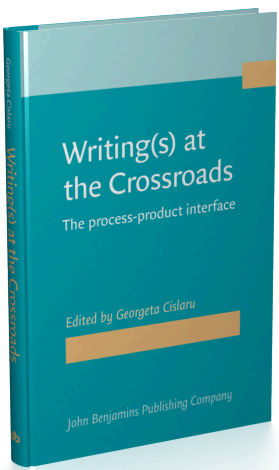
Linguistic forms at the process-product interface

Analysing the linguistic content of bursts of production

Thierry Olive | Centre National de la Recherche Scientifique & Université de Poitiers, France

Georgeta Cislaru | CLESTHIA, Université Sorbonne nouvelle Paris 3

https://doi.org/10.1075/z.194.06oli



Pages 99–124 of
Writing(s) at the Crossroads: The process–product interface
Edited by Georgeta Cislaru
2015. vi, 304 pp.

Linguistic forms at the process-product interface

Analysing the linguistic content of bursts of production

Thierry Olive & Georgeta Cislaru
Centre National de la Recherche Scientifique & Université de Poitiers, France / CLESTHIA, Université Sorbonne nouvelle Paris 3

In the present study, we adopt a cognitive-discursive approach to analyse the linguistic structures of bursts of production in a corpus of reports by social workers about children at-risk. Bursts were identified as periods of fluent writing between pauses of at least two seconds, and were coupled with textometric analyses of the final texts. We focused on repeated segments (RS) of texts, i.e. sequences of at least two linguistic units that are repeated at least twice in the corpus. Preliminary analyses showed that the number of bursts with identical or nearly identical content to repeated segments in the texts was limited. Morphosyntactic and semantic descriptions of RSs and bursts indicated that short and medium-sized bursts mainly corresponded to complete syntactical constituents, whereas short and medium RSs often correspond to incomplete syntactical constituents. All together, this study offers information on the structure of the language that is produced during bursts, and thereby raises further questions about the status of routines both at the discursive and psycholinguistic levels.

Keywords: bursts of writing; repeated segments; routines

1. Introduction: Linguistic forms at the process-product interface

This chapter aims to examine the nature and behaviour of linguistic forms on both sides of the process-product interface: the writing process in a real-life situation vs. text and its specific features. The comparison is structured in relation to a more general question concerning writing practices and the relationship between textual data and data related to the writing process. To what extent are text data relevant to the understanding of writing practices? And conversely, are writing

DOI 10.1075/z.194.06oli

practices predictive of the configuration of the text as a finished product of the writing process? The aim is to add a bit more linguistics and linguistic analysis to the description of the process, and to connect the text more directly to the conditions of its production.

Our study falls within the principles of a pluridisciplinary corpus analysis. The questions addressed in this chapter are anchored to four domains: linguistics, textual genetics, natural language processing (NLP), and psycholinguistics. Here we combine data from all four of these fields to assess how concepts from each of them may be put together to better understand writing practices, but also to produce a new research heuristic.

We collected a corpus of reports on children at risk written by social workers. The corpus was annotated with real-time data recorded by a keylogging program (Inputlog; see Leijten, Van Waes & Van Horenbeeck this volume) while the social workers were typing the reports. In order to detect specific or recurrent linguistic structures, we performed a twofold analysis. First, we analysed the content of the bursts of writing. The term "bursts" of writing refers to strings of text that are produced without major interruption. In other words, bursts are segments of text that are produced between two consecutive pauses. Second, we analysed the content of repeated segments, which are linguistic strings that are reiterated within a text or a corpus.

Our goal was to determine whether bursts and repeated segments are similar by comparing their content. If they are, then bursts and repeated segments, which are behavioural and textual observations respectively, should be considered to reflect similar phenomena in both writing and text. However, if the two do not coincide, they should be considered to reflect distinct writing phenomena. Thus, we compared the linguistic composition of bursts with re-occurring segments of discourse and, more specifically, with the linguistic forms present in repeated segments.

The first Section (§ 2) of the paper presents the notions of burst and repeated segment as well as a brief state of the art of the theoretical questions to which they are related: respectively, writing skills and discourse routines. These are followed by a Section (§ 3) which clarifies the methodological framework and the nature of the corpus. Finally, we present (§ 4) and discuss (§ 5) the results of the analysis, which are both quantitative and qualitative.

2. Bursts of writing and repeated segments of text

2.1 Bursts of writing

At the behavioural level, the activity of a writer can be described as a sequence of periods of handwriting (or typing) – i.e. bursts of production – separated by

pauses. Pauses usually take circa 50% of composing time, and they generally occur for cognitive reasons, although they can also result from socio-psychological or physical causes (see Schilperoord 2002). Cognitively speaking, pauses signal the occurrence of writing processes that cannot be carried out simultaneously with handwriting/typing; they may also be the consequence of memory decay, the writer having forgotten what s/he wanted to write. In the latter case, pauses are used to re-instate the intended message. In sum, pauses "are fundamental moments of conceptualization, formulation or control of the message" (Chanquoy, Foulin & Fayol 1996, 37).

By contrast, bursts of production are moments during which writers produce text as such, making up the remaining 50% of writing activity. Bursts of production are thus periods of handwriting (or typing) during which a segment of text is written. It is important to note that, during bursts of language, writers are not only transcribing what has been prepared earlier in the writing process. Instead, the writing cognitive processes of planning, translating and revising[1] can be implemented while handwriting or typing, at least when these latter skills are sufficiently automatized (Olive 2014). For example, Olive and Kellogg (2002) showed that adult writers, but not 9-year-old children or adult writers using an unfamiliar calligraphy, can simultaneously apply planning, translating and revising during handwriting (for similar evidence with writers' eye movements, see Alamargot, Dansac, Chesnet & Fayol 2007). In fact, in adults, translating occurs mostly during handwriting, whereas planning and revision mainly occur during pauses (Alves, Castro & Olive 2008; Olive, Alves & Castro 2009).

Kaufer, Hayes, and Flower (1986) conducted the first study that investigated bursts of production. They showed that adult writers typically compose by producing segments of text with an average length of 9 words. They also observed that more skilled writers composed using larger bursts (four words more on average) than less skilled writers. Since the texts written by the experts were generally rated of better quality than those composed by novices, the authors interpreted this increase in burst size and length as evidence of more efficient translating processes. This interpretation was later confirmed by Chenoweth and Hayes (2001), who found that undergraduate students are more fluent and produce longer bursts when composing in their first language than in their second language (L2); the same finding was observed with students who were more skilled in L2 in comparison to less skilled

1. "Planning" refers to psychological processes that operate at a conceptual level for retrieving and organizing ideas. "Translating" refers to the psycholinguistic processes that formulate written language (see also Galbraith and Baaijen this volume). Revision processes are engaged when reviewing the text and assessing its match with the writers' communicative goals (see Brunner and Pordeus Ribeiro this volume; Fenoglio this volume). A more detailed account of these writing processes can be found in Alamargot and Chanquoy (2001).

students. Chenoweth and Hayes (2003), as well as Hayes and Chenoweth (2006), completed these findings by showing that impairment in verbal working memory, a system that is required in translating, consistently decreased burst length and writing fluency. Notably, Hayes and Chenoweth (2006) did not find bursts in the production of expert typists, as if these writers were able to prepare their text completely while typing, and thus to compose without pausing. Using a passive-to-active sentence conversion task, Hayes and Chenoweth (2007) also concluded that translating is strongly involved in bursts of execution.

Bursts are also determined by writers' handwriting/typing skills. Several studies have shown that having a low level of handwriting or typing skills directly constrains writing fluency and text quality. The detrimental effect on text quality is due to the resulting need for a large amount of cognitive resources or attention when handwriting, which cannot be devoted to planning, translating or revising (Olive 2014). Accordingly, writers with limited handwriting or typing skills do not have enough processing capacity to activate high-level writing processes in parallel. They therefore produce their text in short bursts, during which they mainly produce the text prepared during a previous pause: i.e. with a *thinking-and-then-writing* strategy. By contrast, because writers with high level of transcription skills need little if any cognitive resources to produce the text itself, they can activate high-level writing processes as they do so, with a *thinking-while-writing* strategy (Olive 2014). Thus, they are better able to produce longer bursts.

Consequently, automatizing transcription also leads to longer language bursts. For example, writers with a high level of typing skills compose in larger bursts, on average three words more (Alves, Castro, Sousa & Strömqvist 2007). Similarly, fourth-graders with a high level of handwriting skill show larger written language bursts, compose text more fluently, and produce better stories (Alves, Branco, Castro & Olive 2011). More recently, in Alves, Olive, and Castro's (2008) study, half of the participants composed by handwriting and the other half by typing. In the handwriting group, handwriting skill was manipulated by asking writers to use either an uppercase cursive script or their usual calligraphy. In the typing group, typing skill was manipulated by using either a normal or a scrambled keyboard layout. In both modalities, the low-skill groups showed similar reliable decreases in burst length – about six words less – and received lower ratings for text quality.

As this short review shows, at least translating and handwriting determine the duration and length of bursts. Writers with a high level of handwriting skills can devote their available cognitive resources to translating, which can then be maintained longer while handwriting. Moreover, a high level of translating skills

allows writers to prepare longer segments of texts concurrently to handwriting. In sum, the greater a writer's level of translating and handwriting skills, the longer the bursts they produce.

Less is known, however, about the content of bursts, and more specifically about the linguistic structure of the portions of text that are produced during bursts. In fact, only one study has analysed the linguistic structures of bursts (Kaufer et al. 1986). The authors showed that these parts or segments tend to correspond to clauses, since they showed a strong tendency to end at clause boundaries, and less so at phrase boundaries. Thus, according to these authors, writers compose sentences by first selecting a topic, and then by producing and evaluating sentence parts that fit grammatically with the part of the sentence that has already been prepared. If the evaluation is negative, the writer has to either revise the current part or produce an alternative part to follow it. If the evaluation is positive, then the sentence part is added to the current sentence that has already been produced, or that is still in the writer's mind (i.e. in verbal short-term memory).

In this context, we explored the process-product interface by investigating for a possible association between bursts of production and the linguistic forms or structures that are produced during these execution periods. In particular, we aimed to determine whether texts are produced in segments of text that share common structural characteristics. The main questions about bursts that we addressed in this study were the following:

- What are the linguistic forms produced in bursts?
- Are there regularities in the content of bursts?
- Is bursts' content predetermined by specific, defined text structure/patterns?
- Are there ready-made linguistic structures that can be retrieved from long-term memory and directly written out?

2.2 Repeated segments

Following Lafon and Salem (1983) and Salem (1986), we define "co-occurrences" as "couples of forms that function almost exclusively within idioms"[2] (Lafon & Salem 1983, 162). This definition selects one specific subtype of co-occurrence, namely, repeated segments (RSs): i.e. strings of at least two graphical units that occur together at least twice in a text or a corpus. RSs represent ready-to-speak

2. "[…] des couples de formes fonctionnant presqu'exclusivement à l'intérieur d'expressions figées" (Lafon & Salem 1983, 162).

units (which are somewhat different from collocations): in the framework of textometry and discourse analysis, they are considered as discourse routines that characterize either a studied language or a type of discourse. In the framework of corpus linguistics, Sinclair (1991, 2004) showed that all linguistic productions, oral or written, are half constituted of prefabricated sequences, following an "idiomaticity principle" (cf. Erman & Warren 2000; Kuiper 2009). Biber's corpus-driven studies on multi-word regular sequences (Biber 2009; Biber et al. 2004) also show a high prevalence of various types of formulaic language in both oral and written corpora.

The repetition principle suggests the hypothesis of a routinization of discourse, as defined by Wray (2002, 9): sequences of words or other units that seem to be prefabricated, which are memorized and reproduced "as is" in the text, and not generated ad hoc. Along the same lines, Mayaffre (2007, 10) wrote "Repeated segments of significant length are linguistic tunnels where the creativity of the speaker/writer is reduced in favour of a kind of recitation."[3] These are strong hypotheses, which we will test here by comparing the bursts and repeated segments in our corpus. Accordingly, repeated segments are viewed as key elements of text organization in the framework of discourse analysis and corpus linguistics,[4] inasmuch as they signal discourse routines related to genre, social sphere of activity, professional domain and occupation, etc.

RSs constitute a "formal" approach to linguistic routines, inasmuch as they are detected by their graphical form. There are several ways to broaden the insights that can be drawn from the study of such units. On the one hand, generalized "grammatical patterns" may be identified as regularity in some subset of the repeated segments (Hunston & Francis 2000). In this we follow Biber's (2009) and Biber et al.'s (2004) work on lexical bundles. On the other hand, the semantic types

---

3. "Les segments répétés de longueur importante sont des tunnels linguistiques dans lesquels la créativité du locuteur recule au profit d'une forme de récitation." (Mayaffre 2007, 10)

4. The study of co-occurrences is "[…] the first thing to do in order to underline semantic nets that are shaped in a text or, more precisely, which *shape the text*; the first thing to do in order to reach the essential features of the text (i.e.: what makes the text a meaningful linguistic sequence (a 'completeness of meaning' [Détrie, Siblot, Verine 2001, 349]) and a coherent and cohesive assembly of words" (Mayaffre 2007, 8).

"[…] premier mouvement pour pointer les réseaux sémantiques qui se forgent dans un texte, ou plus précisément *qui forgent un texte*; le premier mouvement pour toucher à l'essentiel de ce qu'est la textualité (*i.e.*: ce qui fait d'un texte une suite linguistique signifiante ('une complétude de sens' [Détrie, Siblot, Verine 2001, 349]) et un assemblage de mots à la fois cohérent et cohésif)." [our translation]

of repeated segments may be considered, depending on their lexical-grammatical contents and discursive profile (see also Cislaru et al. 2013), such as:

- The "waffle" (doublespeak) RSs, determined by the genre or the topic of discourse (*être en/be in, can+speech verb*).
- RS-genre clichés, related to a type of cognitive activity: analysis, evaluation (*nous avons/we have, nous pensons/we think*).
- RSs representing structural clichés in French (*de la, lieu de, part de, une fois, quant à, en effet*).
- RSs representing individual discourse habits (*ce dernier/cette dernière – the latter,* etc.).

We are concerned here with the particularities of a type of professional discourse, the reports of social workers on at-risk children. The early presence of the RS in the drafts might signal a stereotyped form of discourse, reflecting a strongly constrained professional discourse. At first glance, the drafts of social workers' reports in our corpus do not really seem to correspond to such a discourse type. For instance, the longest RS recorded by the machine contains 11 forms. Whereas a few of them emerge beginning in the first two versions of a report, most of these RSs appear no earlier than version 4–6. This deserves to be underlined, inasmuch as it suggests that ready-mades are not automatically activated in the first stages of the writing process (and they are probably not genuine ready-mades).

2.3 Bursts versus repeated segments

To summarize, by contrasting the defining criteria and interpretive hypotheses regarding bursts and repeated segments, we highlight the differences between the two categories, and show that some of the criteria of definition and identification are undecidable, inasmuch as they need to be tested on corpora, and, concerning bursts more particularly, because they have not yet been submitted to detailed linguistic analysis.

As can be seen in Table 1, bursts and repeated segments are not obviously aspects of the same psycho-linguistic interface. A handful of existing studies run in the same direction. For instance, Schmitt et al. (2004) selected recurrent target clusters from corpora and then tested them during psycholinguistic tasks involving native speakers and second language speakers (see also Schmitt 2004). Their results suggest "it is unwise to take recurrence of the clusters in a corpus as evidence that those clusters are also stored as formulaic sequences in the mind" (Schmitt et al. 2004, 147). Schmid also underlines the weak interpretive impact of the notions of recurrence or frequency, and writes: "we seem

to be quite far from having a good grip on the relation between frequency and entrenchment.[5] This is mainly due to the unclear interaction between absolute and relative frequency, or cotext-free and cotextual entrenchment, respectively." (Schmid 2010, 123).

As suggested by Schmid (2010, 102), "patterns of frequency distributions of lexico-grammatical variants of linguistic units correspond to variable degrees of entrenchment of cognitive processes or representations associated with them". Accordingly, because RSs may constitute prefabricated written forms or discourse routines, from a psycholinguistic point of view, they may refer to language forms that are retrieved in a single block from the writer's long-term memory, and that therefore can be written in a single burst. Consequently, if RSs are discourse routines, then the linguistic form of bursts and of RS may be expected to be relatively similar.

Table 1. Contrasting the features of bursts and repeated segments

Criteria	Bursts	Repeated segments
Recognition	Pauses, real-time data	Repetition and identity
Frequency	Does not apply	Defining
Newly created	Yes (No)	No
Memorization	Does not apply	Presupposed
Familiarity	Not expected	Yes, presupposed
Conventional meaning	?	Yes
Context dependence	?	Yes
Competences required	Writing	Discourse
Standard methodology	Real-time analysis	Corpus-driven

3. Corpus and methodology

3.1 Global description of the corpus and of the method of analysis

Our study is based on a corpus of six reports written by social workers (see Table 2) with a key-stroke logging program (Inputlog; see Leijten & van Waes 2006; Leijten,

5. (Langacker 1987; see Schmid in press for a discussion). "Entrenchment refers to the ongoing reorganization and adaptation of individual communicative knowledge, which is subject to exposure to language and language use and to the exigencies of domain-general cognitive processes and of the social environment." (Schmid in press).

Van Waes and Van Horenbeeck this volume) during their regular activity of monitoring and evaluation of the situation of foster children (see also Brunner and Pordeus Ribeiro this volume, and Cislaru and Lefeuvre this volume, for a detailed presentation of the writing situation). Inputlog is a computer tool that records all actions that a writer performs with the computer when composing a text using a word processor and keyboard. In the present case, each key press, along with its timing, was recorded, as was each move of the mouse in the text or in the menu of the word processor that the writers were using. All textual operations to modify the text were also recorded, as were all interactions with other programs (web browser, email client, etc.) available on the computer.

Table 2. Corpus description

Reports	6 (from 2 to 6 pages)	Sentences	979
Pages	30	Words	13,701
Paragraphs	305	Words per sentence	14

The linguistic analysis, supported by the application of NLP tools to a corpus annotated with real-time data, is quite new in the field, and requires several methodological adjustments. To perform these, we used a natural language processing program developed by Adrien Lardilleux (Lardilleux et al. 2013) to extract bursts from Inputlog's log files, automatic detection of repeated segments (initially extracted by Le Trameur, a textometry tool),[6] and the alignment of repeated segments and bursts, as shown in Figure 1. The right frame gives counts of repeated segments (columns to the left of text) and bursts (columns to the right of text) recorded in a complete file of drafts for one report. The outermost numbers indicate the total number of units that are represented as by a given set of related repeated segments (far left of the frame) and bursts (far right of the frame). In the left frame, the upper part shows context for the selected repeated segments, while the bottom part shows the neighbourhood of the selected bursts, with temporal data (the first column indicates the timing within the time log for the writing session, and the second column the length of the burst in seconds).

6. Developed by Serge Fleury, Université Sorbonne nouvelle Paris 3, http://www.tal.univ-paris3.fr/trameur/ (see Née et al. 2012 for an application on this corpus; see also Doquet and Poudat this volume).

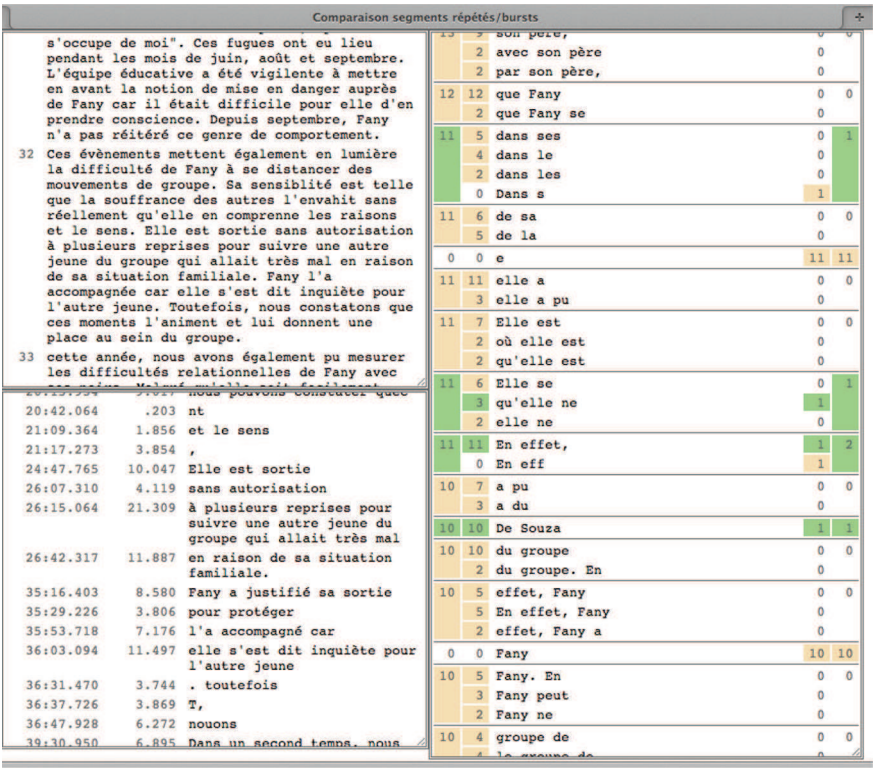


Figure 1. Alignment of bursts and repeated segments, with direct access to the text

Our study is based on hybrid methods of corpus analysis (close to pattern grammar studies: cf. Biber 2009) based on real-situation text production. The list of repeated segments is corpus-driven, and we operate with raw data, without sequence pre-selection. The bursts were produced during a real-time and real-situation writing activity. Both bursts and repeated segments were also described grammatically (in the terms of constituent analysis by using pre-defined grammatical categories) and semantically. Particular attention was paid to cases of homonymy, such as *qu'elle ne* (~ that she not), which introduces a noun determiner as a burst (*le sentiment qu'elle ne* – the feeling that she Verb not…) versus reported speech as a repeated segment (*elle dit qu'elle ne* – she says that she Verb not…). The fact that the same writers produced both bursts and repeated segments is crucial, ensuring that the data for both correspond to the same discourse genre and social activity. Indeed, the homogenous conditions of text production for bursts and repeated segments reinforce the conclusions that can be drawn on the (non-)correlation between the two.

The threshold of frequency for repeated segments was fixed at 2, given the size of the corpus. For the real-time analysis, we opted for pauses between bursts of more than 2 seconds. Bursts were therefore defined as periods of typing separated by pauses longer than 2 seconds. This threshold allowed us to exclude all pauses that resulted from typing movements from the analyses (for example, moving the hands and fingers on the keyboard to reach the next key, or preparing a combination of keys to type a diacritic character), and whose origin thus did not lie in the operations of one of the writing processes.[7]

Table 3 sums up the quantitative characteristics of the studied corpus. The reports were written in 34 sessions of an average of 23 minutes in length, corresponding to a total composing time of 12h57. Within this time, roughly 40% was spent pausing (5h10) and during the remaining 60% (7h47), the social workers typed their text. They produced their text at a speed of 17,6 words per minute, which falls within the normal range for common compositional fluency in adults. Bursts were long (22 sec.) and contained few words (2.6) suggesting that at least some of the writers were not highly skilled typists.

Table 3. Quantitative and temporal parameters associated to the reports

Writing sessions	34		
Total composition time	12h57	Number of analysed RSs	1506
Total pause time	5h10	Number of pauses	5157
Total writing time	7h47	Mean pause length	13 s
Mean session length	23 min.	Mean pre-writing pause	8 s
Writing fluency	17,6 wpm	Mean within-words pause	10 s
Number of analysed bursts	1014	Mean between words pause	11 s
Mean burst duration	22 s	Mean between sentences pause	13 s
Mean burst length	2.6 words	Mean between paragraphs pause	20 s

3.2 Text progression

Text progression is rather linear in contrast to the evolution of drafts, and many authors who work with the text as a finished product consider that "texts are linearly… and also non-linearly developed" (Hoey 2004, 395). However, a detailed study on the rewriting operations involved in each draft for a larger series of social reports (twenty-nine reports in total: see Brunner and Pordeus Ribeiro

7. For a recent discussion on the several ways that this threshold has been defined, see Chenu, Pellegrino, Jisa, and Fayol (2014).

this volume) shows that text progression "in process" is not necessarily linear (see also Fenoglio this volume), and that chunks of text are frequently displaced more than once within the text.

The text progression of the reports we analysed is shown in Figure 2. As can be seen, it took the writers between 1 and 11 sessions to compose their reports, suggesting the adoption of very different writing strategies. It may also be noticed that revision sessions can be detected, i.e. when the curve becomes flat, particularly at the end of writing sessions for the reports that were written over a larger number of sessions. By contrast, some reports increased greatly in length between sessions. In sum, social workers differed in their way of completing their reports.

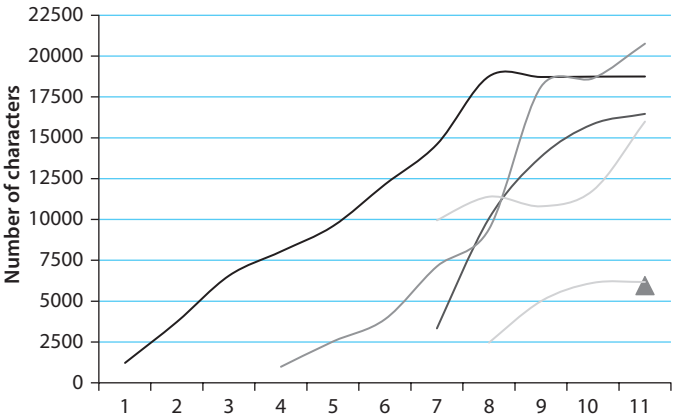


Figure 2. Text progression. The x-axis represents the writing sessions, and the y-axis the number of characters produced. For example, one report was written in 11 sessions, while another was written in a single session

3.3 Pause analysis

Before analysing the content of the bursts and the repeated segments, we first looked at pause data to assess whether the writers who composed the reports could have been differently drawing on the cognitive processes involved in writing. As a first observation, it is interesting to notice that globally, the prewriting pause is the shortest one, even shorter than the mean within-words pause (see Table 3). This may indicate that the writers had already a plan in mind when they began composing their reports. This is not so surprising since the reports include a set of predefined sections on specific topics (life history, daily life in the group, health, school, relations with others, conclusion), which helped the writers structure their reports. This resulted in reduced planning efforts.

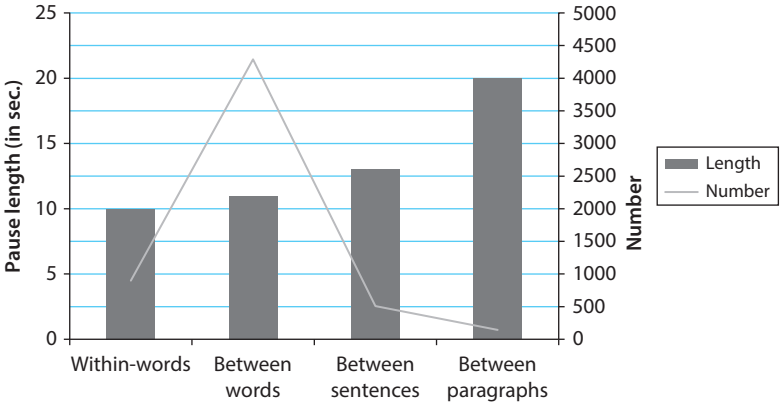


Figure 3. Number and length of pauses longer than 2 seconds

Replicating previous findings on pauses during writing, we observed that the location of pauses in the text strongly influenced their length (Foulin 1995, 1998; Schilperoord 2002; van Hell, Verhoeven & van Beijsterveldt 2008). The shortest pauses were within words, followed by pauses between words, and then pauses between sentences, while the longest pauses were those that preceded a paragraph (see Figure 3). As suggested above, since pause length may be taken as an index of the mental effort that the writer is exerting in constructing the text, our data indicate that pauses before paragraphs involved more mental effort by the writers. This is presumably due to the fact that before writing a paragraph, writers engage planning processes that are cognitively more costly than the formulating processes activated before producing sentences and words (see Olive 2004, 2012 for a review on the cognitive demands of writing processes). Despite the low planning demands of this writing situation, it is nevertheless possible to conclude that the writers studied here engaged cognitive writing processes in a rather standard way.

4. Linguistic analysis

Less than 3% of bursts and repeated segments converged, i.e. were 75 % similar from a formal/graphic point of view. Given that this ratio is very low, we searched for various comparison criteria, in order to establish a more complex linguistic view on the phenomena that the two represent. Our first choice was to contrast syntactic structures that were specific to bursts and repeated segments. This approach is based on corpus linguistic methods, and more specifically on the colligation principle (Hoey 2005; Yamasaki 2008), which allows the identification of pattern types via constituent analysis.

The following is a sample of the most common constructions:

- Noun phrase (NP)
- Prepositional phrase (PP)
- Noun
- Noun phrase + Verb (NP+V, close to sentence-type)
- Verb (auxiliary, participle or other incomplete verb form)
- Verb phrase (VP)
- Adjective
- Adverb
- Adverbial phrase, Connector, Connector & NP or Connector & NP+VP
- Conjunction
- Clause
- Determiner
- NP + preposition
- Preposition + determiner
- Etc.

The classification of lexical strings – either RSs or bursts – demands complex criteria and a number of adjustments to the types of syntactic structures to which they are assigned. The main criterion applied was syntactic saturation, due to its analytical accessibility. This means that we mainly distinguished two categories: saturated strings, which correspond to phrase-type constructions (noun phrases, prepositional phrases, sentences, etc.), and unsaturated strings, which correspond to syntactically irrelevant constructions and units that associate two grammatical groups (phrases), such as *NP+preposition*, or that stop ahead of the boundary of a grammatical group, such as *Preposition+determiner*. However, we are aware that saturation is a notion that is subject to further negotiation. First, syntactic saturation does not always coincide with semantic (i.e. informational) saturation. Some noun phrases, for instance, may be saturated out of context, but unsaturated in discourse use/context; thus, semantically, "*her/his difficulties*" may be either saturated or unsaturated (e.g. "*her difficulties in…*"). We tried to take such cases into account, but a much more thorough semantic analysis is needed. Second, lexical strings are never discursively and interdiscursively saturated: they always maintain and evoke connexions with other words or lexical strings, and this might be an important cognitive/memory factor. Another difficulty is related to the ambiguous status of connectors, which represent discursive functions rather than grammatical categories. Unfortunately, a detailed discourse analysis was not really manageable within the framework of this study. These two main difficulties imposed a limitation on the syntactic criterion. Nevertheless, we took into account a range of semantic data, and thus subdivided the

saturated and unsaturated categories into several subcategories. Full stops and capital letters were treated as graphical criteria, marking sentence boundaries. These criteria allowed us to identify different types of breaks, such as …*other children. She*….

Non-saturated constructions are markers of discontinuity, and can also include:

- Break before and after full stop: … *other children. She*…
- Break after coordination: *Alex shows some signs of sadness and/but* [he]…
- Break after concatenation between a saturated unit and a connector: *She decided to leave. Therefore*…
- Etc.

Some strings, like clauses, may be regarded as saturated although they are not autonomous. Things get more complicated with longer strings that are peculiar to bursts, which can contain saturated sentence-type strings followed by unsaturated strings, for instance.

It is interesting to study the distribution of these constructions in repeated segments and bursts, in order to verify the linguistic particularities of the two categories, by distinguishing: (i) the distribution of each type of construction; (ii) the distribution of saturated vs. non-saturated constructions.

- ***RSs and bursts for each linguistic structure***. Table 4 presents the percentage of saturated and unsaturated linguistic structures in RSs and bursts. Overall, the number of saturated and unsaturated structures in the corpus significantly differed ($\chi^2 = 203.6$, $p < .001$). Accordingly, unsaturated bursts and RSs are more numerous than saturated ones. The distribution of bursts and RSs between these two types of structures also significantly differed ($\chi^2 = 42.9$, $p < .001$). More precisely, around 42.6% of bursts are saturated, whereas 32% of RSs are saturated. This difference between bursts and RSs also indicates that the grammatical structures found in repeated segments of text and in bursts differed.

Table 4. Number and percentage (in parentheses) of saturated and unsaturated linguistic structures among all bursts and repeated segments

	RS	**Burst**
Saturated	483 (32%)	1083 (43%)
Unsaturated	1016 (68%)	1458 (57%)
Total	1499	2541

– ***RSs and bursts with saturated structures***. Sentences and clauses are specific to burst production in our corpus; no saturated pattern of this category is attested as a repeated segment. Although numbers may correspond to a single graphical unit (e.g. *2009*) and could have been produced in a single burst, they could not be classified as repeated segments, which required a string of at least two units. It may be noted that saturated noun phrases and prepositional phrases are the structures most frequently found in both repeated segments and bursts, and they are the most frequent ones. Connectors and adverbials are also common to burst and RSs but are less frequent. Additionally, these three types of structures were more present in bursts than in RSs. Verb phrases, NP + conj + NP, clauses, sentences and linked sentences constructions are not frequently used in RSs. Of these, only verb phrases, clauses and sentences appeared in bursts, albeit to a lesser extent than the types shared with RSs.

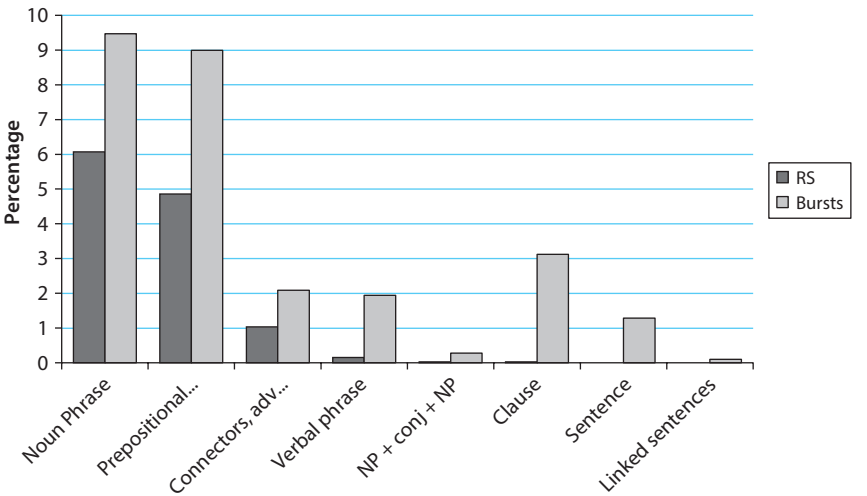


Figure 4. Percentages of bursts and repeated segments represented by different types of saturated linguistic structures

– ***RSs and bursts with discontinuities (unsaturated)***. Unsaturated constructions appeared in both bursts and RSs. Long items and breaks or double breaks in the context of a boundary marker (full stop, connector, conjunction, comma, etc.; see above) were clearly exclusive to bursts. Repeated segments had a greater association than bursts with only two types of discontinuities: incomplete noun phrases and prepositions followed by determiners or various other items.

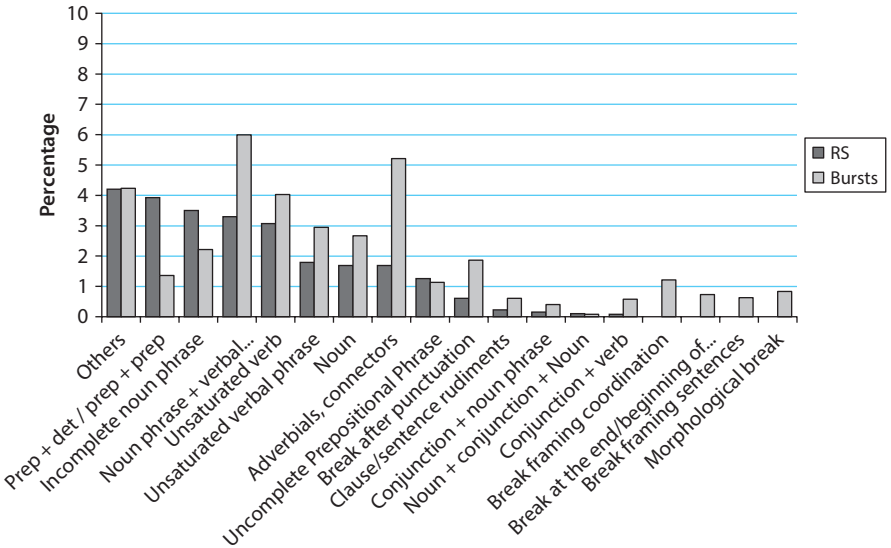


Figure 5. Percentages of bursts and repeated segments in the various unsaturated linguistic structures

5. Discussion

5.1 Saturated and unsaturated patterns

The concept of unsaturated constructions is similar to that of lexical bundles, as defined by Biber (2009). First, lexical bundles are by definition extremely common (in contrast to most idioms and many 'grammar patterns', which tend to be rare). Second, most lexical bundles are not idiomatic in meaning and not perceptually salient. For example, the meanings of bundles like *do you want to* or *I don't know what* are transparent from the individual words. And finally, lexical bundles usually do not represent a complete structural unit (Biber 2009, 283).

Talking about the idiom-collocation principle, Partington (1998, 19 et sq.) suggests that the use of prefabs facilitates communication processing on the part of the speaker as well as the hearer. But what seems to be obvious – at least at first glance – for oral communication does not function with written communication, where process and product are clearly separated both materially and chronologically. According to Biber (2009), lexical bundles in writing, such as the construction *in the light of,* usually serve to bridge pairs of phrases, and are open-choice oriented on their right border. Indeed, some of the unsaturated constructions in our corpus are open-choice repeated segments. For instance, adverbials followed

by prepositions or subordination markers, as well as prepositions followed by determiners or various other items, are "filled with" relevant units in context; in Biber's terms, they provide a kind of "pragmatic head" for larger units, thus assuming the role of "interpretive frames".[8]

From a cognitive point of view, bursts may be considered to function in a similar way, with the writer having to pause in order to choose the contextually relevant development. However, this interpretation is nuanced by the very limited number of units that bursts and repeated segments share. The affirmation of this similarity thus appears to be cognitively valid and semantically weak, as suggested in Section 5.2. Discontinuity constructions highlight cases where the connection between facts is pre-constructed, and only the discourse elements that are to be connected are selected from a list of possibilities.

Cognitive linguistics (see Schmid 2010 and in press for discussion) is interested in the degree of routinization and automation in the formation and use of a unit. The hypothesis that repetition favours entrenchment is close to the assumptions of corpus linguistics (see above). Yet Schmid (2010, 125) sums up his paper as follows: "What I have tried to show here, however, is that so far we have understood neither the nature of frequency itself nor its relation to entrenchment, let alone come up with a convincing way of capturing either one of them or the relation between them in quantitative terms."

The results of the present study can be discussed first in terms of the communicative competence/performance contrast, along the lines of Hymes' (1971) proposals. Hymes distinguished four kinds of competence skills: knowledge of what is formally possible given the constraints of the language system, on feasibility, on appropriateness and, last but not least, on actually produced sequences. Our study focuses on actual performance, and offers the possibility to confront performance with competence hypotheses. It may be assumed, in the light of the results presented here, that the produced sequences can be separated into two distinct classes, those of process-performance and product-performance. The fact that we found a great number of repeated segments which did not have an equivalent burst may reflect strategies of communicative adaptability (Mey 1998, Verschueren & Brisard 2009), which fits particularly well with the fact that most of the repeated segments did not emerge in the first drafts of writing. This point may suggest an overlap between the use of linguistic prefabs and the shaping of text to conform to social norms. Finally, as noted by Schmid,

---

8. See Biber et al. (2004) for the three primary discourse functions for lexical bundles in English: (1) stance expressions, (2) discourse organizers, and (3) referential expressions.

> […] what frequency counts in a corpora reflect more or less directly are degrees of *conventionalization* of linguistic units or structures. Conventionalization, however, is a process taking place first and foremost in social, rather than cognitive, systems, and it requires an additional logical step to assume that degrees of conventionalization more or less directly translate into degrees of entrenchment. (Schmid 2010, 116–117)

The writer may thus search for the appropriate formulae, first describing the individual situation of the youth they are monitoring, and then adapting the particular situation to social norms, in terms of both assessment and language choices. This hypothesis can be verified by tracking the contents of revision bursts (Baaijen, Galbraith & de Glopper 2012; see also, under a different perspective, Galbraith and Baaijen this volume).

5.2 Cognitive-semantic analysis and discussion

A complete semantic analysis of bursts and repeated segments would require an entire study unto itself. Space does not allow us to present such a full semantic description of our data here, but in what follows we will highlight a few relevant semantic phenomena.

First, common RS-burst constructions related to certain specific denotational domains. Noun phrases were frequent among these common constructions, although they also included N + V and adverbial constructions. Among the noun phrases, common RS-burst constructions were often Poss.det. + Noun. Most of these referred to the foster child's family:

ses parents, sa mère, sa famille, ses sœurs, ses inquiétudes pour, de ses sœurs, chez sa mère	*his/her parents, his/her mother, his/her family, his/her sisters, his/her worries for, of his/her sisters, at his/her mother's home*

Items focusing on the child were also frequent; they were often configured as Subject Noun + Auxiliary Verb constructions, although various patterns were attested. Some patterns seem to be gender-oriented, such as the preferential use of *to be* (potentially followed by a characterization) with a female subject, and the preferential use of *to have* (potentially followed by an event-verb) with a male subject:

Eloïse/Fanny/il est; elle peut; se montre plutôt; Kevin/Alexis a; il a demandé; né/e le; Enfin, Fanny; qu'elle doit; qu'elle ne; du jeune	*Eloïse/Fanny/he is; she can; tends to be rather; Kevin/Alexis has; he asked; born on; Finally, Fanny; that she must; that she not; of the youth*

Some constructions shared by bursts and RSs refer to stereotypical realia in social work, such as the collective (a) or an institution (b); characterize the child and support the social worker's evaluation (c); or refer the educator's subjective involvement (d):

(a) groupe de, sur le groupe (x2), du groupe, de l'internat, sur l'internat, aux autres (x2), les autres, des contacts	(a) *group of/from, on the group* (*x2*), *of the group, of the boarding residence; on the boarding residence; to others* (*x2*), *others,* (*of*) *contacts*
(b) les éducateurs, au SAFE, en IME, par le SESSAD, de l'Orangerie, le placement,	(b) *educators, at the SAFE, in IME, by the SESSAD, of the Orangerie, the fostering*
(c) se réfugie dans, la question, les difficultés, les raisons, reste difficile	(c) *to take refuge in, the question, the difficulties, the reasons, remains difficult*
(d) nous observons, nous constatons, nous avons, nous avions, nous lui avons	(d) *we observe, we notice, we have, we had, we had…to him/her*

Complex proper names and other designative phrases (*de Balleroy, Me Alleaume, Mme de Souza, Mme Chaudin, Mme X, Mr Y*, etc.) were also included among common RS-burst constructions: not only did names occur in all parts of the text, but they also seem to have been written within bursts.

All these denotative types are highly entrenched in the situation and professional practice of these writers, putting the child, his/her family and situation, and fostering at the core of the writing process. It is the institution, however, that determines and dominates the denotative domain.

Moreover, some idioms were found in the corpus both as bursts and as repeated sequences – either formulaic, like most of the adverbials below, or some type of verbal lexical bundles. These are the only systemic elements that were at the same time memorized and reproduced as bursts, and produced as repeated segments:

un peu, suite à, à chaque fois, du fait, à ce sujet, lors des, à plusieurs reprises, d'autre part, de ce fait, mais aussi, De plus, En effet (x2), Par ailleurs (x2), Pour autant (x2)	*a bit, following* (*the*), *each time,* (*because*) *of the fact, on this subject, during, several times, on the other hand, as a result/therefore, but also, moreover, indeed, furthermore, nevertheless*
a été (x2), a été prononcé, il faut, se trouve	*has been/was, has been/was pronounced, he must/it is necessary, ~as a matter of fact*

Secondly, the types of patterns that were specific to bursts present some intriguing profiles which merit semantic and cognitive analysis. The cases of discontinuity, which were more peculiar to bursts, are in this respect semantically relevant. For instance, the discontinuity after or on both sides of a full stop, a coordinative marker or a connector indicates that concatenations were anticipated, but

their content was not pre-formulated: ***et*** *son agressivité contenue; importantes.* ***.*** *Le médecin… –>* ***and*** *her contained aggressiveness; great. The physician…*

Many non-saturated noun bursts corresponded to lists or subtitles, and signal the existence of genre-specific pre-defined structures in the reports of social workers.

We also attested two infra-grammatical types of bursts, the plural marker "*s*" and the feminine marker "*e*" as breaking points, which suggest the occurrence of cognitive processing before these two grammatical markers. This is quite surprising, since a pause before morphological markers of gender or number suggests that the marker was not retrieved along with its noun but in a separate step. This suggests that in some cases calculating the morphological marker is effortful for writers, a behaviour that is more typical of novice writers.

6. Conclusion

Less than 3% of the units in the analysed texts were shared by bursts and repeated segments. This means that corpus data are not psycholinguistically valid here, and that the memorization and automatisation principles supposedly associated to clusters and collocations cannot objectively define the recurrent occurrences attested in the body of discourse that we examined.

However, some relevant features can be formulated here:

- The most frequent types of repeated segments of text and of bursts shared the same linguistic structures;
- The most frequent linguistic structures in bursts and RSs were syntactically unsaturated strings;
- The longest bursts were made up of complete syntactic structures such as sentences, and clauses, but these bursts were less frequent.

The "nominal" dimension of the discourse in our corpus is quite intriguing. The prevalence of saturated noun phrases and prepositional phrases (which usually contain a noun phrase) as repeated segments, and the high proportion of noun-based constructions among those that appeared both as repeated segments and as bursts, point to a topic that will be worth examining in depth in following studies.

It must be mentioned that the comparison performed here between RSs and bursts strongly depends on the definition of bursts. As mentioned above, bursts were identified in real-time data by the presence of a fluent transcription period, i.e. without any pause. The question of to define a threshold for identifying pauses is still a matter of debate in psycholinguistics (see Chenu, Pellegrino, Jisa & Fayol

2014). Our analysis of bursts is hence dependent on the two-second threshold used here. The notion of burst thus has to be fixed, and the ideal pause duration settled, before they can be assigned the status of psycholinguistic counterparts of formulaic language. Further research should examine the possible slot choices in open-ended productions, more thoroughly confronting the qualitative contents of bursts and repeated segments.

Finally, our study offers new information about linguistic entrenchment (Schmid, in press). Entrenchment is the psychological consolidation of linguistic structures which, as such, are retrieved from writers' or speakers' long term-memory in a single chunk and consequently expected to be executed fluently. We explored entrenchment by comparing process and product performances, and showed that, in our corpus, RSs and bursts share very few linguistic structures. The relationship between social conventionalization, on the one hand, and cognitive automatisation and retrieval, on the other hand, remains to be clarified. Our study shows that both play a role in both the product and the process of writing.

The prospects for generalization from our data are limited, inasmuch as we were dealing with professional writing, with domain-specific norms, instructions, and constraints (see Cislaru 2014). Nevertheless, it may be that the constructions seen here to occur both as repeated sequences and as bursts play a particular role in communication. In the case of our corpus, they may represent text segments that are conventionally shared by all the addressees of the reports.

Acknowledgments

Our thanks go to the French National Research Agency (ANR) for funding the ECRITURES project (2011–2014); to Adrien Lardilleux, post-doctoral fellow with the ANR ECRITURES project; to Michela Spacagno, Ph.D. student at Université Sorbonne nouvelle Paris 3; and to the members of the ANR ECRITURES Project generally.

References

Alamargot, Denis, and Lucille Chanquoy. 2001. *Through the models of writing*. Dordrecht (NL): Kluwer.

Alamargot, Denis, Christophe Dansac, David Chesnet, and Michel Fayol. 2007. "Parallel processing before and after pauses: A combined analysis of graphomotor and eye movements during procedural text production." In *Writing and Cognition: Research and Applications*, ed. by Mark Torrance, Luuk Van Waes, and David Galbraith, 13–29. Amsterdam: Elsevier. DOI: 10.1108/s1572-6304(2007)0000020003

Alamargot, Denis, Patrice Tellier, and Jean-Marie Cellier (Eds.). 2007. *Writing in the work place*. Amsterdam: Brill. DOI: 10.1163/9789004253254

Alves, Rui A., Marta Branco, Sao Luis Castro, and Thierry Olive. 2011. "Children of high transcription skill compose using bigger language bursts." In *Past, Present, and Future Contributions of Cognitive Writing Research to Cognitive Psychology*, ed. by Virginia W. Berninger, 389–402. New York: Psychology Press.

Alves, Rui A., Thierry Olive, and Sao Luis Castro. 2008. The Transcriber contributes to burst length in written language production. *Paper at the Writing Research Across Borders Conference. Santa-Barbara*. CA, USA. February 22–24.

Alves, Rui A., Sao Luis Castro, and Thierry Olive. 2008. "Execution and pauses in writing narratives: Processing time, cognitive effort and typing skill." *International Journal of Psychology* 43: 969–979. DOI: 10.1080/00207590701398951

Alves, Rui A., Sao Luis Castro, Luisa Sousa, and Sven Strömqvist. 2007. "Typing skill and pause-execution cycles in written composition." In *Writing and Cognition, Research and Applications*, ed. by Mark Torrance, Luuk Van Waes, and David Galbraith, 55–65. Dordrecht: Elsevier Sciences Publishers.

Baaijen, Veerle M., David Galbraith, and Kees de Glopper. 2012. "Keystroke Analysis: Reflections on Procedures and Measures." *Written Communication* 29: 246–277. DOI: 10.1177/0741088312451108

Biber, Douglas. 2009. "A corpus-driven approach to formulaic language in English. Multi-word patterns in speech and writing." *International Journal of Corpus Linguistics* 14 (3): 275–311. DOI: 10.1075/ijcl.14.3.08bib

Biber, Douglas, Susan Conrad, and Viviana Cortes. 2004. "If you look at…: Lexical bundles in university teaching and textbooks." *Applied Linguistics* 25 (3): 371–405. DOI: 10.1093/applin/25.3.371

Chanquoy, Lucile, Jean-Noel Foulin, and Michel Fayol. 1996. "Writing in adults: A real-time approach." In *Writing Research: Theories, Models and Methodology*, ed. by Gert Rijlaarsdam, Huub van den Bergh, and Michel Couzjin, 36–43. Amsterdam: Amsterdam University Press. DOI: 10.5117/9789053561973

Chenoweth, Anne N., and John R. Hayes. 2001. "Fluency in writing." *Written Communication* 18: 80–98. DOI: 10.1177/0741088301018001004

Chenu, Florence, François Pellegrino, Harriet Jisa, and Michel Fayol. 2014. "Interword and intraword pause threshold in the writing of texts by children and adolescents: a methodological approach." *Frontiers in Psychology* 5: 182. DOI: 10.3389/fpsyg.2014.00182

Cislaru, Georgeta. 2014. "Contraintes linguistiques et contextuelles dans la production écrite." *Les Carnets du Cediscor* 12: 55–74.

Cislaru, Georgeta, Frédérique Sitri, and Frédéric Pugnière-Saavedra. 2013. "Figement et configuration textuelle: les segments de discours répétés dans les rapports éducatifs." In *Across the Line of Speech and Writing Variation. Corpora and Language in Use – Proceedings 2*, ed. by Catherine Bolly, and Liesbeth Degand, 165–183. Louvain-la-Neuve: Presses universitaires de Louvain.

Détrie, Catherine, Paul Siblot, and Bertrand Verine. 2001. *Termes et concepts pour l'analyse du discours: une approche praxématique*. Paris: Honoré Champion. DOI: 10.1075/li.34.1.05lee

Erman, Britt, and Beatrice Warren. 2000. "The idiom-principle and the open-choice principle." *Text* 20: 29–62. DOI: 10.1515/text.1.2000.20.1.29

Foulin, Jean-Noel. 1995. "Pauses et débits: les indicateurs temporels de la production écrite." *L'Année Psychologique* 95: 483–504. DOI: 10.3406/psy.1995.28844

Foulin, Jean-Noel. 1998. "To what extent does pause location predict pause duration in adults and children writing." *Current Psychology of Cognition* 17: 601–620.

Hayes, John R., and Anne N. Chenoweth 2006. "Is working memory involved in the transcribing and editing of texts?" *Written Communication* 23: 135–149. DOI: 10.1177/0741088306286283

Hoey, Michael. 2004. "Lexical priming and the properties of text." In *Corpora and Discourse*, ed. by Alan Partington, John Morley, and Louann Haarman, 385–412. Bern: Peter Lang.

Hoey, Michael. 2005. *Lexical Priming: A New Theory of Words and Language*. London: Routledge. DOI: 10.4324/9780203327630

Hunston, Susan, and Gill Francis. 2000. *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam/Philadelphia: John Benjamins. DOI: 10.1017/s0022226701001001

Hymes, Dell. 1971. *On Communicative Competence.* Philadelphia: University of Pennsylvania Press.

Kaufer, David S., John R. Hayes, and Linda Flower. 1986. "Composing written sentences." *Research in the Teaching of English* 20: 121–140.

Kuiper, Koenraad. 2009. *Formulaic Genres*. New York: Palgrave Macmillan. DOI: 10.1057/9780230241657

Lafon, Pierre, and André Salem. 1983. "L'inventaire des segments répétés d'un texte." *Mots* 6: 161–177. DOI: 10.3406/mots.1983.1101

Langacker, Ronald. 1987. *Foundations of Cognitive Grammar*. Stanford: Stanford University Press. DOI: 10.1017/s0022226700000177

Lardilleux, Adrien, Serge Fleury, and Georgeta Cislaru. 2013. "Allongos: Longitudinal Alignment for the Genetic Study of Writers' Drafts." *Computational Linguistics and Intelligent Text Processing*, Springer LNCS 7817: 537–548. DOI: 10.1007/978-3-642-37256-8_44

Leijten, Mariëlle, and Luuk Van Waes. 2006. "Inputlog: New Perspectives on the Logging of On-Line Writing." In *Studies in Writing: Vol. 18. Computer Keystroke Logging and Writing: Methods and Applications*, ed. by Kirk P.H. Sullivan, and Eva Lindgren, 73–94. Oxford: Elsevier.

Mayaffre, Damon. 2007. "L'analyse de données textuelles aujourd'hui: du corpus comme une urne au corpus comme un plan. Retour sur les travaux actuels de topographie/topologie textuelle (partie I)." *Lexicometrica* 9. http://lexicometrica.univ-paris3.fr/numspeciaux/special9/mayaffre.pdf. (March 2, 2012).

Mey, Jacob L. 1998. "Adaptability." In *Concise Encyclopedia of Pragmatics*, ed. by Jacob L. Mey, 5–7. Oxford: Elsevier. DOI: 10.1017/s0022226799277742

Née, Emilie, Erin McMurray, and Serge Fleury. 2012. "Textometric Explorations of Writing Processes: A Discursive and Genetic Approach to the Study of Drafts." *Lexicometrica* JADT 2012 767–778.

Olive, Thierry. 2004. "Working memory in writing: Empirical evidences from the dual-task technique." *European Psychologist* 9: 32–42. DOI: 10.1027/1016–9040.9.1.32

Olive, Thierry. 2012. "Writing and working memory: A summary of theories and of findings." In *Handbook of Writing: A Mosaic of New Perspectives*, ed. by Elena L. Grigorenko, Elisa Mambrino, and David D. Preiss, 125–140. New York: Psychology Press.

Olive, Thierry. 2014. "Toward an Incremental and Cascading Model of Writing: A review of research on writing processes coordination." *Journal of Writing Research* 6: 173–194.

Olive, Thierry, and Ronald T. Kellogg. 2002. "Concurrent activation of high- and low-level production processes in written composition." *Memory & Cognition* 30: 594–600. DOI: 10.3758/bf03194960

Olive, Thierry, Alves Rui A., and Castro Sao Luis. 2009. "Cognitive processes in writing during pauses and execution periods." *European Journal of Cognitive Psychology* 21: 758–785. DOI: 10.1080/09541440802079850

Partington, Alan. 1998. *Patterns and Meanings*. Amsterdam/Philadelphia: John Benjamins. DOI: 10.1017/s0272263101221068

Salem, André. 1986. "Segments répétés et analyse statistique des données textuelles." *Histoire & Mesure* 1(2): 5–28. DOI: 10.3406/hism.1986.1518

Schilperoord, Joost. 2002. "On the cognitive status of pauses in discourse production." In *Contemporary Tools and Techniques for Studying Writing*, ed. by Thierry Olive, and C. Michael Levy, 59–85. Dordrecht: Kluwer Academic Publishers. DOI: 10.1007/978-94-010-0468-8_4

Schmid, Hans-Jörg. 2010. "Does frequency in text instantiate entrenchment in the cognitive system?" In *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches*, ed. by Dylan Glynn, and Kerstin Fischer, 101–133. Berlin: Mouton de Gruyter. DOI: 10.1515/9783110226423.101

Schmid, Hans-Jörg. In press. "A framework for understanding linguistic entrenchment and its psychological foundations in memory and automatization." In *Entrenchment, Memory and Automaticity. The Psychology of Linguistic Knowledge and Language Learning*, ed. by Hans-Jörg Schmid. Walter de Gruyter.

Schmitt, Norbert (Ed.). 2004. *Formulaic Sequences*. Amsterdam, Philadelphia: John Benjamins. DOI: 10.1093/applin/ami018

Schmitt, Norbert, Sarah Grandage, and Svenja Adolphs. 2004. "Are corpus-derived recurrent clusters psycholinguistically valid?" In *Formulaic Sequences: Acquisition, Processing and Use*, ed. by Norbert Schmitt, 127–152. Amsterdam/Philadelphia: John Benjamins. DOI: 10.1075/lllt.9.08sch

Sinclair, John. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press. DOI: 10.1177/003368829302400207

Sinclair, John. 2004. *Trust the Text. Language, Corpus and Discourse*. London/New York: Routledge. DOI: 10.4324/9780203594070

van Hell, Janet G., Ludo Verhoeven, and Liesbeth M. van Beijsterveldt. 2008. "Pause Time Patterns in Writing Narrative and Expository Texts by Children and Adults." *Discourse Processes* 45: 406–427. DOI:10.1080/01638530802070080

Verschueren, Jeff, and Frank Brisard. 2009. "Adaptability." In *Key Notions for Pragmatics*, ed. by Jeff Verschueren and Jan-Ola Östman, 28–47. Amsterdam/Philadelphia: John Benjamins. DOI: 10.1075/hoph.1.02ver

Wray, Alison. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press. DOI: 10.1017/cbo9780511519772

Yamasaki, Nozomi. 2008. "Collocations and colligations associated with discourse functions of unspecific anaphoric nouns." *International Journal of Corpus Linguistcs* 13(1): 75–98. DOI: 10.1075/ijcl.13.1.05yam