistique de cette segmentation n'ont pas encore été élucidés, même si plusieurs hypothèses ont pu être formulées (voir Cislaru & Olive 2018). Une de ces hypothèses prend appui sur la nature des unités linguistiques par lesquelles débute ou finit un jet textuel, sur le modèle des groupes morphosyntaxiques : l'attention est alors portée sur les déterminants, les signes de ponctuation, le rapport entre sujet et verbe, etc. Mais ce focus soulève d'autres questions concernant les mécanismes sous-tendant l'accrétion des jets textuels en tant que regroupements sémantiquement et cognitivement pertinents : au-delà des relations de dépendance syntaxique et surtout lorsque les frontières des jets ne sont pas calquées sur les frontières des groupes, qu'est-ce qui assure les liens internes des jets textuels ?

3 Étude de cas : les jets textuels comportant un *et*

Nous avons porté notre attention sur l'unité linguistique *et*, nous proposant d'observer son emplacement au sein des jets textuels et le rôle qu'elle est susceptible de jouer en tant qu'outil relationnel à la fois sémantique et syntaxique.

3.1 *Et* : caractéristiques linguistiques et spécificités d'usage

La conjonction *et* présente plusieurs caractéristiques qui justifient l'étude de son comportement au niveau du processus.

Du point de vue de la production, cette unité se présente comme un outil d'enchaînement efficace et facilement disponible ; conjonction la plus fréquente du français, *et* jouit du statut d' « archi-connecteur » (Bronckart & Schneuwly 1984 ; Schneuwly & Bronckart 1986). Ainsi, il est souvent convoqué à l'oral pour assurer la continuité discursive, servir de joncteur entre deux constituants (Bilger 1999) ou commencer une « phrase ». Dans les productions orales, *et* assure la continuité discursive ; son emploi relève d'une « routine énonciative » nécessaire au maintien de la continuité du monologue (Mouchon, Fayol & Gombert 1991 cités par Favart & Passerault 1999 : 159). Si l'on ne saurait confondre production orale et écriture, il est néanmoins possible de

rapprocher, à des fins heuristiques, le processus d'écriture réalisé en temps réel de la linéarité progressive des discours oraux.

Et peut exprimer à lui seul une palette de relations différentes : addition, succession, chronologie, opposition, conséquence, etc. (cf. Riegel *et al*. 2005 [1994] : 880 ; voir aussi Charaudeau 1992 : 503-504 ; Rousseau 2007). Grâce à ses propriétés syntaxiques et sémantiques, cette conjonction est privilégiée dans le processus d'acquisition du langage. Elle est ainsi l'unité assurant la cohésion du texte la plus largement utilisée dès les débuts de l'acquisition de l'écrit (Favart & Passerault 1999, Favart & Chanquoy 2007 : 54), avoisinant les 50% (Paolacci & Favart 2010). Même si son utilisation diminue ensuite au profit de connecteurs plus diversifiés sémantiquement (Fayol 1986, Favart & Passerault 1995), *et* reste très fréquent dans le discours oral comme dans le discours écrit.

Le fonctionnement de *et* comme connecteur en début de phrases est donc particulièrement visible chez les enfants pendant les premières années d'acquisition de l'écrit, où *et* assume non seulement le même fonctionnement que celui qu'il a chez les jeunes enfants à l'oral, mais aussi tient lieu de tout autre connecteur, sur le mode de l'accumulation dans une démarche d' « énonciation première » (cf. Doquet 2011 : 190) : « il y a eu une baston entre des garcons ils se sont mis des patates **he** [=et] j'etait pas la ».

Compte tenu de ces différentes particularités, *et* devrait bénéficier d'une facilité de traitement se répercutant sur la dynamique du processus et sur la segmentation des jets textuels, y compris dans des textes produits par des adultes. La conjonction constitue dès lors un candidat idéal pour aborder la gestion processuelle des jets textuels. Ainsi, la coordination aux niveaux intra- et extraphrastique, recouvrant des valeurs différentes (cf. 3.2 sur l'annotation), pourrait donner lieu à des comportements différents. Au-delà de cette distinction, certaines relations sémiotisées par *et*, telles les relations temporelles, consécutives ou disjonctives, ne sont pas symétriques (ces particularités sont signalées aussi bien pour le français que pour l'anglais : Hobaeck Haff 1987, Sag 2005, Renner 2007). Au niveau de la production, cette asymétrie pourrait donner lieu à un haut degré d'implicitation entre l'unité à gauche et l'unité à droite du connecteur entraînant des répercussions sur l'incidence des segmentations pausales, avec potentiellement un phénomène d'attraction entre les deux éléments.

Si l'on considère, à la suite de Lang (1984), que la coordination est caractérisée par un trait sémantique commun aux deux segments coordonnés, on saisit d'emblée le rôle que le *et* est susceptible de jouer au niveau de l'agrégation des *chunks*. Se basant sur des données issues de l'anglais, Lang définit la valeur relationnelle de la coordination comme une entité conceptuelle dont le sens est différent de celui de la conjonction de coordination elle-même ; il appelle ce trait sémantique commun *Common Integrator*.

A l'inverse, dans leur étude prospective de grammaire linéaire (*Linear Unit Grammar*, portant sur l'anglais également) qui cherche à rendre compte du traitement en production en s'appuyant sur le regroupement en unités de communication en réception, Sinclair et Mauranen (2006) stipulent que *et* doit être traité séparément, en tant qu'unité autonome ou, dans tous les cas, distincte et non intégrée aux séquences que la conjonction relie.

Des études antérieures de l'actualisation du coordonnant *et* dans l'écriture enregistrée en temps réel montrent une réalité contrastée (Cislaru & Olive 2016). Ainsi, chez des scripteurs expérimentés comme les travailleurs sociaux ou les étudiants de Master, le connecteur peut être actualisé :

- seul, en tant que jet textuel unique (moins de 3%)
- en début d'un jet textuel, à l'instar des usages à l'oral ou dans l'écriture débutante (30%)
- en fin de jet textuel, comme une ouverture sémiotisant une attente de complétion (moins de 10%)
- à l'intérieur d'un jet textuel, sous le format *X et Y* (environ 60% des cas).

Les résultats ont été contrastés avec un corpus d'écrits d'élèves faisant ressortir une inversion des conditions d'actualisation, avec une préférence marquée pour le format *et X* dans les écrits d'élèves (2/3 des jets textuels).

Ces données tendraient à confirmer, d'une part, la nature du traitement de l'information et le rôle joué par le coordonnant dans l'écriture débutante et, d'autre part, le potentiel intégrateur du coordonnant dans le *chunking* (cf. Lang *op. cit*). Nous nous proposons de vérifier ces hypothèses statistiquement, à partir d'un corpus d'écrits d'étudiants, en procédant dans un premier temps à une annotation des emplois extra- et intraphrastiques. Il s'agit d'examiner si le niveau des unités impliquées a une incidence sur la composition des jets textuels.

3.2 Annotation des jets textuels

Notre corpus de travail comportait 604 occurrences de *et*, pour lequel nous n'avons pas considéré de formes graphiques alternatives. Pour que l'ensemble des exemples puissent être annotés manuellement, une concordance de type KWIC a été générée. À la différence de concordances « classiques », les opérations d'édition sont représentées visuellement. Ainsi, il ne s'agit pas d'un texte linéaire mais d'une photographie d'un texte en cours de rédaction comportant les retouches. Nous allons l'illustrer à l'aide de l'exemple suivant :

Exemple 2
facile ␣ d'access ␣ ⌫⌫ ␣ **et** ␣ facile ␣ a⌫à ␣ utiliser.

Les espaces sont représentées visuellement, de même que les suppressions (⌫) et les substitutions. Dans cet exemple 2, on voit que la chaine « s␣ » a été supprimée et le « a » sans accent a été remplacé par un « à » avec accent. Il s'agit d'un cas assez simple et relativement facile à interpréter. Cependant, de nombreux exemples sont loin d'être aussi faciles à décoder :

Exemple 3
Pour ␣ conclr⌫ure, ␣ on ␣ ␣ pourra ␣ dire ␣ u⌫que ␣ la ␣ législation ␣ du ␣ cannabis ␣ est ␣ un ␣ sujet ␣ portan⌫⌫⌫⌫⌫⌫⌫ammenant ␣ ⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫ju⌫**et** ␣ d'actualité

Ici, il faut comprendre que le *et* en question fait partie de la suite « un suj**et** d'actualité », ce qui en fait un exemple à rejeter dans le cadre de notre étude. L'annotation manuelle est donc particulièrement complexe car il ne suffit pas de simplement lire les exemples : de fait, il faut les reconstruire partiellement.

Pour éviter un trop grand nombre d'erreurs liées à la matérialité du texte, nous avons préféré ne pas nous lancer dans une annotation trop fine et nous nous sommes restreints à distinguer seulement deux catégories, à savoir les joncteurs intraprastiques et les connecteurs extraphrastiques. Les joncteurs intraphrastiques sont des unités qui permettent de coordonner deux constituants à l'intérieur d'une même position syntaxique. C'est ce qui est appelé « joncteur de liste » dans Bilger (1999) et plus largement dans le cadre de l'approche pronominale. Ainsi, dans l'exemple ci-dessous, « et » permet de coordonner le constituant « diminuer le nombre de fumeur » avec « avoir plus d'impact » dans le même paradigme de constituants dépendants du verbe « pouvoir » :

Exemple 4
ces messages peuvent avoir plus d'impact **et** ainsi diminuer le nombre de fumeur

Nous avons aussi inclus dans cette catégorie les cas de coordination de propositions lorsque le sujet est mis en facteur commun. C'est le cas dans l'exemple ci-dessous :

Exemple 5
Ce cout est assez élévé **et** est obligatoire pour les étudiants n'ayant pas les bourses

Ici, il y a deux verbes principaux, mais le second ne pourrait pas fonctionner en l'absence du sujet du premier verbe.

Pour ce qui est des connecteurs extraphrastiques, il s'agit de tous les cas où « et » permet d'articuler deux phrases ou, plus généralement, deux unités textuelles.

Exemple 6
Il existe le préservatif masculin, **et** depuis quelques années, il existe le préservatif féminin également.

Comme on l'aura compris avec l'exemple 6, il ne s'agit pas d'une annotation basée sur les marques graphiques, la présence d'une ponctuation forte ou d'une virgule ne modifiant nullement notre analyse. Si deux propositions sont reliées par *et* sans unité rectrice permettant de considérer qu'elles appartiennent à un même paradigme syntaxique, nous avons annoté ce *et* comme un connecteur extraphrastique.

Au cours de la lecture de la concordance, nous avons éliminé un certain nombre d'exemples que nous ne parvenions pas à interpréter. Nous avons également écarté les exemples en *et ce* et *et cela*, qui fonctionnent différemment des précédents. Nous aurions également pu créer une catégorie supplémentaire pour les cas comme « Un mot de plus **et** je m'en vais » mais ce type d'emploi n'est pas attesté dans notre corpus.

Après lecture de l'ensemble des lignes de la concordance, nous avons retenu 475 occurrences (381 occurrences intraphrastiques et 94 occurrences extraphrastiques).

4 Analyse statistique des jets textuels : application à la répartition des occurrences de *et*

Comme on vient de le constater, les jets textuels posent un problème épistémique majeur. En effet, ils se caractérisent à la fois par leur complexité, leur diversité, et leur hétérogénéité. Pris dans leur individualité, la logique de chaque jet s'enracine tout à la fois dans le contexte textuel dans lequel il s'insère, dans son contenu linguistique propre, et dans le contenu linguistique à venir, qui mobilise des processus cognitifs pouvant favoriser l'interruption du jet en cours. Saisir les conditions de réalisation et les motivations d'un jet individuel est donc une véritable gageure, et de multiples explications sont parfois possibles, sans que l'on puisse toujours déterminer laquelle prévaut.

Alternativement, il est possible de considérer les jets via une approche statistique, c'est-à-dire d'essayer de dégager des traits récurrents dans la constitution et la délimitation des jets. En ce cas, l'objectif n'est plus de saisir la logique du jet dans son individualité, mais des principes valables dans leur globalité, et supposés généralement à l'œuvre. Parmi ces observations statistiques, on peut notamment considérer la position de certains types d'unités linguistiques vis-à-vis des jets : en début de jet, en fin de jet, en milieu de jet, entre deux jets, ou formant un jet à elles seules, comme il a été suggéré plus haut pour la conjonction *et*. Un test multinomial permet alors en principe d'évaluer la significativité de ces observations[2].

Cependant, un tel test présente deux écueils : d'une part, les observations seront mécaniquement moins susceptibles d'être significatives si elles sont associées à un faible nombre d'occurrences ; d'autre part, il porte sur une distribution aléatoire des occurrences dans un texte dont la segmentation est fixe. Or, les occurrences sont très sensibles à

l'organisation linguistique (par exemple, on trouvera difficilement un verbe en début de phrase, alors que les circonstants et marqueurs discursifs y sont fréquents), si bien qu'une distribution aléatoire des occurrences n'a aucune chance d'être réaliste. De ce fait, il est préférable de conserver le texte tel quel, et au contraire de considérer une distribution aléatoire des pauses de production.

Là encore, tous les emplacements du texte ne sont pas équivalents du point de vue des pauses. En effet, certains facteurs élémentaires relatifs à la structuration du texte, comme les fins de phrase, les fins de paragraphe, etc., rendent les pauses plus probables, même si ces facteurs structurels sont loin de pouvoir rendre compte de la segmentation en jets de la production textuelle. Pour considérer une distribution aléatoire réaliste, il convient de tenir compte de ces facteurs qui constituent une sorte d'explication « au premier ordre » de la segmentation. Autrement, les observations linguistiques, en ce qu'elles refléteront déjà l'effet de ces facteurs, seront toutes significatives, et l'hypothèse nulle ne pourra pas servir de crible suffisamment sensible pour signaler la pertinence ou non de ces observations. Si au contraire l'hypothèse nulle tient compte de l'effet des facteurs, l'interaction observée des jets et de leur contenu linguistique, lorsqu'elle apparaît significative, nous renseignera bien quant à la motivation linguistique de ces jets de production, permettant d'avancer un niveau plus loin dans l'explication. Il est donc nécessaire de considérer des segmentations pausales aléatoires dont les propriétés relativement aux facteurs structurels les plus élémentaires reprennent celles de la segmentation pausale véritable.

A cette fin, nous avons mis en place une procédure s'inscrivant dans le cadre méthodologique dit Monte Carlo, et qui consiste à générer sous l'hypothèse nulle un large nombre de segmentations alternatives et aléatoires des textes pour construire une distribution à laquelle comparer ensuite l'observation empirique. Nous nous sommes en outre fixé le but d'élaborer une hypothèse nulle aussi riche et informée que possible quant à la structure de chacun des textes, de sorte que la significativité soit spécifiquement sensible aux idiosyncrasies de la forme linguistique étudiée (ici la conjonction *et*), et non pas aux logiques sous-jacentes plus générales de la segmentation en jets.

Nous présenterons ici cette méthode, et en particulier l'hypothèse nulle sur laquelle elle repose, en elle-même révélatrice des principaux caractères statistiques des jets textuels. Nous l'appliquerons ensuite au cas de la conjonction *et* qui nous préoccupe ici, en contrastant ses emplois intraphrastiques et extraphrastiques. Les différences de comportement observées corroborent, incidemment, le caractère polyfonctionnel de cette conjonction.

4.1 Méthode d'investigation statistique des jets textuels

La méthode d'inspiration Monte Carlo que nous utilisons pour quantifier la pertinence des observations effectuées sur les jets textuels consiste en trois étapes : l'élaboration d'une hypothèse nulle, la génération de segmentations aléatoires du texte suivant cette hypothèse nulle, et finalement la comparaison entre la distribution ainsi obtenue et les données empiriques, propres à la segmentation par le scripteur.

Dans ce qui suit, notre approche consiste à modéliser chaque texte indépendamment et isolément. Ce choix s'explique par la très forte variabilité interindividuelle entre les scripteurs. Par exemple, certains écrivent le texte d'une traite, d'autres multiplient les va-et-vient dans le texte, donnant lieu à des dynamiques d'écriture très disparates (Leblay & Caporossi 2015). Nous estimons donc plus robuste de caractériser la segmentation de chaque texte séparément, plutôt que de mettre en commun des données obéissant à des logiques de production hétérogènes.

4.1.1 Élaboration de l'hypothèse nulle

L'hypothèse nulle que nous considérons repose sur l'idée que les jets seraient des objets aléatoires, sans contenu linguistique déterminé ; dès lors, on peut définir la probabilité d'une observation vis-à-vis d'une telle segmentation aléatoire. Si l'observation dévie de la segmentation aléatoire, c'est qu'elle manifeste une sensibilité des jets au contenu linguistique relatif à cette observation. Cependant, avec une hypothèse nulle trop sommaire, par exemple en supposant que la segmentation réelle peut s'expliquer par une probabilité constante et homogène à travers le texte de s'arrêter après chaque événement de frappe, il est à peu près inévitable que toute observation, en ce qu'elle manifestera certaines sensibilités de cette segmentation au plus bas niveau du contenu linguistique du texte (à commencer par le fait qu'il est plus probable de stopper la production entre deux mots plutôt qu'en plein milieu d'un mot), se révèlera significative.

Dès lors, pour saisir en quoi les jets sont sensibles à leur contenu linguistique, il convient d'enrichir l'hypothèse nulle pour évacuer leur sensibilité aux facteurs explicatifs les plus élémentaires. L'hypothèse nulle reviendra donc toujours à considérer que la segmentation est aléatoire, mais cette dimension aléatoire sera désormais non plus homogène à travers le texte, mais sensible au contexte local : par exemple à la fin d'un mot, à la fin d'une phrase[3]. Cela suppose donc d'évaluer l'impact d'un nombre donné de facteurs sur la probabilité d'interrompre le flux de production par une pause.

Les facteurs que nous avons retenus comme étant les plus déterminants pour rendre compte de la segmentation textuelle sont les suivants : 1 - pause après chaque événement de frappe ; 2 - pause après chaque mot ; 3 - pause après chaque marqueur de ponctuation faible (,) ; 4 - pause après chaque marqueur de ponctuation forte (., !, ?, …) ; 5 - pause avant un événement de révision (le scripteur efface une partie du texte précédemment produite). D'autres facteurs seraient certainement pertinents, en particulier le passage d'un groupe syntaxique à un autre ; cependant, en l'absence d'annotation entièrement fiable du texte à ce stade (le caractère incomplet des unités produites et les révisions multiples ne permettent pas une annotation automatique robuste), nous nous sommes limités à des facteurs directement définis à partir des formes linguistiques.

Hormis le facteur 1, toujours actif, nous avons défini ces facteurs comme indépendants, avec priorité donnée à la révision : ainsi, si une révision intervient immédiatement après une ponctuation, l'événement est caractérisé comme révision, et non comme étant associé au marqueur de ponctuation. De même, le facteur 2 correspond plus précisément à la production d'une espace entre deux caractères alphabétiques[4]. Les marqueurs de ponctuation médiane (:, ;) ne sont pas pris en compte, car une trop grande proportion de textes de notre corpus ne présente aucun de ces marqueurs. Le dernier facteur considéré (pause avant révision) est peut-être plus sujet à caution, puisque l'on peut s'interroger sur les raisons pour lesquelles la révision, tout comme la pause, intervient là et pas ailleurs ; c'est-à-dire qu'il faudrait éventuellement mettre révisions et pauses sur un même plan, l'une et l'autre devant résulter de facteurs considérés comme explicatifs. Cependant, comme la série des révisions est donnée, et que les événements de révision sont souvent précédés de pauses, il fait sens d'intégrer ce phénomène à l'hypothèse nulle[5]. Mettre de côté l'effet des révisions reviendrait à introduire une disparité indésirable entre les segmentations aléatoires et la segmentation d'origine.

Une fois ces facteurs identifiés, il est possible d'en évaluer l'impact sur la segmentation en jets, pour chaque texte considéré séparément. Pour ce faire, on effectue une régression multinomiale de la variable dépendante encodant les pauses. On associe à chaque événement de frappe une valeur binaire, codant 1 si l'événement est suivi par une pause et donc termine un jet ; ceci constitue un vecteur Y dont la longueur est égale au nombre des événements, et qui fournit la variable dépendante du modèle. Les prédicteurs, correspondant à chacun des cinq facteurs énoncés ci-dessus, sont également des vecteurs de variables binaires, codant pour le contexte. Le prédicteur 1 est toujours égal à 1 ; les autres prédicteurs sont égaux à 1 si l'événement correspond au facteur considéré (fin d'un mot,

ponctuation faible ou forte, dernier événement avant une révision, etc.), et construits comme orthogonaux deux à deux ainsi qu'il a été précisé plus haut. Ces cinq vecteurs constituent la matrice des prédicteurs *X*. Le modèle utilisé est alors le suivant :

$$Y = 1 / [1 + \exp(-\beta X)] \quad (1)$$

où β est un vecteur de taille 5 qui pondère chacun des facteurs. Le modèle optimise ces poids pour chaque texte afin de prédire au mieux *Y* à l'aide de la matrice des prédicteurs *X*. Ce sont ces poids qui permettent de caractériser et d'informer l'hypothèse nulle, dont le détail est spécifique à chaque texte.

En soi, ce modèle de régression logistique renseigne déjà quant à certaines propriétés statistiques des jets. Plus précisément, il informe de la probabilité qu'une pause survienne dans les contextes correspondant aux facteurs listés. Ces probabilités, moyennées au niveau du groupe, sont répertoriées dans le Tableau 1 ; la valeur *p* associée au facteur y est également reportée. La valeur *p* n'est pas calculée au niveau des poids, mais des probabilités, dans la mesure où elles s'écartent de la probabilité « standard » du facteur 1. La valeur *p* correspond alors au test *t* associé à la différence entre la probabilité associée à chaque facteur et la probabilité standard au niveau du groupe. Il apparaît que tous les facteurs sont fortement significatifs ; s'agissant d'une hypothèse nulle, on s'est en effet assuré de considérer des facteurs évidents, dont l'effet potentiel sur la segmentation en jets n'est pas en doute. Il ressort donc qu'une ponctuation forte induit une pause dans 57% des cas ; une fin de mot non accompagnée de ponctuation, dans 7% des cas.

Tableau 1. Probabilité d'une pause en fonction du contexte

Contexte	**Probabilité moyenne**	**Valeur *p***
après chaque événement	0.01	X
après un mot	0.07	4e-41
ponctuation faible	0.16	6e-22
ponctuation forte	0.57	1e-43
avant révision	0.22	5e-44

4.1.2 Génération de segmentations aléatoires alternatives

L'hypothèse nulle revient donc à se munir d'un modèle probabiliste lequel, à chaque événement, attribue la probabilité qu'il soit associé à une pause, délimitant ainsi le jet textuel de production en cours. Ce modèle est informé par le texte en ce qu'elle est entraînée de manière distincte pour chacun des textes du corpus, caractérisé alors de manière unique par la matrice des facteurs *X* et par la pondération β des différents facteurs.

Ce modèle probabiliste permet ainsi la génération aléatoire de segmentations alternatives de ces textes. Pour générer une segmentation aléatoire alternative *Y**, on tire *N* nombres aléatoires compris entre 0 et 1 correspondant à chacun des *N* événements composant le texte : si le nombre tiré est inférieur à la probabilité de pause

définie localement, alors *Y** est codé 1 pour cet événement. Cela définit une nouvelle segmentation du texte, aléatoirement générée suivant les probabilités déterminées par l'hypothèse nulle.

Nous avons généré 20 000 segmentations aléatoires, pour chaque texte. Cela permet de comparer la segmentation réelle à un nombre très vaste de segmentations obéissant à des principes généraux comparables, et s'appliquant au même texte. De là, on peut estimer la portée statistique des observations effectuées sur le texte.

4.1.3 Calcul de la valeur p

On peut en effet répéter l'observation d'intérêt sur chacune de ces segmentations fictives des textes du corpus (ce qui suppose néanmoins de pouvoir procéder à cette observation de manière automatisée), et en déduire une distribution statistique pour l'observation considérée. On obtient alors une distribution statistique sur la valeur associée à l'observable linguistique. Ici, nous considérons le ratio d'occurrences de *et* se trouvant respectivement en début, en milieu et en fin de jet ; ce ratio, qui dépend de la position relative des pauses et des occurrences concernées, dépend donc de la segmentation du texte, et peut être calculé pour chacune des segmentations aléatoires alternatives générées. Pour obtenir la valeur *p* associée à l'observation empirique réelle, il suffit alors de considérer où se situe la valeur associée dans la distribution aléatoire obtenue[6].

Pour un texte unique et considéré isolément, cette méthode présente l'inconvénient de livrer des résultats excessivement quantifiés. Ici, on considère le ratio d'occurrences, par exemple en début de jet. Or, ce ratio ne peut prendre que que $n + 1$ valeurs, où n est le nombre d'occurrences de l'unité considérée ; s'il n'y a que 3 occurrences au total, le ratio peut seulement valoir 0, 1/3, 2/3 ou 1 selon qu'on ne trouve aucune, une, deux ou trois occurrences en début de jet. La distribution statistique va alors se concentrer sur ces valeurs. Cela réduit la possibilité pour une observation empirique d'être significative ; en effet, l'analyse perd en granularité, ce qui ne permet pas d'obtenir des valeurs *p* faibles.

De ce fait, il est préférable de considérer l'ensemble des textes offerts par le corpus. Ceci peut s'effectuer de deux manières : soit en mettant en commun tous les textes pour les considérer comme un seul et vaste jeu de données (ce qui revient à augmenter n), soit en calculant l'observable sur chacun des textes, pour en considérer ensuite la moyenne au niveau du corpus. C'est cette seconde option que nous avons privilégiée, pour les raisons évoquées plus haut : les comportements individuels sont trop disparates (à commencer par la longueur des textes) pour qu'une hypothèse nulle calculée à partir de l'ensemble des textes puisse capturer efficacement les spécificités linguistiques de chacun. Pour générer la distribution correspondante à l'ensemble des textes individuels, il faut alors se munir de M copies alternatives du corpus, où chaque texte est aléatoirement segmenté suivant la structure de probabilité qui lui est propre.

4.2 Application aux occurrences de *et*

Notre méthode ainsi définie, nous pouvons l'appliquer aux 381 occurrences de *et* intraphrastiques et aux 94 occurrences de *et* extraphrastiques, quant à leur positionnement par rapport aux jets textuels. Nous rappelons que ces occurrences peuvent se situer en début de jet, en fin de jet, en milieu de jet, ou composer un jet à elles seules. On considère que l'occurrence est en début de jet (resp. en fin de jet) si le premier (resp. le dernier) caractère alphanumérique de l'occurrence coïncide avec le premier (resp. le dernier) caractère alphanumérique du jet. Il existe une cinquième possibilité, celle d'une occurrence à cheval entre deux jets, mais nous n'en avons pas observé dans notre corpus. Dans le cas de *et*, cette possibilité correspond exactement à trouver le *e* dans un premier jet et le *t* dans un second ;

ce type de production est absent de nos données, en raison probablement de l'unité constituée par le mot, ainsi que l'aisance à le traduire mécaniquement au clavier.

De prime abord, il semble attendu que la plupart des occurrences se retrouvent en milieu de jet : c'est ce que prévoit une répartition purement au hasard des pauses, *et* étant un mot particulièrement court comparé aux 35 événements composant un jet en moyenne. Pour cette raison, la significativité des observations doit être évaluée vis-à-vis d'une distribution de référence, ici générée par la méthode précédemment présentée. Ce qui importe donc, ce n'est pas la répartition des occurrences entre les quatre catégories identifiées, mais la mesure dans laquelle la répartition observée s'écarte de ce à quoi l'on est en droit de s'attendre compte tenu de certaines caractéristiques basiques de la segmentation en jets textuels (notre hypothèse nulle).

Avant de présenter l'analyse des deux types d'occurrence, intraphrastique et extraphrastique, on peut faire remarquer que leur répartition au niveau du groupe est significativement différente. Par souci de simplicité, la notion de répartition est ici ramenée à la proportion ρ d'occurrences à l'intérieur du jet *versus* à la frontière du jet. Pour évaluer la significativité de cette différence, nous avons d'abord posé l'hypothèse nulle que les deux sous-populations d'occurrences sont deux échantillons indépendants d'une même population. Cette hypothèse nulle est obtenue en calculant la probabilité ρ^* qui maximise le log-likelihood de l'observation de chacune des proportions intraphrastiques et extraphrastiques, respectivement ρ_E et ρ_I. La valeur p de la différence entre les deux proportions (ρ_E - ρ_I) est alors directement donnée par la probabilité jointe des deux distributions binomiales de même paramètre ρ^* correspondant à chacune des deux sous-populations concernées. Il en ressort que la différence est hautement significative (p = 0.004), ce qui justifie de traiter séparément les deux types d'occurrence lors de l'analyse.

4.2.1 Occurrences intraphrastiques

Les résultats concernant les occurrences intraphrastiques sont récapitulés dans le Tableau 2. Nous avons répertorié les ratios mesurés pour la segmentation réelle du texte, l'intervalle de confiance associé pour chacun d'eux dans la distribution Monte-Carlo (les intervalles de confiance correspondant à l'intervalle entre les quantiles 0,025 et 0,975), et la valeur p, c'est-à-dire la probabilité d'obtenir sous l'hypothèse nulle une valeur plus extrême que le ratio observé. On y lit ainsi que, si la très large majorité des occurrences se retrouvent en milieu de jet (78%), cette répartition est attendue compte tenu de notre hypothèse nulle. En revanche, les occurrences intraphrastiques se retrouvent plus souvent au début et plus rarement à la fin des jets que ce que prévoit une répartition aléatoire.

Tableau 2. Répartition des occurrences intraphrastiques vis-à-vis des jets

Position	**Pourcentage observé**	**Intervalle de confiance Monte-Carlo**	**Valeur *p***
en début de jet	18%	[7%– 14%]	0.0003
en milieu de jet	78%	[76%– 85%]	0.21
en fin de jet	3,9%	[4,1%– 11%]	0.03
Seule	0%	[0%– 2%]	0.59

Ce résultat peut être interprété comme la manifestation de la possibilité qu'a cette conjonction d'adjoindre un élément additionnel à un schéma déjà présent. En effet, une pause qui accompagne l'adjonction d'un nouveau contenu (comme ici l'élément introduit par *et*) peut s'expliquer de deux manières : soit il y a volonté de la part du locuteur d'adjoindre, mais le contenu reste à déterminer ; soit le contenu est activé cognitivement, mais le locuteur doit mettre en œuvre une stratégie discursive pour l'intégrer au discours. Or, selon la première interprétation, on devrait observer des *et* en fin de jet, ce qui est significativement peu probable : la seconde interprétation, selon laquelle la conjonction permet de résoudre le problème de l'intégration, est donc favorisée, même si elle reste à discuter dans une optique plus qualitative.

Un résultat empirique supplémentaire permet de souligner la fonction cohésive du *et* intraphrastique. Pour chaque occurrence, nous avons calculé le quantile correspondant à la longueur du jet associé, c'est-à-dire sa position dans la distribution de toutes les longueurs du jet du texte considéré. Pour chaque texte/scripteur, nous avons ensuite moyenné ce quantile sur toutes les occurrences ; puis nous avons pris la moyenne de ce quantile moyen au niveau du groupe, sur tous les scripteurs. Le résultat obtenu est 0.81, c'est-à-dire que les jets où figurent les occurrences intraphrastiques de *et* se situent en moyenne aux quatre cinquièmes de la distribution ; il n'y a, en moyenne, qu'un cinquième de jets plus longs. Cependant, ce résultat peut être l'effet d'un mécanisme simple : un jet long abrite plus de mots qu'un jet court, donc pour n'importe quelle unité linguistique considérée, il est plus probable de la rencontrer dans un jet long que dans un jet court. Pour pouvoir interpréter ce résultat, nous l'avons rapporté à la distribution obtenue sur M segmentations aléatoires et alternatives des textes du corpus. L'intervalle de confiance est [0.73 ; 0.78] et la valeur p associée à l'observation empirique est de 0.0001. Le résultat est confirmé comme extrêmement significatif. La présence d'une occurrence intraphrastique de *et* est donc bien associée à des jets plus longs en moyenne, ce qui renforce l'interprétation de cette conjonction comme favorisant la continuité de la production en agglomérant des *chunks*.

4.2.2 Occurrences extraphrastiques

Les résultats concernant les occurrences extraphrastiques sont récapitulés dans le Tableau 3. On y retrouve le résultat précédent selon lequel la conjonction se trouve plus fréquemment en début de jet. En revanche, il apparaît que les emplois extraphrastiques sont moins fréquemment intégrés au jet qu'attendu. De plus, les occurrences isolées du reste de la production, composant un jet à elles seules, se retrouvent dans une proportion significativement élevée.

Tableau 3. Répartition des occurrences extraphrastiques vis-à-vis des jets

Position	**Pourcentage observé**	**Intervalle de confiance Monte-Carlo**	**Valeur *p***
en début de jet	36%	[13%– 29%]	0.0007
en milieu de jet	52%	[61%– 80%]	0.0001
en fin de jet	2%	[1%– 12%]	0.12
seule	10%	[0%– 1%]	0.0002

Ces résultats sont plus difficiles à démêler. Les emplois extraphrastiques de *et* apparaissent comme hétérogènes vis-à-vis de la production langagière : leur faible proportion en milieu de jet (comparativement à une distribution aléatoire) indique qu'ils participent plus volontiers des phénomènes de frontière. Le détail des occurrences montre qu'il peut s'agir d'occurrences introduisant une nouvelle phrase (*et* semblant alors servir de marqueur discursif par défaut), ou de l'articulation d'idées disparates, comme ici :

Exemple 7
Cette décision nous concerne tous [pause] et [pause] notre Rédaction vous expose ici [pause] les différents points de vue que tout à chacun pourrait avoir [pause] de cette mesure.

D'une manière révélatrice, la pause la plus longue de cette séquence se situe *après* l'occurrence du *et* extraphrastique ; le *et* ici semble donc indiquer la volonté d'élaborer sur ce qui a été dit, alors même que l'apport informationnel est encore à déterminer. Le *et* extraphrastique permet alors tout à la fois d'indiquer une volonté de poursuivre le propos, et de délayer la réalisation d'un énoncé encore en construction. Parmi les occurrences en début de jet, un nombre important d'entre elles sont immédiatement précédées d'une révision ; la conjonction, en particulier dans ces emplois extraphrastiques, apparaît alors comme un appui automatisé permettant la reprise de la production après interruption.

Par ailleurs, nous avons également considéré le quantile moyen de la longueur des jets associés aux occurrences de *et* extraphrastique (moyenné au niveau des occurrences pour chaque scripteur, puis au niveau du groupe). Celui-ci est de 0.74, à comparer avec l'intervalle de confiance sous hypothèse nulle [0.71 ; 0.77]. Le résultat n'est pas significatif, avec une valeur p de 0.84. Inversement, cette absence de résultat indique en creux que les occurrences extraphrastiques ne jouent pas non plus un rôle disruptif dans la production (autrement, on s'attendrait à des jets plus courts qu'attendu sous hypothèse nulle). Cette observation, couplée aux préférences du *et* extraphrastique pour les frontières de jet, renforce donc l'interprétation de cette conjonction comme permettant une reprise de la production, jouant ainsi le rôle d'amorce pour le contenu produit après la pause.

5. Conclusion

Nos résultats permettent d'abord de vérifier que la segmentation pausale du flux de production écrite ne relève pas d'un processus aléatoire, uniquement sensible à des considérations structurelles d'ordre très général, mais est bien sensible au contenu linguistique associé. Pour le montrer, nous avons étudié la conjonction *et*, dont le rôle discursif polyfonctionnel est reconnu, en considérant sa répartition vis-à-vis des frontières des jets textuels. Nous avons élaboré une hypothèse nulle regroupant l'ensemble des facteurs de segmentation les plus élémentaires, afin de voir si une telle hypothèse permettait de couvrir les observations relatives à la forme choisie. Il est apparu que le comportement de *et* vis-à-vis de la segmentation textuelle, que ce soit dans ses emplois intraphrastiques ou extraphrastiques, déroge aux prédictions induites par l'hypothèse nulle et se révèle être spécifique de cette forme. La segmentation pausale empiriquement attestée ne peut pas être expliquée par les facteurs élémentaires considérés et n'est pas indépendante du contenu linguistique. Cela confirme la valeur épistémique des jets relativement aux processus de production textuelle, en tant qu'unité d'analyse linguistique pertinente.

La distinction entre emplois extraphrastiques et intraphrastiques nous a permis de mettre en évidence une certaine corrélation entre le degré de dépendance sémantico-grammaticale et la préférence pour un emplacement à l'intérieur du jet textuel : ainsi, les occurrences en emploi extraphrastique sont peu intégrées aux jets, montrent une préférence accrue pour

l'actualisation en borne gauche et comptent également des actualisations autonomes, à la différence des occurrences en emploi intraphrastique. Les deux types d'occurrence reflètent ainsi des usages distincts de la conjonction : les premières servent d'appui discursif pour s'engager à élaborer le propos, ou le reprendre après une interruption du processus de production, les secondes permettent d'intégrer un contenu déjà là et de renforcer la cohésion du contenu. Quoi qu'il en soit, cette étude montre l'éclairage mutuel entre, d'une part, la considération du contenu linguistique et la segmentation du processus de production : le contenu linguistique permet de saisir ce qui motive la segmentation, en même temps que le rôle joué dans cette segmentation informe quant au fonctionnement sémantique et pragmatique des unités linguistiques.

Les perspectives ouvertes par cette étude sont avant tout méthodologiques, visant à tirer profit de différentes approches statistiques afin de résoudre des problèmes d'identification et de qualification des unités d'analyse linguistique potentielles. De manière indirecte, c'est la segmentation, principe de base de l'analyse linguistique, qui est ainsi questionnée. Enfin, des analyses qualitatives plus fines, prenant en compte les valeurs spécifiques de la conjonction – et de toute autre observable linguistique, le cas échéant – permettront de compléter les résultats obtenus en fournissant des jalons pour la motivation linguistico-cognitive de la segmentation.

Cette étude a bénéficié du support de l'Agence Nationale de la Recherche dans le cadre du projet Pro-TEXT ANR-18-CE23-0024-01.

Références bibliographiques

Alves, R. A., Castro, S.L., Sousa, L., Strömqvist, S. (2007). Typing skill and pause-execution cycles in written composition. In M. Torrance, L. Van Waes, D. Galbraith (G. Rijlaarsdam éditeur de série) *Writing and Cognition, research and applications*. Dordrecht : Elsevier Sciences Publishers, p. 55-65.

Bilger, M. (1999). Coordination : analyses syntaxiques et annotations. *Recherches sur le français parlé, 15*, 255-272.

Bronckart, J.-P., Schneuwly, B. (1984). La production des organisateurs textuels chez l'enfant. In M. Moscato, G. Pieraut-Le Bonniec (éds) *Le Langage : construction et actualisation*. Rouen : P.U.R., p. 165-178.

Charaudeau, P. (1992). *Grammaire du sens et de l'expression*. Paris : Hachette.

Christiansen, M. H., Chater, N. (2016). *Creating Language: Integrating Evolution, Acquisition, and Processing*. Cambridge : The MIT Press.

Cislaru, G., Olive, T. (2016). Les automatismes du scripteur : jets textuels spontanés dans le processus de production écrite, le cas des constructions coordinatives. *Congrès Mondial de Linguistique Française*. Tours, 5 juillet 2016.

Cislaru, G., Olive, T. (2018). *Le Processus de textualisation*. Bruxelles : De Boeck.

Doquet, C. 2011. *L'Écriture débutante*. Rennes : Presses universitaires de Rennes.

Doquet, C., Leblay, C. (2014). « Temporalité de l'écriture et génétique textuelle : Vers un autre métalangage ? » *Congrès Mondial de Linguistique Française*, Berlin, Allemagne. SHS Web of Conférences - Vol. 8 - 4e Congrès Mondial de Linguistique Française.

Dragsted, B. (2005). Segmentation in translation: Differences across levels of expertise and difficulty. *Target*, *17:1*, p. 49-70.

Favart, M., Chanquoy, L. (2007). Les marques de cohésion comme outils privilégiés de la textualisation : une comparaison entre élèves de CM2 et adultes experts. *Langue Française, 155(3)*, p. 51-58

Favart, M., Passerault, J.-M. (1995). Evolution du rôle fonctionnel des connecteurs et de la planification du récit écrit chez les enfants de 7 à 11 ans. *Revue de Phonétique Appliquée, 115-117*, p. 198-212.

Favart, M., Passerault, J.-M. (1999). Aspects textuels du fonctionnement et du développement des connecteurs : approche en production. *L'Année Psychologique, 99*, p. 149-173.

Fayol, M. (1986). Les connecteurs dans les récits écrits : Étude chez l'enfant de 6 à 10 ans. *Pratiques, 49*, p. 101-113.

Foulin, J.-N. (1995). Pauses et débits : les indicateurs temporels de la production écrite. *L'Année Psychologique, 95*, p. 483-504.

Gobet, F., Lane, P., Croker, S., Cheng, P., Jones, G., Oliver, I., Pine, J.. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences, 5 (6)*, p. 236–243.

Grésillon, A. (2016 [1994]). *Éléments de critique génétique*. Paris : CNRS Éditions.

Hobaek-Haff, M. (1987). *Coordonnants et éléments coordonnés*. Copenhague – Paris : Solum Forlag – Didier Erudition.

Immonen, S., Mäkisalo, J. (2010). Pauses reflecting the processing of syntactic units in monolingual text production and translation. *Journal of Language and Communication Studies, 44*, p. 45–61.

Lang, E. (1984). *Semantics of Coordination*. Amsterdam : John Benjamins.

Leblay, C., Caporossi, G. (2015). A graph theory approach to online writing visualization. In G. Cislaru (éd) *Writing(s) at the CrossRoads*, Amsterdam/Philadelphia : John Benjamins Publishing Company, p. 171-181.

Lebrave, J.-L. (1987). *Le Jeu de l'énonciation en allemand d'après les variantes manuscrites des brouillons de H. Heine*. Thèse de Doctorat d'État : Université Paris-Sorbonne.

Leijten, M., Van Waes, L. (2013). Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes. *Written Communication, 30*, p. 358– 392.

Matsuhashi, A. (1981). Pausing and planning: The tempo of written discourse production. *Research in the Teaching of English, 15 (2)*, p. 113-134.

Medimorec, S., Risko, E. F. (2017). Pauses in written composition: on the importance of where writers pause. *Reading and Writing: an Interdisciplinary Journal, 30*, p. 1267–1285.

Mouchon, S., Fayol, M., Gombert, J.E. (1991). L'emploi des connecteurs dans les récits : une tentative de comparaison oral/écrit. *Repères, 3*, p. 87-98.

Olive, T. (2012). Writing and working memory: A summary of theories and of findings. In E. Grigorenko, E. Mambrino, D. Preiss (éds) *Handbook of Writing: A mosaic of new perspectives*. New York : Psychology Press.

Paolacci, V., Favart, M. (2010). Traitement des marques de cohésion par les jeunes scripteurs : l'utilisation de la ponctuation et des connecteurs à l'entrée en sixième. Approche linguistique, cognitive et didactique. *Langages, 177*, p. 113-128.

Renner, V. (2007). Coordination formelle et subordination sémantique dans le lexique : l'exemple du l'hendiadys en anglais. In D. Banks (éd.) *La coordination et la subordination dans le texte de spécialité*. Paris : L'Harmattan, p. 77-84.

Riegel, M., Pellat J.-C., Rioul, R. (2005 [1994]). *Grammaire méthodique du français*. Paris : Presses Universitaires de France.

Rousseau, A. (2007). La coordination : approche méthodologique, critique et raisonnée des questions essentielles. In A. Rousseau, L. Begioni, N. Quayle, D. Roulland (éds) *La Coordination*. Rennes : Presses universitaires de Rennes, p. 18-57.

Sag, I. A. (2005). La coordination et l'identité syntaxique des termes. *Langages,* 160, p. 110-127.

Schilperoord, J. (1996). The distribution of pause time in written text production. In G. Rijlaarsdam, H. Van den Bergh, M. Couzijn (éds) *Theories, Models and Methodology in Writing Research*. Amsterdam : Amsterdam University Press, p. 542-556

Schneuwly, B., Bronckart, J.-P. (1986). Connexion et cohésion dans quatre types de textes d'enfants. *Cahiers de linguistique française, 7*, p. 279-294.

Sinclair, J., Mauranen, A. (2006). *Linear Unit Grammar: Integrating Speech and Writing*. Amsterdam – Philadelphia : John Benjamins.

Spelman Miller, K., Sullivan, K.P.H. (2006). Keystroke logging: an introduction. In K. P. H. Sullivan, E. Lindgren (éds) *Computer Keystroke Logging: Methods and Applications*. Oxford : Elsevier, p. 1–9.

Van Hell, J., Verhoeven, L., van Beijsterveldt, L. (2008). Pause time patterns in writing narrative and expository texts by children and adults. *Discourse Processes, 45*, p. 406–427.

Wengelin, Å. (2006). Examining pauses in writing: Theory, methods and empirical data. In Kirk Sullivan, Eva Lindgren (éds) *Computer key-stroke logging and writing*, Brill, p. 107-130.

[1] S'agissant de données de production enregistrées en temps réel, de nombreuses scories orthographiques ou grammaticales peuvent persister ; nous avons fait le choix de citer les données brutes, sans correction.

[2] Si le texte présente m jets et M mots, alors il existe 2m emplacements correspondant à une frontière de jet (avant ou après la pause), et la probabilité pour un mot d'être à une frontière de jet est donnée par p = 2m/M. La probabilité d'observer, parmi les n occurrences de *et* correspondant au texte, k de ces occurrences en frontière de jet, est alors donnée par la probabilité binomiale $B(k,n,p) = C(k,n) * p^k * (1 - p)^{n-k}$, où C(k,n) est le coefficient binomial. De là, on peut alors évaluer l'hypothèse nulle selon laquelle les occurrences de *et* sont distribuées au hasard relativement aux frontières des jets. Si on note $S(k,n,p) = \Sigma_{k'\leq k} B(k',n,p)$, la distribution cumulative associée à la loi binomiale, alors l'hypothèse nulle est préférée si la valeur observée d'occurrences k^* en frontières de jet appartient à l'intervalle de confiance [k_{min} ; k_{max}], où $S(k_{min},n,p) = 0.025$ et $S(k_{max},n,p) = 0.975$. On peut généraliser ce raisonnement à des catégories multiples (à l'intérieur d'un jet, en début de jet, en fin de jet, etc.), le test devenant alors multinomial et non plus simplement binomial.

[3] Les relations entre les mots, à l'intérieur des phrases, constituent une observable importante mais non exploitable dans l'état actuel du corpus, qui est en phase d'annotation morphosyntaxique et en dépendances.

[4] Notons une difficulté technique : une pause après un mot peut survenir avant ou après la frappe du caractère d'espacement (de même qu'une pause associée à un événement de ponctuation peut intervenir avant ou après celui-ci). Pour pouvoir prédire la pause dans l'un et l'autre cas, il faudrait donc coder 1 l'événement précédant l'espace, et coder 1 l'événement correspondant à l'espace, ce qui affaiblirait d'autant le pouvoir prédictif de ce facteur. Pour y remédier, le vecteur *Y* est alors manipulé pour que les pauses coïncidant avec un espacement aient toujours lieu après ce caractère ; ainsi seul l'événement d'espacement est codé 1, sans que cela ne laisse de côté des événements où la pause a eu lieu avant l'espace (puisqu'on l'a artificiellement déplacée après cette espace).

[5] Une autre limitation technique est posée par les révisions. De manière générale, les événements de révision se succèdent immédiatement ; par exemple pour effacer un mot, il est nécessaire de répéter le caractère de suppression autant de fois qu'il y a de lettres dans ce mot. Le pouvoir prédictif du facteur « l'événement a lieu avant une révision » serait donc faible si tous ces événements étaient conservés. En pratique, des événements de suppression successifs sont donc résumés en un seul (ce qui revient à ne pas considérer un événement comme « avant révision » si cet événement est lui-même une révision).

[6] En effet, la valeur p correspond à la probabilité d'obtenir sous l'hypothèse nulle une observation plus extrême que celle considérée. Supposons que l'observation réelle se situe dans le quantile q de la distribution. La valeur p sera alors égale à 2 * q ou 2 * (1- q), selon que l'on considère la probabilité d'observer une proportion plus faible que la proportion réelle, ou plus élevée que la proportion réelle (la multiplication par 2 permet d'assurer que le test est bilatéral). Notons ici que la valeur p minimale est fixée par le nombre de segmentations M, la valeur p la plus petite étant égale à 2* 1/M (ici 0.0001), dans le cas où l'observation réelle est en-deçà de l'observation la plus basse sur les segmentations aléatoires (où au-dessus de l'observation la plus haute).