Clitic Subjects as Landmarks in the Writing Production Process: A Study based on a Keylog-derived Corpus of Writing Bursts

Quentin Feltgen[1] & Florence Lefeuvre[2]

[1] Ghent University, Ghent | Belgium

[2] University Sorbonne Nouvelle, Paris | France

Abstract: Bursts of writing, extracted from online recordings of the writing process, have proved an invaluable vantage point into the cognitive mechanisms at work during written language production. Crucially, they show that writers, much like speakers, produce language through a sequence of small 'chunks', patterns-like groupings of words that do not necessarily match the structures of theoretical grammars. As such, they are intriguing objects, whose linguistic properties are yet to be understood. To contribute to this endeavor, we track all instances of French so-called clitic subjects in a corpus of 81 keylogs of short essays written by undergraduate students in experimental conditions. We show that these clitic subjects are attracted to the burst-initial position, favoring resumption of the production after revision events. Moreover, they also act like discursive hubs in that writers are more likely to revise up to a clitic subject and restart from there, possibly relying on an entirely different structure. Therefore, they play the role of landmarks in the writing process, from which information can flow, and to which writers can get back to develop alternative discursive strategies. These results hint that the writing process and the information structure of the product are likely to be intimately intricated.

Keywords: writing process, bursts, clitic subjects, revisions, product-and-process, keylogging

journal of
WRITING RESEARCH

Feltgen, Q, & Lefeuvre, F. (2025). Clitic subjects as landmarks in the writing production process: A study based on a keylog-derived corpus of writing bursts. *Journal of Writing Research*, *16*(3), 433-460. https://doi.org/10.17239/jowr-2025.16.03.04

Contact: Quentin Feltgen, Ghent University, Department of Linguistics, Blandijnberg 2, 9000 Gent | Belgium – quentin.feltgen@gmail.com - https://orcid.org/0000-0001-7369-075X

1. Introduction

Writing relates the cognitive dynamics of the process, which involves planning, translating, and revising, and the material of the product, in which meaning is cast into linguistic structures. The dynamics of the writing process are known to follow specific patterns; in particular, they involve an alternance of cognitive-related pauses and short streaks of production called bursts (Alves & Limpo, 2015). It also features extensive revisions that may intervene at different points of the textualization process, rely on different triggers, and target different types of structures (Conijn et al., 2022). These dynamics can be thoroughly explored through the use of keylogging, that is, live recordings of the timestamps of each keystroke during a typing process (Strɯmqvist et al. 2006). The temporal characteristics of these keylogs, and the Inter-Key Intervals (IKI) in particular, in turn reflect the cognitive mechanisms that underlie the writing process (Galbraith & Baaijen, 2019). However, the relationship between the linguistic product and the process' dynamics have only recently started to become the focus of scientific investigation (Cislaru, 2015). Preceding works have in particular explored the linguistic contents of the bursts (Cislaru & Olive, 2018a) and how revisions interact with the information structure of the product (Bowen & Van Waes, 2020; Bowen & Thomas, 2020).

In this paper, we offer to go beyond these pioneering works by focusing on the idiosyncratic behavior of the French clitic subjects in a keylogging corpus, both with respect to bursts and to revision events. To understand in a finer way the sensitivity of the bursts to these specific linguistic units, we consider the clitic subjects' attraction toward the different positions offered in a burst. We then offer a rigorous Monte-Carlo-based statistical method to provide a reference distribution to which the actual observation can be compared. To better characterize the relationship between clitic subjects and the complex revision and reformulation events that reshape the linguistic material, we define a number of possible roles with respect to these revision events and evaluate the attraction of the clitics toward each of these roles. By combining these two series of findings, we show that clitic subjects are landmarks in the writing process, from which alternative discursive strategies can be explored, and that offer a stepping stone for textual production to resume.

2. Background and research questions

2.1 The writing production process

The writing process has been described by the still popular model of Flower & Hayes (1981), that consists in a set of interacting cognitive processes, namely planning, translating, and reviewing. Planning pertains to goal-setting, generating ideas and overall discourse strategy; translating refers to the articulation of thoughts into language;

reviewing highlights the ability to monitor and revise the past output of the writing process. On top of this, grapho-motor skills are required for the written output to be actually produced. The latter has been made more explicit in a refined version of the model by Chenoweth and Hayes (2001) that now includes four “modules”, a proposer (that generates a concept), a translator (that turns it into an utterance), a reviser (that evaluates the utterance), and a transcriber (that actually produces it). However, the relationship between these components is neither linear nor sequential, and their relative importance vary during the production process (Kellogg, 1987). Notably, these models all highlight the fundamentally dynamical nature of writing.

One of the key features of these dynamics is that the writing process does not flow homogenously, but proceeds through an alternance of pauses and production bursts (Matsuhashi, 1981), that often relate to clauses (Kaufer et al., 1986), although more recent evidence suggests that these bursts do not necessarily correspond to ‘saturated’ (=phrase-type) syntactic segments, both with respect to constituent-style grammar (Cislaru & Olive, 2018b) and Construction Grammar (Gilquin, 2020). Bursts may also overlap with prefab units (Mutta & Salminen 2021), also called lexical chunks or lexical bundles, that is, set of words that are entrenched as such and correspondingly uttered in a holistic way (Blumenthal-Dramé, 2017). However, these prefab units only account for a limited proportion of the bursts – about 5% in (Cislaru & Olive, 2018b). Therefore, although the bursts contents are generally believed to influence the duration of the preceding pause – what Schilperoord coins the Butterworth’s pause paradigm (1996: 294) –, there is no clear consensus regarding how and which information is packaged into production bursts.

Alternative ways have been offered to model the production process so as not to rely on the behavioral alternance of pauses and bursts. For instance, in a sentence-centric perspective (Ulasik et al. 2025), the production process is akin to a text history, described as a series of operations that transform the states of the sentences that made up the text. In this view, successive micro-events in the linear production sequence are grouped insofar as they target the same sentence and act on it in the same way (revision, deletion, insertion, addition, etc.)

2.2 Bursts of writing and linguistic units

Although the bursts have emerged as dynamical units in the writing process, with the potential to reflect and inform on the underlying cognitive mechanisms of that process, they have recently received additional spotlight, due to their alleged relevance in laying the basis of an overall description of language (for a recent review, see Vasylets & Marín 2025, this issue). According to Linear Unit Grammar (Sinclair & Mauranen 2006), language is articulated into “chunks” of information that are linearly encoded into speech or writing. Note that this notion of a chunk does not coincide with the lexical chunks defined above but is a reference to the idea that information is handled by repackaging it into larger units, called chunks in Miller’s foundational paper (1956). These chunks constitute, according to Sinclair & Mauranen, the foundational units of a language's grammar. Even though they are not properly defined, Mauranen argues that they can be partly identified in oral speech through repetitions, rephrasing, and dysfluencies (2016).

That bursts of writing may be considered as linguistic units has already been advocated (Olive & Cislaru 2015). Cislaru & Olive (2018a) have thus equated writing bursts to "linguistic units of writing performance". This identification between production bursts and linguistic units draws from Chafe's hypothesis about speech that "each intonation unit verbalizes the information active in the speaker's mind at its onset" (1994: 63), and that these intonation units correspond to "linguistic expressions of information" (1994: 69), although he stresses that intonation units are only one possible way of segmenting language into units, alongside segmentations according to phonemes, syllables, words, sentences, etc. (1994: 58).

This relationship between intonation units and linguistic (more precisely, syntactic) units has been more thoroughly investigated by Degand & Simon (2009). They distinguish three types of discourse units: a) one intonation unit is composed of several syntactic units, b) one syntactic unit is produced across several intonation units, and c) one intonation unit exactly matches one syntactic unit. From different oral corpora, they show that a potential mismatch between the two (e.g. an intonation unit overlaps two syntactic units but its boundaries coincide with neither of them) is fairly rare (only 5% of the total). In the same spirit, the interaction between disfluencies (discourse markers, filled pauses, and unfilled pauses) and clauses has been investigated through a monitoring of the disfluencies' positions in both clauses and dependency units (Crible et al. 2017), showing a clear attraction to clause boundaries. Despite these possible relationships, prosodic and syntactic units have mostly acted as two different, and possibly conflicting, perspectives on textual segmentation (Lefeuvre & Moline 2011).

Even though one cannot rely on intonation to segment the writing production process into units, the interspersed pauses that define the bursts of writing are a close counterpart to the oral disfluencies. Therefore, these results from the oral literature hint at a possible correspondence between the writing production units and the linguistic units lying at the core of Linear Unit Grammar.

2.3 Clitic subjects in French

In this study, we decided to focus on French clitic subjects to study their behavior with respect to the writing process. The clitic subjects form a closed set and can be easily listed: *je*, *tu*, *il*, *elle*, *on*, *nous*, *vous*, *ils*, *elles*. Clitic subjects, referred to as such for phonetic reasons – Le Goffic (1993: 27) also refers to them as "atonal" –, are a phenomenon circumscribed to a small number of Romance languages (Poletto & Tortora 2016). They need to obey the tight constraint that they are only produced joint to a conjugated verb (Feltgen et al. 2023): they cannot occur in isolation (although *elle(s)*, *nous*, *and vous* are identical to their tonic counterparts) and only a restricted set of units can separate the clitic subject and the verb (other pronouns, adverbial or personal, and the expletive negation marker *ne*, all of these being clitics as well). In that sense, they differ from the full personal pronouns that are found in Germanic languages.

Their status has been heavily debated (De Cat 2005), between a “syntactic” reading as full-fledged subjects and a “morphological” one, according to which the clitic subjects actually play the role of agreement markers. The latter reading is reinforced by the phenomenon of subject reduplication, that is, a subject clitic may be produced between the verb and a preceding noun phrase with which it shares the same referent, and therefore seemingly plays no role but that of a verbal agreement marking (Culbertson & Legendre 2008). However, these observations are based on oral data: in written language, agreement marking at the end of the verbs is still effective and reduplication is marked, even in more spontaneous registers like texting (Stark 2013).

The choice to focus on clitic subjects is motivated by a variety of reasons. First of all, we wanted to focus on the sensitivity of the bursts to the syntactic function of the subjects. However, the data cannot be properly annotated through automatic tools, insofar as it presents too much spelling variations, incomplete syntactic structures, and an intricate structure of revisions and rephrasing; besides, the dataset is too large for manual annotation to be a workable option. As such, syntactic information cannot be reliably tracked. The clitic subjects, however, can be readily identified based on their surface form: only a small subset of them have homonymic counterparts (namely *elle(s)*, *nous*, and *vous*), and those four forms remain infrequent enough that they can be easily sorted out manually.

Moreover, encompassing all subjects would yield methodological and conceptual issues. Indeed, the range of complexity of subjects widely differ from one instance to another (see examples 1-4, drawn from our corpus): the subject noun phrase may be enriched with a genitive (1), multiple adjectives (2), a pragmatic revision of the adjective (3), or a subordinate clause (4). Since our perspective is that of the production process, this variety of situations would incur too wide a variety of cognitive processes and cognitive efforts, obscuring the relationship between the linguistic role of these units and their processual properties. Even if we focused on simple [determiner + noun] phrases alone, the retrieval effort associated with the lexical search would be a critical confounder to our analysis. Clitics therefore allows to investigate the subject role while bypassing the issue of the processing effort diversity displayed by the full range of subjects.

(1) l'augmentation du tabac
the [cost] increase of tobacco

(2) Les transports en commun moins cher ou bien gratuit
free or cheaper public transportation

(3) certaines firmes nationales, voir internationales
some national, or even international, companies

(4) la première disposition qui a été mise en place
the first provision to be implemented

Furthermore, the French language is highly observant of the Theme-Rheme structure, at least in speech where it seems to drive the intonation structure (Le Goffic 1993: 14). The literature already suggests that the bursts of writing may be sensitive to the theme-rheme partition (Bowen & Van Waes 2020), in that revisions target themes preferentially. Since clitics act as minimal theme, whether this preference for revising themes holds with subject clitics is certainly an intriguing research venue.

2.4 Research questions

In this study, we address the following research questions:

RQ 1: Are the bursts sensitive to the linguistic properties of their contents?

We tackle this question by focusing on subject clitics, in order to assess whether these elements, defined by their syntactic role, follow some idiosyncratic behavior with respect to the bursts, which would in turn reflect a sensitivity of the burst toward the linguistic material they produce.

RQ2: What role do the syntactic components play in the revision processes?

To answer this, we investigate the interplay between subject clitics and revisions, to assess the extent to which clitic subjects elicit revisions (due to the phrase boundary between the clitic and the verb) or serve as landmarks when revising (the written product is revised up to a clitic to restart from there).

3. Methods

3.1 Corpus and occurrence extraction

The empirical study of bursts of writing has considerably benefitted from live data recording software, be it for handwriting with the use of HandSpy (e.g. Alves & Limpo 2015, Limpo & Aves 2017), or for typing with the use of Inputlog (Leijten & Van Waes 2013). However, bursts remain a theoretical construct, and need to be properly extracted. In Inputlog, the software records timestamps of each keystroke, which allows to collect the IKI (Inter-Key Intervals). Bursts are commonly defined based on a set threshold: if an IKI is longer than this threshold, then it is considered a pause. Bursts are immediately derived from there as the output of the writing process between two such pauses.

The proper setting of this threshold has been much debated. The common consensus is to consider a 2s threshold, notably to ensure comparability across studies (Wengelin 2006). This threshold has been nonetheless criticized for its lack of adaptability to the variability among writers (Dragsted 2005), especially when non-typical populations are considered (e.g. children). Moreover, basic IKI features may vary across tasks (Conijn et al. 2019), and IKI located at key syntactic boundaries (e.g. sentence ends) may often fall below the 2s threshold (Medimorc & Risko 2017). To overcome these issues, attempts have been made to derive an individualized threshold for each participant, for each text. To achieve so, the key idea is to rely on the underlying features of the IKI distribution, with the assumption that one could distinguish several components corresponding to different kinds of cognitive processes. These have nonetheless proven unsuccessful so far (Baaijen et al. 2012, Hall et al. 2022): although a two or three--modes Gaussian mixture seems to be a convincing fit, the models are either inconsistent across individuals (three modes models) or lead to define much shorter pauses than what has been typically considered as such in the literature (two modes models).

In our study, we only consider undergraduate students (therefore a homogenous population, typical of psycholinguistic studies), that engages in only one task (so the reference is the same for all textual productions). Accordingly, we should not be overly affected by the population heterogeneity and task variability issues. Nevertheless, our data relies on threshold individualization to some extent, although still in keep with the 2s reference, as detailed in section 3.1.2.

3.1.1 Keylogs recording

We rely on experimental data collected by Bouriga (2020) from undergraduated students in Psychology, who performed the writing task as part of their evaluation. The task consisted in a short prompt asking them to write an essay on a given topic to which they were assumed to be familiar. The topics ranged from tobacco smoking restrictions to road safety regulation. They had 15 minutes to write the essay on a computer keyboard. The timestamps of each keystroke, as well as non-keyboard events, were recorded thanks to the Inputlog software (Leijten & Van Waes 2013). 81 texts have been recorded in this way, from as many different participants.

3.1.2 A corpus of bursts

We did not, however, directly worked with the keylogs corpus, but relied instead on a bursts corpus derived from the former, and due to Olive & Bouriga (2022). A burst is a string of keyboard events such that the timestamp difference between each two consecutive events (the so-called IKI) is lower than a set threshold. An IKI greater than this threshold is then deemed a pause in the production process. Note that non-keyboard events may still be interspersed among the keyboard events. The pause-defining threshold have been individualized to an extent while setting a common 2s reference, in accordance to the usual value in the literature (Rшnneberg et al. 2022). All IKI data are pooled together across all 81 participants and the quantile corresponding to 2s is extracted from that pooled distribution. This quantile is then applied to each individual participant. Therefore, although they have different thresholds, they all share the same proportion of pauses.

The corpus totals 240,000 events, 33,760 words, and 6,409 bursts.

3.1.3 Bursts representation

Each text is represented by a series of chronologically-arranged bursts, featuring all keyboard events that compose the burst, including spacing events (␣) and revision events, marked by one or several pressings of the backspace key (⌫), as illustrated in (5):

(5) Pourtant␣⌫,␣o⌫comme␣tou⌫⌫⌫⌫⌫⌫⌫⌫⌫il␣ne␣faut␣pas␣oublier
que␣cette␣plante␣est␣aussi␣ue⌫ne␣drogue␣

Here, the characters "␣", "o", "comme␣tou", and "e" have been deleted, so the resulting sentence added to the text would be (6):

(6) Pourtant, il ne faut pas oublier que cette plante est aussi une drogue
Yet, one must not forget that this plant is also a drug.

Note that the bursts are arranged according to the chronology of the production process, which is not necessarily equivalent to the textual chronology of the product: writers may, for instance, come back to a previous point in the text to add or modify content. In that case, bursts that are produced later (and indexed as such in our corpus) will contribute to parts of the text that may come prior in the final product than the contribution of earlier bursts.

3.1.4 Occurrence extraction

We then extracted all occurrences of clitic subjects from the corpus of bursts. The list of clitic subjects is the following: *je* (first person singular), *tu* (second person singular, informal), *il* (third person singular, masculine; non-referential clitic), *elle* (third person singular, feminine), *on* (generic third person singular), nous (first person plural), *vous* (second person plural; second person singular, formal), ils (third person plural, masculine/mixed), *elles* (third person plural, feminine). All these forms are non-ambiguous and code solely for a subject function, with the exception of *elle(s)*, *nous* and *vous* which may also stand for their clitic object or tonic counterparts. These occurrences (69 out of 849 occurrences) have been disambiguated manually.

For each occurrence, we recorded its position within the burst: the occurrence can be split (if the alphabetic characters of the occurrence span more than one burst), alone (if all alphabetic characters of the burst belong to the occurrence), or in the beginning (resp. in the end) of the burst if the occurrence is not split and if there is no alphabetic character before (resp. after) it in the burst. Finally, the occurrence is within the burst in all other cases (all alphabetic characters of the occurrence are contained within the burst and there is at least one alphabetic character in the burst both before and after the occurrence).

3.2 Significance of the clitic subjects positioning

3.2.1 Issue

One key issue with the bursts is how difficult it may be to assess the statistical significance of an observation. For instance, let us consider the proportion of capped letters found at the beginning of bursts. This proportion turns out to be quite high, simply because a full stop often entails a pause. Therefore, a simple Fisher test (capped/non-capped letter vs. at the beginning/not at the beginning of a burst) would be highly significant, even though the observation is trivial based on the most basic segmentation features of the text (by text, we refer, here and in what follows, to the chronological sequence of characters during the production, not to the product itself). To remedy this, we follow a Monte-Carlo-based approach. We generate a large number (20,000) of alternative segmentations of the text into bursts, accounting for a certain number of basic features (to be described below in 3.2.3.), and then we repeat the observation on these alternative segmentations. This allow us to build a distribution to which the actual observed value can be compared, in order to assess its significance.

Note that the minimum p-value is set by the number of random alternative segmentations: if the observation is above or below that of all the random alternatives, then we indicate that the p-value is below 2/20,000, that is, below 0.0001.

3.2.2 Model

To each keyboard event, we associate a variable $\boldsymbol{y}$ equal to 1 if the event is followed by a pause, and 0 otherwise. We also associate to it a vector x coding for features that we detail below. The text is then described by a vector Y coding for all pauses and a matrix X coding for all features, for all keyboard events. We then fit a logistic regression model $Y \sim s(\beta X)$, where $s(.)$ is the sigmoid function and β the model parameters, the weight of the features that are fitted according to the logistic regression. The vector $s(\beta X)$ is therefore a vector of probabilities – the probability for a pause to occur after each keyboard event. From then on, we can sample this probability to generate an alternative vector $\tilde{Y}$, which provides an alternative segmentation of the text following the model.

3.2.3 Pause-inducing factors

We consider a number of factors that may favor the occurrence of a pause:

- Baseline: active after each keyboard event. This is the most basic hypothesis, according to which the segmentation would be entirely random.
- Between words: active between two words. A pause between words may be due, for instance, to planning and lexical retrieval. One difficulty is that the pause may be diversely located (e.g. before or after the typing of the spacing character). To harmonize things, we moved all spacing character beginning a burst at the end of the previous burst; in the situation where multiple spacing characters had been typed

sequentially, we kept only one. A spacing character that is not both preceded and followed by an alphabetic character is not considered as a between words event. The spacing characters surrounding punctuation marks of any kind have also been removed.

- Soft punctuation: active after a soft punctuation mark (comma, colon, semicolon, dash, parenthesis).
- Sentence end: active after a period or any sentence-ending punctuation mark (exclamation and interrogation marks).
- Revision: active prior to a revision event. In case of multiple successive revision events (which occur very often, any time when more than a letter is deleted at the same time), we condensed the whole string of revision events into a single one.
- Resumption: active after a revision event (as defined above)

These factors are exemplified in Table 1 for the following sequence:

(7) ,et␣jouerait␣un␣rôle␣é⌫certzin⌫ain␣ro⌫ôe⌫le␣écologique.
, and would play a role some ecological role.

Note that the spacing character after the comma has been deleted due to our burst manipulation (a spacing character remains only when no other separating character is present), and that a single revision character (⌫) may stand for the deletion of an entire sequence (rôle␣é). The revisions mostly pertain to spelling considerations, except the first one, where ***écologique*** seems to have been initiated before the writer decided to downplay the "ecological role" by deleting everything up to the determiner ***un*** and adding ***certain*** ('***un certain***' = 'some').

One caveat is that, in this scheme, within word pauses are only due to the baseline probability of pausing, which may be an oversimplifying hypothesis considering that morphological boundaries are likely to trigger pauses as well, especially so in synthetic languages (Ivaska et al., 2025). In French, however, this issue remains limited, although verbal and nominal agreements may elicit pauses (cf. the first pause to occur in the *Y*-line in Table 1).

3.2.4 Factor strengths

We display on Table 2 the strength (***β***-weight) of each of the model's factors and represent them on Figure 1. A *β*-weight greater than 0 indicates that the activation of the factor increases the probability to pause, as compared to the baseline probability. All these factors are highly significant at the group level, with the exception of the Soft Mark factor (notably, it is affected by 7 outliers that produced very few soft marks and never introduced a pause afterward, leading to a highly negative *β*-weight; if these 7 outliers are removed, then the factor is significantly influential).

Table 1. Coding pause-inducing factors over a text excerpt.

Character	**,**	**e**	**t**	**␣**	**j**	**o**	**u**	**e**	**r**	**a**	**i**	**t**	**␣**	**u**	**n**	**␣**	**r**	**ф**	**l**	**e**	**␣**	**й**	**⌫**	**c**	**e**	**r**	**t**	**z**	**i**	**n**	**⌫**	**a**	**i**	**n**	**␣**
Y	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
baseline	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
between	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**
soft mark	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
hard mark	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
revision	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**
resumption	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**

Character	**r**	**o**	**⌫**	**ф**	**e**	**⌫**	**l**	**e**	**␣**	**й**	**c**	**o**	**l**	**o**	**g**	**i**	**q**	**u**	**e**	**.**
Y	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**
baseline	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
between	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
soft mark	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
hard mark	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**
revision	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
resumption	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

Table 2. β-weights of the different factors impacting pause probability in our model

	Baseline	Between words	Soft mark	Hard mark	Revision	Resumption
Mean	-5.1	2.4	1.6	5.6	3.6	2.6
Standard deviation	0.4	0.5	5.9	2.7	0.5	0.6
Minimum	-6.2	1.1	-18.0	2.8	2.4	1.0
Maximum	-4.4	3.7	5.1	27.7	5.0	3.9
Confidence level	[-5.9; -4.3]	[1.4; 3.4]	[-10.2; 13.4]	[0.2; 11.0]	[2.6; 4.6]	[1.4;3.8]
Probability to pause	0.007	0.07	0.16	0.57	0.20	0.09
Pause proportion	14%	29%	4%	11%	28%	14%

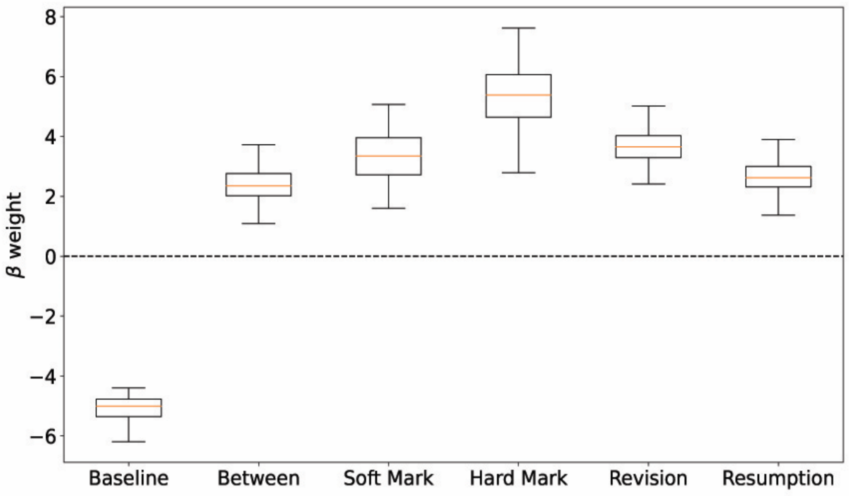

Figure 1: Boxplots of the β-weights of the different factors that impacts pause probability in our model. The box itself represents the Inter-Quartile Range and the line within the box shows the median. All points further away from the box by more than 1.5 times the Inter-Quartile Range are considered outliers and not represented. This includes 7 points for the Soft Mark factor and 1 point for the Hard Mark factor (for which the β -weight is much higher).

The strongest factor is the hard mark, leading to a pause probability above 50 %. This is a highly relevant feature in the model for the study of clitics, since subjects tend to be produced as the first element in a sentence in French. The revision factor is high as well. As for the between words factor, it is associated to a probability of only 7%; however, since this factor is very often active, it is the one that accounts for the largest proportion of pauses (29%), on par with the proportion of pauses due to revision processes (28%).

Interestingly, the 'between words' and the 'resumption' factors have very similar statistical features; they are, furthermore, highly correlated ($r = 0.56$, $p = 6e^{-8}$). This hints that resuming production after a deletion sequence is not fundamentally different from a between words context.

3.2.5 Group level

Each individual text has been fitted separately by the model. For each of the five burst-related positions (cf. *supra* 3.1.4.), we recorded the ratio of clitic subjects' occurrences in that position and averaged these ratios at the group level. Next, we generated an alternative random segmentation for each text, computed the ratios for each of the five positions, and averaged them on the group level. We repeated this random generation 20,000 times. This way, we generated a distribution for these five group-level-averaged ratios under the set of individual models of text segmentation. We therefore assess the significance of the group level averages, not the significance of the ratios at an individual level.

Note that an alternative approach could have been used, consisting in designing a model directly at the group level, including random effects to account for individual variations. However, it is unclear which random effects should be considered on top of the fixed effects, and this would require an extensive model selection analysis.

3.3 Annotation

3.3.1 Annotation values

The results that we will outline in section 4 show a strong presence of clitic subjects in the beginning position. To understand this affinity, we decided to syntactically annotate the content immediately preceding bursts starting with a clitic subject. The annotation variable could take the following values:

— end of a sentence: the clitic subject begins a new sentence;

(8) Il␣est␣mis␣⌫⌫⌫⌫⌫⌫⌫⌫n'est␣⌫⌫⌫⌫⌫⌫est␣⌫⌫⌫⌫peut␣être␣mis␣à␣disposition␣en␣libre␣accès␣dans␣les␣rues,␣les␣lycées,␣d⌫⌫⌫␣ou␣dans␣les␣grandes␣surfaces.␣**On**␣peut␣donc␣y␣avoir␣accès␣f⌫gratuitement.
[The condom] (is made) > (is not) > (is) > can be made freely available in streets, high schools, or in supermarkets. One may then freely access to it.

— juxtaposition: the clitic subject begins a new main clause, but directly follows another main clause by concatenation, rather than starting a new sentence proper;

> (9) Il␣m'apparaît␣que␣les␣r⌫⌫⌫⌫⌫ce␣système␣ne␣réduit␣pa␣⌫s␣de␣manière␣significative␣la␣problématique␣de␣la␣vitesse,␣**elle**␣,n⌫⌫ne␣fait␣que␣dép⌫⌫⌫la␣déplacer.
> It seems to me that this system does not significantly reduce the speed issue, it only shifts it further away.

— conjunctions (mostly with et, 'and');

> (10) Cependant,␣**il**␣y␣aurait␣surement␣des␣cons"⌫équences␣sur␣le␣marché␣du␣cannabis␣.
> Nonetheless, there would certainly be consequences for the cannabis market

— relative pronoun: the clitic subject is then the subject of a subordinate clause (relative pronouns can be subjects of their clause, but then they are never followed by a clitic subject);

> (11) d'autres␣pensent␣au␣contraire␣qu'**elle**␣n'est␣pas␣utile.
> Others, on the other hand, think it is not useful.

— subordinate clause: the clitic subject is then the subject of the main clause;

> (12) si␣tous␣les␣étudiants␣bénéficient␣de␣ce␣moyen,␣**il**␣y␣aura␣un␣surplus␣de␣population␣dans␣less␣tr⌫⌫⌫⌫␣transport étudiantes␣s␣en␣commun
> If all students benefit from this [transportation] mean, there will be a population increase in the public transportation for students.

— framing adverbials, as defined by Charolles & Vigier (2005), e.g. de nos jours ('nowadays');

> (13) Au␣vu␣de␣sa␣nocivité␣pour␣la␣société;␣**il**␣et⌫st␣important␣de␣faire␣ce␣qui␣est␣nécessaire␣pour␣amener,␣du⌫oucement,␣une␣trabs⌫⌫nsition␣générale.
> Given its harming factor for society, it is important to do whatever necessary to smoothly bring an overall change.

— revision event (any occurrence of a deletion);

(14) il␣est␣utilisé⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫il␣est␣donc␣relativement␣méconnu␣ quant␣aux␣maus⌫x␣qu'il␣induit.
(it is used) > it is therefore not so well known with respect to the harm it causes

— discontinuity.

The latter category occurs whenever there is a mismatch between the continuity of the process and that of the product; that is, whenever the segment that is immediately produced prior to the clitic in the chronology of the actual production belongs to another section of the text. This may happen, e.g., if the writer stops, revises an earlier segment, and then restarts where they had left.

An underlying hypothesis of this annotation scheme is that each of these values corresponds to a different cognitive operation, e.g. producing a framing adverbial, ending a sentence, or performing a deletion sequence. Therefore, our annotation aims at capturing the latest cognitive event preceding the production of the clitic (even though cognitive events that do not translate in a keyboard interaction may take place in between). As such, these annotation values are mutually exclusive.

This manual annotation has been furthermore double checked by three different annotators.

3.3.2 Automatization

The annotation cannot be automatized: the raw content of the burst is too heterogenous (multiple spelling errors, revisions, etc.) to be automatically annotated. However, to provide a comparison point, we could automatize the annotation of full sentences and revision events. To do so, we consider the last character produced before the clitic (in the situations where the clitic occupies the beginning position of the burst it belongs to); we skip spacing characters, soft and medium punctuation marks. If the last character is a strong punctuation mark (resp. a backspace keystroke), the occurrence is categorized as preceded by a sentence's end (resp. a revision event).

Note that this automatized annotation differs from the manual annotation; in particular, since the randomly generated segmentation relies on a simplified version of the bursts where revision events have been contracted (e.g. ⌫⌫⌫ > ⌫), some minor revision events (e.g. the deletion of an extra spacing character) are not considered relevant in the manual annotation, but they cannot be sorted out easily in the automatized annotation. This issue remains nonetheless unimpactful as long as the output of the automated annotation over the randomly generated alternative segmentations is only compared to the output of the same automated annotation procedure over the original segmentation, instead of comparing it to the (more reliable) output of the manual annotation.

3.4 Relationship with revisions

There are different ways with which a linguistic item may interact with revisions: they can follow a revision (resumption), be part of a revised segment, and in that case, they may trigger the revision (the revision starts immediately after they have been produced), be a landmark for the revision (the text is revised up to and including the linguistic item), or serve as a boundary for the revision process to stop (the revision stops short of deleting the item). The linguistic item may also be produced as part of the segment replacing what has been revised (that is, beyond the resumption position), but we shall not investigate this latter possibility, mostly because of the fundamental difficulty to properly distinguish replaced contents from genuinely new production building upon it.

3.4.1 Clitics as resumption

Clitics may immediately follow a revision event and be used to resume production. However, what counts as a revision event is debatable. For instance, a large part of revision events are immediate spelling corrections (e.g. smarth⌫phone), and therefore the production process needs not be resumed in this case. We therefore define a revision event (for this analysis only) as any deletion of at least two words; concretely, of at least two non-zero sequences of alphabetic characters separated by at least a spacing character (the separation may also include punctuation marks). The clitic is considered to be the element that resumes production if no other alphabetic character is produced between the end of the revision event and the production of the clitic.

To assess the significance of the relationship between clitics and resumptions, we performed an exact two-sided Fisher test pitting clitic subjects vs. all other words (a word is any sequence of alphabetic characters between two non-alphabetic characters), and resumption events vs. any other word production events. Furthermore, we tested how robust the results were by varying this definition and considering at least two, and at least three, separate spacing characters in the sequence.

3.4.2 Revision landmark

Similarly, we considered instances where the clitic subject is part of the deleted sequence, sticking to the same definition of a revision event. We specifically distinguished the case where the clitic is the last element deleted (the revision runs "up to the clitic"), in which case we say that the clitic is the *landmark* of the revision event. Similar to resumption, we tested the results with different definitions of what constitutes a revision event (deletion of at least two, at least three, or at least four words). We also assessed significance with an exact two-sided Fisher test.

This *landmark* situation is illustrated in the occurrence below:

(15) Pour␣une⌧majorité⌧de⌧jeunes⌧de⌧notre⌧société,⌧le⌧cannabis␣à␣⌫⌫a␣
un⌧usage␣ré créatif,␣il␣est␣utilisé⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫il␣est␣donc␣
relativement⌧méconnu⌧quant⌧aux⌧maus⌫x␣qu'il⌧induit.
For a majority of young people in our society, cannabis has a recreative use, (it is used) > it is therefore relatively poorly known with respect to the drawbacks it entails.

In this example, the revision deletes the whole rheme up to the clitic ***il*** (which is deleted as well), and restarts with the same clitic, referencing the same element.

We may also consider the boundary word, that is, the word that stands just prior to the deleted sequence. For instance, in the following:

(16) **ils␣s'ince**⌫⌫⌫⌫⌫⌫peuvene⌫⌫nt⌧s'incrire⌧entre⌧200␣à␣400␣⌫euros
they (range) > may range between 200 and 400 euros
the clitic *ils* is the boundary of the revision (the element before which the writer stops to restart production).

In our analysis, we shall consider both cases, landmark and boundary.

3.4.3 Revising after a clitic

We also considered the situation where the clitic subject is revised immediately after being produced. This is especially relevant to test the cohesion with the verbal element: if the clitic subject is produced cohesively with the verb, much like an agreement marking, then production should not get interrupted by a revision immediately after the clitic and prior to the verb. We consider that an immediate revision occurs as long as no alphabetic character is produced between the end of the clitic production and the start of the revision sequence, and as long as all the alphabetic characters composing the clitic were deleted. To account for this, we require a spacing character to be included anywhere in the deleted sequence but the last position, or a spacing character to appear in the 'boundary' position immediately prior to the deleted material.

To assess the significance of the relationship between clitics and immediate revisions, we compared this (through an exact two-sided Fisher test) to the behavior of several sets: the set of two-characters word, the set of four-character word or less, and the set of all words (defined as mentioned above). This way, we account to some extent for the limited size of the clitics, which may otherwise increase the likelihood of deleting them entirely in a single event.

4. Results

4.1 Burst sensitivity to clitics

We collected 780 occurrences of clitic subjects in our corpus, and recorded their positions within the bursts, as reported in Table 3.

The clitic subject occurrences are mostly found within burst, as expected: there are far more within burst positions available than any other positions, as indicated by the expected values (between 0.68 and 0.74 with a 95% probability).

Table 3. Observed group level averaged ratios of occurrences falling in each of the five positions, and their comparison with the confidence interval of a distribution of random segmentations.

Position	**beginning**	**within**	**end**	**alone**	**split**
Mean ratio	0.24	0.70	0.03	0.02	0.0
95% CI	[0.15; 0.21]	[0.68; 0.74]	[0.06; 0.10]	[0.00; 0.03]	[0.0; 0.02]
***p*-value**	< 0.0001	0.59	< 0.0001	0.61	0.01

The two main results, though, are the significantly high (resp. low) ratios of occurrences at the beginning (resp. at the end) of the bursts. Note that these ratios cannot be attributed to the fact that subjects often initiate sentences, since the model already accounts for the sentence-end higher probability of pausing when generating random segmentations. The scarce presence of the clitics in the burst-final position may also reflect a strong level of cohesion with the verb that usually follows it.

To understand better the high proportion of clitic subjects at the beginning of the bursts, we annotated the last element produced in the burst immediately preceding the clitic subject, as described in 3.3.1. The results of the annotation are shown in Table 4.

From the table, it appears that in 59% of the cases (sentence end, juxtaposition, and conjunction), the clitic initiates a new clause; in 10% of the cases, the clitic is only preceded by a framing element (subordinate clause or framing adverb). In 28% of the cases, the clitic corresponds to a resumption of the production after a deletion (23%), or when resuming to that point in the text after the production of a segment somewhere else in the text (5%).

Table 4. Proportion of syntactic roles found prior to the production of clitic subjects in beginning position.

Syntactic role	Ratio
Sentence end	0.40
juxtaposition	0.07
conjunction	0.12

relative pronoun	0.03
subordinate clause	0.07
framing adverb	0.03
revision	0.23
discontinuity	0.05

The proportion of 'excess' clitic at the beginning of bursts should be somewhere between 12.5% and 37.5% (based on the [15%-21%] confidence interval for the expected ratio of burst-initial clitic subjects in Table 3, and with 12.5 = (24-21)/24*100 and 37.5 = (24-15)/24*100). Therefore, the occurrences due to juxtaposition and conjunction, tallying to 19% of these occurrences, can explain this excess proportion, since neither juxtaposition nor conjunction are accounted for in the random segmentation model. This observation suggests that pauses are more sensitive to clauses than they are to sentences. In that sense, the bursts are effectively sensitive to their linguistic contents.

We now compare the results of the automatic annotation (see 3.3.2.). For the original segmentation, we find a proportion of 34% of clitic subjects in beginning position preceded by a sentence end, and a proportion of 35% preceded by a revision (for the mismatch between the manual annotation and the automatic annotation, see 3.3.2.). Over the randomly generated alternative segmentations, these proportions become equal to 39% (CI: [33%-45%]) and 34% (CI: [28% - 40%]). Therefore, the observed proportions do not deviate significantly from their random counterpart.

Given the previous conclusion, this is actually puzzling: we know that, in the occurrences of the original segmentation, a good chunk of them are due to a pause at the clause level, which the model typically does not capture and assumedly ignore. Therefore, starting from the random distribution, if we add then these extra occurrences, the proportion of occurrences after a sentence end and after a revision should drop to 30% (CI: [24%-36%]) and 26% (CI: [20%-33%]). Hence, the observed proportion of sentence ends before burst-initial clitics (34%) is consistent with the hypothesis, but the observed proportion of revisions (35%) is abnormally high.

Therefore, the only interpretation consistent with all these results is that the abnormally high burst-initial proportion of clitics is due to both a sensitivity to clause boundary in the production process, and to a abnormally high proportion of revisions preceding the burst-initial clitics. In the next section, we investigate the latter more thoroughly.

4.2 Clitics relationship with revisions

4.2.1 Resumption

Clitic subjects are found in a resumption situation (first produced word after a revision event) in a proportion greater than expected: 28 of them were expected for the whole corpus and 55 of them were found, with a corresponding Fisher's test p-value of $2.3e^{-6}$. This holds even if we restrict the set of revision events to those deleting three words or more (46 clitic objects are found in a resumption position, while only 22 are expected on average, with a Fisher's test p-value of $3.1e^{-6}$), and to those deleting four words or more (23 found, 11 expected, $p = 0.002$).

This finding is in line with the hypothesis of the previous section, where we mentioned that the proportion of revisions before burst-initial clitics was likely higher than normal. Even though we consider all clitic subjects here (not only those in burst-initial position), the observation holds that these linguistic item have an affinity for resumption contexts. Since a pause is likely after a revision event (see 3.2.4.), resumption often initiates a new burst, hence the high proportion of clitic subjects in the burst-initial position due to a prior revision event.

4.2.2 Revision landmark

Since we consider, in this section, revision events that delete a sequence of character spanning at least two words, it is worth considering where does the deletion stops – both with respect to the last element erased (the 'landmark' element), and with respect to the element in front of which the revision stopped (the 'boundary' element).

The clitic subject is found 58 times in the 'landmark' position ($p = 2.5e^{-7}$), and 32 times in a boundary position ($p = 0.43$), while 28 of them were expected in both cases. This result holds when considering only larger revision elements, spanning at least three words: 22 expected, 42 landmark ($p = 7.1e^{-5}$), 30 boundaries ($p = 0.10$); or at least four words: 11 expected, 24 landmark ($p = 0.0007$), 18 boundaries ($p = 0.07$).

This asymetry between the landmark and the boundary positions (although it tends to disappear when increasingly large revision events are considered) is certainly intriguing. Before trying to interpret it, an easy explanation that would have nothing to do with the linguistic properties of clitic subjects must be ruled out. Indeed, writers often delete the first word of a sentence when they failed to produce a capital letter in the first place. Since clitic subjects may often initiate a sentence, there could be a large number of 'landmark' deletions that only corresponds to a case correction. However, we found only 2 such replacement among our 'landmark' occurrences, so leaving these two aside would not change the significance of the result.

Therefore, there is a marked tendency to delete linguistic material up to (and therefore including) the clitic subject, while there is no particular preference for stopping deletion in front of clitic. By contrast, there is a significant tendency to stop in front of a definite article ($p = 0.01$). By surveying the occurrences of deletion up to a clitic, it appears that in many cases, the writing then branches off to an entirely new discursive strategy. For instance:

(17) On␣sait␣qu'il␣y␣a␣des␣abus␣
⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫Des⌧gens⌧en⌧meurt⌧⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫␣
meurent␣tous␣les␣ans␣à␣cause␣de␣cela␣⌫
(We know this may lead to excessive) > People (die from it) > die every year because of it.

Coupled with the previous observation that clitic subjects facilitate resumption, they furthermore appear as 'landmarks' in the writing process, in the sense that they signal specific loci of the text from where it is convenient for the writer to restart, and from where new discursive strategies may arise.

4.2.3 Revising after a clitic

We now consider whether the writing production process is likely to be disrupted by a revision after a clitic subject, which would then be immediately deleted. Since shorter words are more likely to be entirely deleted, we first consider only the revision events where the first deleted word is only 2 letters. Among these, 15 are clitic subjects, while 26 of them were expected based on the number of 2-letters clitic subjects, so this is significantly low (p = 0.01). Similarly if we only consider revisions such that the first deleted word is four letters are less (the longest clitic subject, *elles*, is 5-letter long, but they account for less than 3% of the total of subject clitics, so most of them are 4-letter long or less), then we found 24 of them, but 42 are expected, which is again significantly low (p = 0.002). Finally, if we put no constraint on the length of the first deleted word, we find the same 24 clitic subjects deleted, which is significantly less than the 38 expected (p = 0.01).

This result offers an interesting contrast with the previous ones. Indeed, this time, the clitic subjects are significantly infrequent in that position; it means that the clitic subjects facilitate the production flow. This aligns well with the previous finding that they are a landmark in revision sequences, in the sense that writers delete linguistic material up to them to restart from a clitic subject. It also resonates with the fact that clitic subjects are seldom found at the end of the writing bursts.

This result offers an interesting contrast with the previous ones. Indeed, this time, the clitic subjects are significantly infrequent in that position; it means that the clitic subjects facilitate the production flow. This aligns well with the previous finding that they are a landmark in revision sequences, in the sense that writers delete linguistic material up to them to restart from a clitic subject. It also resonates with the fact that clitic subjects are seldom found at the end of the writing bursts.

5. Discussion

Our study has relied on two major ways to assess the disfluencies and dynamics of the writing process: the segmentation of the process into bursts, and the revision events. We focused on the behavior of selected linguistic items, the clitic subjects, which act as minimal theme elements in French. Our results show a consistent picture: clitic subjects are significantly attracted toward the burst-initial position and repelled from the burst-final one. This hints at a facilitating role in production resumption. Furthermore, when they appear in the burst-initial position, they are often preceded by a revision event, which reinforces the conclusion that they facilitate resumption. The high proportion of peripheral framing elements before the pause also hints that the clause plays a stronger role than the sentence in the writing production dynamics.

Subject clitics also show an idiosyncratic yet consistent behavior with respect to revisions. They are often found immediately after a revision event, irrespectively of their relative position to the burst boundaries, which confirms their role as a production resumption facilitator. Similarly, production seldom stops after a clitic subject, and revision events immediately afterward are significantly rare. Finally, revisions often run until they reach a clitic subject and delete it. Interestingly, they do not significantly stop short of deleting it, but go all through the way of erasing it. One may consider that this is the flipping side of the propension to restart with a clitic subject after a revision; however, a closer look at the actual examples shows that these revisions often lead to a completely different utterance strategy, possibly involving a full-fledged noun phrase subject instead of a clitic. This observation suggests that the use of the clitic subject ties to the translating component of the writing process, since the replacing sequence typically carries a similar meaning than the one that got deleted (there is no change with respect to the planning component).

Moreover, writers engage in highly complex editing operations when they revise their text. Production may occasionally be a process of ebb and flow, with many different linguistic structures being considered and written before settling on the final one, as the following example illustrates:

(18) Il␣y␣a␣aussi␣de⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫␣␣Le␣préservatif␣⌫⌫⌫⌫⌫⌫⌫⌫⌫
⌫⌫⌫⌫⌫⌫Il⌧existe⌧le⌧préservatif␣féminin⌧et⌧le⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫⌫masculin,⌧et⌧
dep⌫⌫⌫⌫,␣depuis⌧quelques⌧années,⌧il⌧existe⌧le⌧préservatif⌧féminin⌧
également

(There is also) > (The condom) > There exists a (female condom and a) > male condom, and since a few years, there exists a female condom as well.

The purpose of this sentence is to highlight the existence of a female condom alongside the male one. Yet, this example illustrates how writers may face uncertainty with respect to the theme/rheme structure of the sentence, framing the female condom first as a rheme by relying on an impersonal structure (assuming *des préservatifs féminins*, 'a female condom', is indeed the initially intended continuation of the first deleted segment), next as the theme of the sentence (here again, assuming the condom of the second segment is the female one), and finally as the rheme again, in an anaphoric conjunction with the male condom.

In these complex reformulation processes, writers need to rely on specific linguistic "hubs" where they can come back to branch off more easily towards a different strategy. In many respects, the choice of a clitic subject already engages in such a strategy; they may be used in a referential way to assess thematic continuity; they may be used in an impersonal way to introduce epistemic value; they may engage the writer in giving their own opinion with the use of firstperson pronouns. As such, clitic subjects act as landmarks in the writing process. Interestingly, this specific feature of the production dynamics has also been identified for oral language (Blanche-Benveniste 2010), which suggests that, despite the differences between the two media (especially for the clitic subjects, that may have a different status in speech), their respective translation strategies may rely on similar patterns.

Our study has, nevertheless, several limits. The first one is intrinsic to the specific type of data that we used – keylogs. Keylogs data are extremely difficult to handle appropriately, especially when writers jump across different points in the text, sometimes back and forth within a burst. Since results can only achieved at the statistical level, a high degree of automatization is required, yet the range of phenomena and textual oddities that may occur makes the analysis difficult and unreliable at times. We tried to check insofar as possible that the output of the automatization was close to what a manual extraction would have done, but discrepancies may have persisted.

The second limit ties to the choice to focus on clitic subjects. We may especially wonder whether subjects, in that they are mostly clause-initial in French, play a similar role. Since clitic subjects may be swapped for a full subject after a revision event (and the reverse is true as well), there is good reason to expect that subjects would also be attracted to the burst-initial position and may facilitate resumption after a revision event. However, being less versatile than the clitic subjects, they probably offer less flexibility with respect to the discursive strategy. More broadly, we may wonder to which extent the a-thematicity of the clitic subject makes it a specific tool as compared to more fully realized expressions of the theme. In any way, these results show in a very clear way that both the bursts of writing and the revision sequences, which are highly distinctive markers of the dynamics of the writing process, are also clearly sensitive to the linguistic material they contribute to produce and reshape.

6. Conclusion

In this paper, we focused on French clitic subjects, minimal realizations of the thematic component of an utterance, to study their behavior with respect to the writing process. We showed that these clitic subjects favor the burst-initial position and are more frequent than expected when the writers resume their production after a deletion sequence. They also seldom lead to an immediate revision and are averse to the burst-final position. This shows that both the bursts of writing and the revisions, which are empirical signatures of the writing process, are sensitive to the linguistic structure of the writing product. We also showed that revision events are attracted to the clitic subjects and tend to delete linguistic contents up to this element. This suggests that the clitic subject acts as a landmark in the textual process, from which the writer can branch off toward alternative translating strategies.

Acknowledgements

This research is part of the ANR Research Project Pro-TEXT (ANR-18-CE23-0024-01). Quentin Feltgen acknowledges funding from Ghent University (BOF.PDO. 2022.0001.01). The authors warmly thank Dominique Legallois for his instrumental scientific input and for his contribution to the annotation procedure.

Data availability statement

The keylogs data and the bursts have been collected within the Pro-TEX ANR Research Project and are available upon request from the project website (https://pro-text.huma-num.fr/).

References

Alves, R. A., & Limpo, T. (2015). Progress in written language bursts, pauses, transcription, and written composition across schooling. *Scientific Studies of Reading*, *19*(5), 374-391. https://doi.org/10.1080/10888438.2015.1059838

Baaijen, V. M., Galbraith, D., & De Glopper, K. (2012). Keystroke analysis: Reflections on procedures and measures. *Written Communication*, *29*(3), 246-277. https://doi.org/10.1177/0741088312451108

Blanche-Benveniste, C. (2010). Lexique et grammaire dans les reformulations [Grammar and lexis in reformulations]. In M. Candea & R. Mir-Samii (Eds.), *La rectification à l'oral et à l'écrit. Hommage à Marie-Annick Morel* (pp. 77-89). Ophrys.

Blumenthal-Dramé, A. (2017). Entrenchment from a psycholinguistic and neurolinguistic perspective. In H.-J. Schmid (Ed.), *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge* (pp. 129–152). De Gruyter Mouton. https://doi.org/10.1037/15969-007

Bouriga, S. (2020). Papier-crayon vs. йcran-clavier : Effets sur le coыt cognitif et sur la dynamique de la production de textes [Pen and paper vs. keyboard and screen : Effects on cognitive demand and on text production dynamics]. [Unpublished doctoral dissertation]. https://theses.hal.science/tel-03795276v1

Bowen, N. E. J. A., & Thomas, N. (2020). Manipulating texture and cohesion in academic writing: A keystroke logging study. *Journal of Second Language Writing*, *50*, 100773. https://doi.org/10.1016/j.jslw.2020.100773

Bowen, N., & Van Waes, L. (2020). Exploring revisions in academic text: Closing the gap between process and product approaches in digital writing. *Written Communication*, *37*(3), 322-364. https://doi.org/10.1177/0741088320916508

Chenoweth, N. A., & Hayes, J. R. (2001). Fluency in writing: Generating text in L1 and L2. ***Written communication***, ***18***(1), 80-98. https://doi.org/10.1177/0741088301018001004

Chafe, W. (1994). ***Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing***. University of Chicago Press.

Charolles, M., & Vigier, D. (2005). Les adverbiaux en position prйverbale: portйe cadrative et organisation des discours [Adverbials in preverbal position: framing scope and discourse organization]. ***Langue française***, ***148***, 9-30. https://doi.org/10.3917/lf.148.0009

Cislaru, G. (Ed.). (2015). ***Writing (s) at the crossroads: The process–product interface***. John Benjamins Publishing Company. https://doi.org/10.1075/z.194

Cislaru, G., & Olive, T. (2018a). ***Les processus de textualisation : Analyse des unités linguistiques de performance écrite*** [Theg textualization processes: Analysis of written performance units]. De Boeck.

Cislaru, G., & Olive, T. (2018b). Bursts of written language as performance units for the description of genre routines. In Legallois, D., Charnois, T., & Larjavaara, M. (Ed.), The Grammar of Genres and Styles (pp. 220-248). De Gruyter Mouton. https://doi.org/10.1515/9783110595864-010

Conijn, R., Roeser, J., & Van Zaanen, M. (2019). Understanding the keystroke log: the effect of writing task on keystroke features. ***Reading and Writing***, ***32***, 2353-2374. https://doi.org/10.1007/s11145-019-09953-8

Conijn, R., Speltz, E. D., Zaanen, M. V., Waes, L. V., & Chukharev-Hudilainen, E. (2022). A product- and process-oriented tagset for revisions in writing. ***Written Communication***, ***39***(1), 97-128. https://doi.org/10.1177/07410883211052104

Crible, L., Degand, L., & Gilquin, G. (2017). The clustering of discourse markers and filled pauses: A corpus-based French-English study of (dis)fluency. ***Languages in Contrast***, ***17***(1), 69-95. https://doi.org/10.1075/lic.17.1.04cri

Culbertson, J., & Legendre, G. (2008). Qu'en est-il des clitiques sujet en franзais oral contemporain ? [What about clitic subjects in Present-Day Spoken French?]. Durand, J., Habert, B., & Laks, B. (Ed.) ***Congrès Mondial de Linguistique Française - CMLF'08*** (pp. 2663-2674), Institut de Linguistique Franзaise. https://doi.org/10.1051/cmlf08308

De Cat, C. (2005). French subject clitics are not agreement markers. ***Lingua***, ***115***(9), 1195-1219. https://doi.org/10.1016/j.lingua.2004.02.002

Degand, L., & Simon, A. C. (2009). On identifying basic discourse units in speech: theoretical and empirical issues. ***Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics***, ***4***. https://doi.org/10.4000/discours.585

Dragsted, B. (2005). Segmentation in translation: Differences across levels of expertise and difficulty. ***Target. International Journal of Translation Studies***, ***17***(1), 49-70. https://doi.org/10.1075/target.17.1.04dra

Feltgen, Q., Lefeuvre, F., & Legallois, D. (2023). Sujet clitique et dynamique de l'йcrit : un йclairage par les jets textuels [Clitic subjects and writing dynamics: A perspective based on bursts]. ***Discours***, 32. https://doi.org/10.4000/discours.12509

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. ***College composition and communication***, ***32***(4), 365-387. https://doi.org/10.2307/356600

Galbraith, D., & Baaijen, V. M. (2019). Aligning keystrokes with cognitive processes in writing. In Lindgren, E., & Sullivan, K. (Ed.), ***Observing writing*** (pp. 306-325). Brill. https://doi.org/10.1163/9789004392526_015

Gilquin, G. (2020). In search of constructions in writing process data. ***Belgian Journal of Linguistics***, ***34***(1), 99-109. https://doi.org/10.1075/bjl.00038.gil

Hall, S., Baaijen, V. M., & Galbraith, D. (2024). Constructing theoretically informed measures of pause duration in experimentally manipulated writing. ***Reading and Writing***, ***37***(2), 329-357. https://doi.org/10.1126/science.aac4716

Ivaska, I., Toropainen, O., & Lahtinen, S. (2025). Pauses during a writing process in two typologically different languages. *Journal of Writing Research*, *16*(3), 405-431. https://doi.org/10.17239/jowr-2025.16.03.03

Kaufer, D. S., Hayes, J. R., & Flower, L. (1986). Composing written sentences. *Research in the Teaching of English*, 121-140. https://www.jstor.org/stable/40171073

Kellogg, R. T. (1987). Effects of topic knowledge on the allocation of processing time and cognitive effort to writing processes. *Memory & cognition*, *15*, 256-266. https://doi.org/10.3758/BF03197724

Lefeuvre, F., & Moline, E. (2011). Unitйs syntaxiques et unitйs prosodiques: Bilan des recherches actuelles [Syntactic and prosodic units: Synthesis of current research]. *Langue française*, (2), 143-157. https://doi.org/10.3917/lf.170.0143

Le Goffic, P. (1993). *Grammaire de la phrase française*. Hachette Supйrieur.

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication*, *30*(3), 358-392. https://doi.org/10.1177/0741088313491692

Limpo, T., & Alves, R. A. (2017). Written language bursts mediate the relationship between transcription skills and writing performance. *Written Communication*, *34*(3), 306-332. https://doi.org/10.1177/0741088317714234

Mauranen, A. (2016). Temporality in speech–linear unit grammar. *English Text Construction*, *9*(1), 77-98. https://doi.org/10.1075/etc.9.1.05mau

Ulasik, M.A., Mahlow,C., & Piotrowski, M. (2025). Sentence-centric modeling of the writing process. *Journal of Writing Research*, *16*(3), 497-532. https://doi.org/10.17239/jowr-2025.16.03.05

Matsuhashi, A. (1981). Pausing and planning: The tempo of written discourse production. *Research in the Teaching of English*, 113-134. https://www.jstor.org/stable/40170920

Medimorec, S., & Risko, E. F. (2017). Pauses in written composition: On the importance of where writers pause. *Reading and Writing*, *30*, 1267-1285. https://doi.org/10.1007/s11145-017-9723-7

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81. https://psycnet.apa.org/doi/10.1037/h0043158

Mutta, M., & Salminen, S. (2021). Les sйquences prйfabriquйes dans la production йcrite dans le cas de scripteurs finnophones de franзais et suйdois L2 [Prefab sequences in written production for finnophone typers in French and L2 Swedish]. *Synergies pays riverains de la Baltique*, *14*(20), 11-26.

Olive, T. & Cislaru, G. (2015). Linguistic forms at the process-product interface: Analysing the linguistic content of bursts of production. In G. Cislaru (Ed.), *Writing(s) at the Crossroads: The process-product interface* (pp. 99-123). John Benhamins Publishing Company. https://doi.org/10.1075/z.194.06oli

Olive, T. & Bouriga, S. (2022, June 20-22). *Effects of cognitive demands of planning on bursts when writing with a pen or with a computer* [Conference presentation]. SIG Writing Conference 2022, Umee University, Sweden.

Poletto, C., & Tortora, C. (2016). Subject clitics. *The Oxford guide to the Romance languages* (pp. 772-785). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199677108.003.0047

Rшnneberg, V., Torrance, M., Uppstad, P. H., & Johansson, C. (2022). The process-disruption hypothesis: how spelling and typing skill affects written composition process and product. *Psychological Research*, *86*(7), 2239-2255. https://doi.org/10.1007/s00426-021-01625-z

Schilperoord, J. (1996). *It's about time: Temporal aspects of cognitive processes in text production*. Rodopi.

Sinclair, J. M., & Mauranen, A. (2006). *Linear unit grammar*. John Benjamins Publishing Company. https://doi.org/10.1075/scl.25

Stark, E. (2013). Clitic subjects in French text messages. In Jeppesen Kragh, K., & Lindschouw, J. (Eds.), *Deixis and pronouns in Romance languages* (pp. 147-169). John Benjamins Publishing Company. https://doi.org/10.1075/slcs.136.09sta

Strumqvist, S., Holmqvist, K., Johansson, V., Karlsson, H., & Wengelin, E. (2006). What keystroke logging can reveal about writing. In Sullivan, K. P. H., & Lindgren, E. (Eds.), *Computer key-stroke logging and writing* (pp. 45-71). Brill. https://doi.org/10.1163/9780080460932_005

Vasylets, O., & Marin, J. (2025) Linguistic and behavioral alignment in writing: A scoping review. Journal of Writing Research, 16(3), 375-404. https://doi.org/10.17239/jowr-2025.16.03.02

Wengelin, E. (2006). Examining Pauses in Writing: Theory, Methods and Empirical Data. In Sullivan, K. P. H., & Lindgren, E. (Eds.), *Computer key-stroke logging and writing: Methods and applications* (pp. 107–130). Elsevier. https://doi.org/10.1163/9780080460932_008