The Processing of Multiword Units by Learners of English: Evidence from Pause Placement in Writing Process Data

Gaëtanelle Gilquin

Centre for English Corpus Linguistics, Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgium; gaetanelle.gilquin@uclouvain.be

Abstract: Different methods and sources of information have been proposed in the literature to study the processing of language and, in particular, instances of formulaic language such as multiword units. This article explores the possibility of using pause placement in writing process data to determine the likelihood that a multiword unit is processed as a whole in the mind. The data are texts produced by learners of English and corresponding keylog files from the Process Corpus of English in Education (PROCEED). N-grams are selected on the basis of the finished texts and retrieved from the keylogging data. The pause placement patterns of these n-grams are coded and serve as a basis to compute the Pause Placement and Processing (PPP) score. This score relies on the assumption that n-grams which are delineated but not interrupted by pauses (hence taking the form of 'bursts of writing') are more likely to be processed holistically. The PPP score points to structurally complete n-grams such as *in fact* and *first of all* as being more likely to be processed holistically than structurally incomplete n-grams such as *that we* and *to the*. While the results are plausible and can be further substantiated by characteristics of specific n-grams, it is acknowledged that additional effects might also be at work to explain the results obtained.

Keywords: multiword units; n-grams; holistic processing; writing processes; keylogging; fluency; pauses; bursts of writing; learner language; learner corpus

1. Introduction

The issue of language processing is one that has attracted a great deal of attention among specialists in several disciplines. While experts in neuroimaging, for example, have greatly improved our understanding of this cognitive phenomenon by directly observing brain processes (see, e.g., Sabourin 2009), linguists also have an important contribution to make, using their own tools and techniques. In this article, language processing is approached on the basis of keylogging data representing writing processes, by considering how pauses in the writing process could help identify chunks of language that are processed holistically.

The object of study is formulaic language, and more precisely multiword units in English. Although the term 'multiword unit' can cover different phenomena, here the focus is on n-grams, that is, recurrent sequences of words. These n-grams are extracted from a corpus of texts written by French-speaking learners of English, the Process Corpus of English in Education (PROCEED; Gilquin 2022), as well as from the corresponding keylog files, thus revealing the processes through which these n-grams are produced.

Following a similar claim made for speech (Dahlmann and Adolphs 2007, 2009), the article explores the plausibility of using pauses in writing process data as indicators of holistic processing. The assumption is that multiword units that are produced in the form of bursts of writing, that is, with delineating pauses and no interrupting pauses, are more likely to be processed and stored as wholes in the mind. The so-called Pause Placement and Processing (PPP) score is developed to assess this likelihood.

The article starts with an overview of the literature on the processing of multiword units, including the use of pauses in speech as potential indicators of holistic storage (Section 2). Then, Section 3 deals with research on writing processes and shows how the concept of a burst of writing could be relevant to the investigation of multiword unit processing. Such an approach is adopted in the empirical study reported on in the following sections. Section 4 introduces the data, taken from PROCEED, as well as the methods, which involve extracting n-grams from the product and process data and coding them for pause placement patterns. Section 5 presents the main results of the study, first focusing on the overall distribution of pause placement patterns, then considering individual n-grams and finally proposing the PPP score as a way of assessing the likelihood with which an n-gram could be processed holistically. Section 6 offers a discussion of the results and Section 7 concludes the paper.

2. Multiword Units and Their Processing

2.1. Approaches to Multiword Units

Durrant and Mathews-Aydınlı (2011) distinguish between three types of approaches to formulaic language: phraseological approaches, frequency-based approaches and psychological approaches. While phraseological and frequency-based approaches use non-compositionality and statistical co-occurrence, respectively, as defining criteria for formulaic language, psychological approaches define formulas as "strings of linguistic items which speakers remember and process as wholes, rather than constructing them 'online' with each use" (Durrant and Mathews-Aydınlı 2011, p. 59).

Durrant and Mathews-Aydınlı (2011) recognize that there is some overlap between these three orientations. In particular, they underline the common idea that formulas, despite being analyzable into several components, are "better left unanalysed" (Durrant and Mathews-Aydınlı 2011, p. 59)—an idea that is clearly reflected by the term 'multiword unit'. Siyanova-Chanturia and Martinez (2015) make a similar point when they compare Sinclair's (1991) idiom principle and Wray's (2002) definition of formulaic sequence:

> The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments. (Sinclair 1991, p. 110)

> [A formulaic sequence is] a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored, retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar. (Wray 2002, p. 9)

While Sinclair's (1991) idiom principle can originally be related to the frequency-based approach and Wray (2002) adopts the psychological approach, both suggest that multiword units are processed holistically, as noted by Siyanova-Chanturia and Martinez (2015, p. 551): they are "semi-preconstructed" (Sinclair 1991, p. 110)/"prefabricated" (Wray 2002, p. 9) and "constitute single choices" (Sinclair 1991, p. 110)/are "stored, retrieved whole from memory at the time of use" (Wray 2002, p. 9). Many other scholars, working within different frameworks, have made similar claims about multiword units. As early as 1933, Palmer defined collocation as a succession of words that "must or should be learnt, or is best or most conveniently learnt as an integral whole or independent entity, rather than by the process of piecing together their component parts" (Palmer 1933, p. 4). Pawley and Syder (1983), in their seminal paper on the formulaic nature of language, argued for the existence of what they called "lexicalized sentence stems", which "the speaker is able to retrieve [...] as wholes or as automatic chains from the long-term memory" (Pawley and Syder 1983, p. 192). The idea that certain word sequences are processed as wholes is also evoked in current usage-based theories. In Cognitive Grammar, for example, it is said that "speakers learn as fixed units a large number of conventional expressions that are nevertheless fully analyzable and regular in formation" (Langacker 1987, p. 42). In Construction Grammar, language is said to be entirely made up of constructions, which are defined as "*stored* pairings of form and function" (Goldberg 2003, p. 219; emphasis added).

2.2. Empirical Studies of the Mental Processing of Multiword Units

Scholars who believe in the existence of multiword units do not all explicitly argue for their psychological reality and their holistic storage in the mental lexicon. Among scholars who make such a claim, not all provide evidence for this. Siyanova-Chanturia and Martinez (2015) point out that both Sinclair's (1991) idiom principle and Wray's (2002) definition of formulaic sequence lack "clear and strong empirical evidence to back them" (Siyanova-Chanturia and Martinez 2015, p. 551).

Yet, there have been attempts to investigate the mental processing of multiword units empirically, both in native language (L1) and in second/foreign language (L2) (see Conklin and Schmitt 2012; Siyanova-Chanturia 2013; Siyanova-Chanturia and Martinez 2015 for reviews). Empirical evidence has come from corpora, mainly in the form of frequency of occurrence. It is often assumed that highly recurrent sequences of words like n-grams are stored as wholes in the mental lexicon. This can be illustrated by the following quote, taken from Biber et al. (2004):

> Frequency data [...] are one reflection of the extent to which a sequence of words is stored and used as a prefabricated chunk, with higher frequency sequences more likely to be stored as unanalysed chunks than lower frequency sequences. (Biber et al. 2004, p. 376)

Importantly, Biber et al. (2004) emphasize that frequency is not the only criterion for holistic storage, giving the example of idioms, which are "usually rare but clearly prefabricated" (Biber et al. 2004, p. 376). Sometimes, simple measures of frequency are replaced or complemented by association measures such as mutual information, which can also be extracted from corpora. Wahl's (2015) study of bigrams, for example, starts from the assumption that strength of association is an "index of mental storage" (Wahl 2015, p. 191). Corpus-derived quantitative measures, however, should be interpreted with caution when it comes to psychological matters like holistic storage. Schmitt et al. (2004, p. 147) stress that "it is unwise to take recurrence of clusters in a corpus as evidence that those clusters are also stored as formulaic sequences in the mind".

Experimentation has been proposed as another empirical method for investigating the holistic nature of multiword units, viewed by some as more compatible with the study of mental phenomena than corpus data. Sometimes, experimentation takes as a starting point a set of expressions shown to be frequent in corpora, with the aim of testing the relevance of frequency as a criterion for holistic storage. Experimental methods used in the analysis of multiword units include auditory word-monitoring tasks (e.g., Sosa and MacFarlane 2002), oral dictation tasks (e.g., Schmitt et al. 2004), lexical decision tasks (e.g., Cieślicka 2006), self-paced reading (e.g., Kim and Kim 2012), eye tracking (e.g., Siyanova-Chanturia et al. 2011b) and electrophysiological measures (e.g., Tremblay and Baayen 2010). Relying on one or several experimental methods, studies have brought to light processing advantages for (frequent) multiword units, in the sense that they tend to be processed more quickly than novel strings of language, are easier to recall and reproduce, are read faster (with fewer and shorter eye fixations), etc. Such evidence has led many scholars to argue that multiword units are stored as wholes in the mental lexicon. In some cases, the results are less clear or indicate both holistic storage of the multiword units and storage of the individual words making them up (cf. Tremblay and Baayen 2010).

As regards the comparison between L1 and L2, some studies find the same processing advantages for multiword units among L1 and L2 speakers (e.g., Jiang and Nekrasova 2007; Conklin and Schmitt 2008), thus pointing to holistic storage of multiword units among L2 speakers too. Other studies highlight differences between L1 and L2 speakers. Ellis et al. (2008), for example, show that L1 speakers' processing of multiword units is mainly determined by the strength of association (mutual information), whereas for L2 speakers it is mainly the frequency of occurrence that determines processing. The results of Siyanova-Chanturia et al. (2011a) suggest that, unlike L1 speakers, L2 speakers do not process idiomatic expressions more quickly than novel phrases.

Despite some mixed results, the literature predominantly brings out the processing advantages of multiword units, especially among L1 speakers, but possibly among L2 speakers too. Such advantages are often taken as evidence for the psychological reality of multiword units. Conklin and Schmitt (2012, p. 54) state that "given all the evidence for the processing advantages of formulaic language […], it is difficult to believe that it does not somehow exist in the mind". However, scholars such as Siyanova-Chanturia and Martinez (2015) insist that the higher speed of processing of multiword units does not necessarily imply that they are stored holistically in the mind. Findings based on experimentation, like those based on corpus studies, should therefore be interpreted cautiously, and if possible, converging evidence based on various types of data should be sought (cf. Ellis and Simpson-Vlach 2009).

2.3. Pauses as Potential Indicators of Holitistic Storage

While much of the literature on the processing of multiword units has focused on how frequent they are and/or how quickly they can be comprehended, some studies with speech as their object of investigation have explored their phonological features. It has been claimed that multiword units should display "phonological coherence" (Peters 1983, p. 10), which involves having a single intonation contour and being produced with no hesitations. The focus here will be on the potential role of hesitations within multiword units which, as will be suggested in Section 3.2, could be relevant to the investigation of writing too, through the notion of writing fluency.

Pawley (1986, p. 107) notes that "pauses within lexicalized phrases are less acceptable than pauses within free expressions". This feature has been taken as possible evidence for the holistic storage of multiword units. Thus, Wray (2002) notes the following:

> It certainly seems a reasonable hypothesis that, if formulaic sequences are retrieved whole from memory (or at least with less recourse to on-line rule application and lexical retrieval than novel utterances), they should be produced more fluently than novel ones. (Wray 2002, pp. 35–36)

A few studies have tested—and confirmed—this hypothesis by examining whether multiword units are likely to be interrupted by pauses in speech. Van Lancker et al. (1981) examine the features of identical phrases with a literal and idiomatic meaning (e.g., *skating on thin ice*) produced orally by speakers asked to convey their contrasting meanings. They find that pauses are five times as frequent in the literal utterance as in the idiomatic one. Erman (2007) looks at (filled and silent) pauses in a corpus of spoken English and, for each of them, determines whether they are related to an open slot (non-prefabricated language) or a restricted one (prefabricated language). She shows that pauses are more frequent in connection with open slots (88.7%) than restricted slots (11.3%). Flinn (2023) investigates four-grams in a corpus of spoken English and discovers that silent pauses within four-grams occur with a frequency of 12.73 per 100,000 words, to be compared with 2205.85 pauses per 100,000 words in the whole corpus. In Schneider (2014), it is demonstrated that sequences of words that are likely to co-occur in a corpus of English telephone conversations (as determined by frequency or association measures) tend to be produced fluently, i.e., without being interrupted by filled pauses, silent pauses or discourse markers.

In their comparison of L1 and L2 speakers of English, Schmitt et al. (2004) briefly consider disfluent features in the participants' reproduction of multiword units in an oral dictation task. They show that hesitations, stutters and false starts are twice as frequent in the L2 participants' multiword units as in those of the L1 participants. Rajtar (2016) comes to a similar conclusion with respect to silent pauses on the basis of naturally occurring language from corpora.

It has also been suggested that while hesitations are unlikely to be found within multiword units, they could occur at their boundaries. Boomer (1965) argues for the presence of hesitations at the beginning of what he calls "encoding units". His argument is that if hesitations occur "at points where decisions and choices are being made" (Boomer 1965, p. 148), they should predominantly be found near the beginning of multiword encoding units. Raupach (1984) investigates formulae, which he sees as "indicators of processing units" (Raupach 1984, p. 116). In order to identify these formulae, he suggests looking for "utterance sequences delimited by hesitation phenomena" (Raupach 1984, p. 128).

Schneider's (2014, p. 248) study appears to support this view of multiword units delimited by hesitations (including pauses), as it reveals a tendency for hesitations to "fall at the boundaries of frequency-derived units, which often, but not always, coincide with traditionally-assumed phrases". The delineation of multiword units by pauses on both sides, combined with their lack of interruption, leads to the concept of "pause-defined unit" (Brown et al. 1980) or "fluent unit" (Pawley and Syder 1983), which often serves as a measure of fluency (cf. mean length of runs, where 'runs' refer to "continuous speech with no pauses or hesitations" (Olynyk et al. 1987, p. 124)).

Dahlmann and Adolphs (2007, 2009) rely on both the absence of interrupting pauses and the presence of delineating pauses to study the holistic storage of multiword units. Their assumption is that "pauses are indirect indicators of prefabricated language and holistic storage" (Dahlmann and Adolphs 2007, p. 50). They put this assumption to the test by examining pause placement patterns in a spoken corpus of L2 English (Dahlmann and Adolphs 2007) and L1 English (Dahlmann and Adolphs 2009). In Dahlmann and Adolphs (2007), the focus is on two frequent three-grams, namely *I don't know* and *I think I*, while in Dahlmann and Adolphs (2009), the focus is on the sequence *I think*.

What Dahlmann and Adolphs (2007, p. 53) call the "'ideal' form", that is, the use of the multiword unit with pauses on both sides and no pauses in between, is relatively infrequent in their corpus data, representing 21% with *I don't know* and 9% with *I think I* in L2 English and 11% with *I think* in L1 English. However, if we consider cases where the multiword unit is not interrupted by pauses and an immediately adjacent pause is found on at least one of its sides, the proportions rise to 72% for *I don't know*, 43% for *I think I* and 51% for *I think*. The differences in proportions could be due to the nature of the multiword units (structurally complete or not)

and the status of the speakers (L1 or L2), with higher proportions for structurally complete multiword units than for structurally incomplete ones (compare *I don't know* and *I think I*) and higher proportions among L2 speakers than among L1 speakers (compare *I don't know* and *I think*).

Importantly, cases where the structurally complete multiword units are interrupted by pauses are extremely rare: 1% with *I think* and no occurrences with *I don't know*. With *I think I*, they represent 18%. It should also be emphasized that, as pointed out by Dahlmann and Adolphs (2009, p. 138), "[t]he absence of a pause […] does not exclude the possibility that there might in fact be a boundary", since the multiword units under study could be embedded within larger chunks of language. This could potentially explain the findings for L1 vs. L2 speakers described above, namely that structurally complete multiword units not interrupted by pauses and with an immediately adjacent pause on at least one side are proportionally less frequent in L1 speech (51%) than in L2 speech (72%). It actually appears from the results of Dahlmann and Adolphs (2007, 2009) that cases where the multiword units are not surrounded by pauses on any side are proportionally more frequent in L1 speech (48%) than in L2 speech (28%), which could point to the processing of larger chunks of language in L1 than in L2.

The evidence gathered by Dahlmann and Adolphs (2007, 2009) about pause placement patterns in speech leads them to suggest that *I don't know* and *I think* might be stored holistically in the mind. On the other hand, they are more cautious when it comes to the interpretation of the results for *I think I*.

It must be acknowledged that not everybody agrees that pause placement can serve as evidence for the holistic storage of multiword units, at least not necessarily among all speakers. The notion of phonological coherence, which includes the lack of interrupting pauses, was introduced by Peters (1983) in the context of child language research. She cites a study by Rosenberg (1977) to claim that "for adults hesitation pauses are not reliable indicators of the size and nature of encoding units" (Peters 1983, p. 10). As for the delineation of multiword units by pauses, Raupach's (1984) claim was made in the context of L2 production, but he points out that his approach may give less reliable results for L1 speakers and highly competent learners.

Lin (2018) is skeptical about the possibility of relying on pauses to prove the psycholinguistic reality of multiword units among adult L1 speakers and proficient learners, although her main argument is that the speech of these two groups will be more fluent and hence less likely to contain enough pauses for such an analysis to be valid. Her empirical study seems to confirm this, with only nine out of sixty-two multiword units being completely delineated by pauses, and eleven multiword units being interrupted by pauses. However, her study is based on the transcription of a single lecture by one person amounting to less than 9000 words, which may not be sufficient to draw reliable generalizations. In addition, as shown above, other studies have demonstrated the relevance of pause placement patterns for the study of multiword unit processing, with corpora of either L1 or L2 speech providing enough evidence for this type of analysis. However, it is fair to say that such evidence should ideally be combined with additional criteria such as intonation contours.

Discussing the use of phonological coherence to identify multiword units, Wray (2002, p. 35) states that this kind of approach "is, of course, restricted to the spoken language"—although she admits that punctuation and layout in written texts could somehow reflect similar characteristics. However, provided one can observe the process through which a text is written, notions such as fluency and pauses can be applied to writing too, and the development of technologies such as screencasting and keylogging has made it easier to investigate such aspects, as outlined in the next section.

3. Writing Processes

3.1. Writing Process Research

While speech is necessarily approached as an ongoing process, due to the online nature of spoken production, writing has mostly been studied as a finished product, on the basis of texts in their final version, ready to be presented to the reader. However, as argued by several scholars over the last few decades (e.g., Hairston 1982; Prior 2004; Révész and Michel 2019), the investigation of writing processes, that is, the different steps involved in the composition of a text until it reaches its final state, has a great deal to offer.

Attention to writing processes began in pedagogy, with the development of the so-called writing-as-a-process movement in the 1960s (see Grabe and Kaplan 1996). This movement advocated focusing on the composition process in writing instruction and viewed planning, drafting and revision as essential parts of this process. In the 1970s, this interest in writing processes spread to research (see, e.g., Graves 1975), and methods were introduced to observe writing processes, such as think-aloud protocols, which involve asking subjects to express their thoughts aloud while writing a text (e.g., Hayes and Flower 1983), and stimulated recalls (e.g., Stolarek 1994), in which some stimulus (such as a video of the writer's hand during the writing session) is used retrospectively to help subjects recall their thoughts while they were writing. Recording of handwriting is another way in which writing processes have been observed, either through videotaping of the writer at work (e.g., Pianko 1979) or, more recently, through special software directly capturing the handwriting movements (e.g., Ductus, described in Guinet and Kandel 2010). Technological developments have led to more methods of observation, including screencasting (recording of the screen activity), keylogging (recording of the keyboard activity) and eye tracking (recording of eye movements); see, e.g., Sullivan and Lindgren (2006), Breuer (2017), Gánem-Gutiérrez and Gilmore (2018), Takayoshi (2018) or Lindgren and Sullivan (2019).

Grabe and Kaplan (1996, p. 89) situate the origin of writing process research in the field of cognitive psychology. Indeed, writing process research, which can now arguably be seen as a field of its own, still has obvious links with cognitive psychology, including in terms of methods (writing process research mainly relies on experimental methods, some of which are used in psychology too), concepts that are of importance in both fields (e.g., working memory) and contributors to writing process research, many of whom

have training in psychology. Cognitive models of writing processes have been developed, most notably by Flower and Hayes (see Flower and Hayes 1981a and further developments of the model), and cognitive strategies have been studied on the basis of writing process data (e.g., Bloom 2008). More generally, writing processes have been thought to give access to various cognitive aspects, such as self-monitoring, attentional capacity or processing, despite the difficulties involved in "aligning" writing process measures with cognitive processes (Baaijen et al. 2012).

3.2. Writing Fluency and Bursts of Writing

While the notion of fluency is more typically associated with speech, it is also relevant to writing. Writing is considered fluent if the composition process runs smoothly, with few pauses and revisions (see, e.g., Abdel Latif 2013). In writing, a pause corresponds to a period of time during which the writer is not actually writing. Fluency, including the presence of pauses and revisions, can be observed in writing process data, for example on a screencast video. It can also be measured—with high accuracy and taking multiple factors into account—on the basis of keylogging data (cf. Van Waes and Leijten 2015).

Pauses have received a great deal of attention in writing process research, one reason being that they arguably "offer observable clues to the covert cognition processes which contribute to discourse production" (Matsuhashi 1981, p. 114). Scholars have considered the frequency of pauses, but also their duration and location. Matsuhashi (1981), for example, shows that writers tend to pause longer when they write for the purpose of generalizing or persuading than for reporting. Her results also reveal that pausing time is usually longer before T-units than within T-units and longer before T-units starting a paragraph than before other T-units. Subsequent research has confirmed that "pause rates and durations in written composition are not random" (Medimorec and Risko 2017, p. 1269), varying according to genres, text boundaries and individual writing styles. Differences between L1 and L2 pausing behaviors have also been brought to light. Spelman Miller (2000), for instance, demonstrates that L2 writers pause longer than L1 writers, especially between clauses and sentences.

Particularly interesting for our purposes is the concept of a burst of writing, which is a segment of text written fluently, in one go. It can be said to correspond to pause-defined or fluent units in speech (see Section 2.3). Bursts of writing have been identified in different ways. The first studies of bursts relied on think-aloud protocols, and, in effect, bursts were identified in writers' spoken production while they were writing (cf. Kaufer et al. 1986; Chenoweth and Hayes 2001). Later, however, bursts came to be identified on the basis of keylogging data, showing writing processes in real time (cf. Hayes and Chenoweth 2006). In Chenoweth and Hayes (2001), a distinction is drawn between P-bursts, which end with a pause, and R-bursts, which end with a revision. The length of bursts is calculated by counting the number of new words in each segment of the protocol (repeated words are disregarded). In Breuer (2019), bursts are identified in keylog files as segments ending with a pause, a revision or a movement of the mouse to another place in the text. In Cislaru and Olive (2018), only pauses are taken into account to identify bursts, which are defined as any text produced between two pauses. Different thresholds can be chosen to determine pauses, and hence bursts: in Breuer (2019), for example, the minimum threshold is one second, whereas in Cislaru and Olive (2018), it is two seconds. The length of bursts can be counted in words (as in Chenoweth and Hayes 2001), in characters (as in Breuer 2019) or in milliseconds (cf. Van Waes and Leijten 2015). In the present study, bursts will be identified as segments of text produced between two pauses of at least two seconds.

Bursts have mainly been used to investigate writing fluency by measuring their mean length in writing process data: the longer the bursts are, the more fluent the writing is. It thus appears that expert writers produce longer bursts on average than novice writers (Kaufer et al. 1986) and that L1 texts are written in longer bursts than L2 texts (Van Waes and Leijten 2015). Occasionally, the linguistic structure of bursts has been investigated, which has revealed that bursts have their own logic and act as linguistic units in writing processes, which may be different from the units traditionally recognized in grammar (Cislaru and Olive 2018). Given that pauses are assumed to reflect cognitive processes (see above), bursts as demarcated by pauses have also been considered for their possible contribution to the study of language processing. Chenoweth and Hayes (2003), for example, have shown that average burst length decreases with reduced working memory capacity. Cislaru and Olive (2017, p. 15) suggest that "the spontaneity of burst production may, until proven otherwise, indicate a certain degree of automatization, which would result in retrieval from long-term memory rather than the implementation of text generation processes" (my translation). This quote, which may remind us of Sinclair's (1991) idiom principle or Wray's (2002) definition of formulaic sequences, mentioned in Section 2.1, underlines the potential of bursts for the study of multiword units.

3.3. This Study

The objective of the present study is to build on previous research on speech and writing fluency to explore the use of pauses in writing process data—including in the form of bursts—as indicators of holistic processing. As is the case in the study of Dahlmann and Adolphs (2007), who made a similar claim for speech, this study is carried out on frequent n-grams. While such sequences may not all have psychological reality (Section 2.2), they are arguably more likely to be processed as wholes than word sequences taken randomly from a corpus. These n-grams are selected on the basis of finished texts and then retrieved from the corresponding keylog files, which means that the analysis also makes it possible to approach the process/product interface (Cislaru 2015). In doing so, it is quite similar in spirit to the work of Olive and Cislaru (2015), who compared repeated segments in product data and bursts of writing in process data. In Olive and Cislaru (2015), however, all bursts were considered, not just those corresponding to repeated segments as is the case here. As for the repeated segments from the finished texts, in Olive and Cislaru (2015), they were only retrieved from the process data if they took the form of bursts of writing, while here, their occurrences outside bursts are also examined in the keylog files.

As is the case in Dahlmann and Adolphs (2007), the data used in this study represent L2 English. These are the data that constitute the core of the Process Corpus of English in Education (Section 4.1). Using L2 data also has the advantage of avoiding one of the main criticisms leveled against the use of pauses in speech to test the psychological reality of multiword units among adult L1 speakers, namely the (supposedly) low frequency of pauses. It has been estimated that between 50% (Olive and Cislaru 2015, p. 101) and 70% (Flower and Hayes 1981b, p. 229) of composing time among adult L1 writers is actually spent pausing. Since L2 writers have been shown to generally pause more than L1 writers (Section 3.2), pauses in L2 writing process data are expected to be frequent enough to allow for a pause-based analysis.

4. Data and Methods

4.1. The Process Corpus of English in Education

The data used in this study come from the Process Corpus of English in Education (PROCEED; Gilquin 2022), which is a corpus representing argumentative essay writing by university students who are learners of English as a foreign language (EFL). The corpus includes L2 English data as well as data produced by the same students in their L1, but only the former are exploited here. In addition to the finished texts, PROCEED includes, for each text, a keylog file, produced by Inputlog (Leijten and Van Waes 2013), and a screencast video, recorded by means of OBS Studio (Bailey and OBS Studio Contributors 2012). The corpus also comes with metadata, which provide (socio)linguistic information about the students and their results on some cognitive tests.

The study relies on a sample of 42 texts (14,420 words) from PROCEED, all written by intermediate to advanced EFL learners with French as an L1. A so-called linear analysis of the corresponding keylogging data was carried out by means of Inputlog. Linear analysis represents all the keys struck on the keyboard as well as all the pauses. Its output is illustrated in (1). The figures between curly brackets, in blue, represent the duration of the pauses in milliseconds (ms). Following the criteria for identifying bursts of writing applied by Chenoweth and Hayes (2001) and Olive and Cislaru (2015), a minimum threshold of 2000 ms was set for pauses to appear in the linear analysis. The indications between square brackets, in gray, show the actions carried out such as deletion ([BACK]) or capitalization ([CAPS LOCK]), while the interpuncts (·) stand for spaces. The remaining elements represent the text typed. In effect, any text occurring between two pauses corresponds to a burst of writing. The following bursts can be observed in (1) (after implementation of the revisions): "Furthermore,", "people are not used to", "take action", "when something bad happens" and ".".

(1)   {2144}[CAPS LOCK]F[CAPS LOCK]urte[BACK]hermore,·{11144}people·are·not·use·[BACK]d·to·{4016}take·action·{5664}when·something·bad·happens{3776}.·{5080}

4.2. Extraction and Coding of N-Grams

Two- to six-grams were automatically extracted from the finished texts (product data) of the PROCEED sample by means of Word-Smith Tools (Scott 2008) and were then searched for in the output of the linear analysis (process data). In an attempt to focus on n-grams that are sufficiently frequent in the (product and process) data, but also representative and varied enough, twenty n-grams were selected for further investigation according to the following criteria: frequency (at least five occurrences in the product data and ten occurrences in the process data), range (the n-grams had to occur in at least five different texts), grammar (mix of syntactic sequences such as *of the* and more lexical ones such as *in my opinion*), semantics (no topic-dependent n-grams like *change the world*) and length (n-grams of various lengths). The twenty selected n-grams are listed alphabetically in (2).

(2)   a good; a lot of; able to; first of all; for example; for instance; I think; in conclusion; in fact; in my opinion; of the; on the other hand; some people; such as; that we; the most; the same; to sum up; to the; us to.

While the extraction of the n-grams from the product data was fully automatic, their retrieval from the process data had to be semi-automatic. This is because the output of the linear analysis includes all sorts of elements that may interrupt the n-grams and preclude their automatic identification, as shown in (3) to (5). In (3), the production of *I think* involved some revision. Although the typo was immediately corrected by the writer, who replaced *j* with *h*, the trace of this revision, through the use of the delete key, prevents automatic retrieval of the n-gram. In (4), it is the indication of the pause that stands in the way of the n-gram *in the*. In (5), no element interrupts the n-gram *in my opinion*, but the misspelling (*opnion*) makes automatic retrieval impossible too.

(3)   I·tj[BACK]hink
(4)   i{2350}n·the
(5)   In·my·opnion

The retrieval of such forms was facilitated by the use of VariAnt (Anthony 2017), a program that relies on the formal similarity between a target word and the other words in the corpus to suggest variants. VariAnt was used here to detect potential variants of the words making up the selected n-grams in the linear analysis output. For example, it detected that *hink* was a potential variant of *think* and *opnion* a potential variant of *opinion*, which made it possible to retrieve (3) and (5). Since not all forms proposed by VariAnt were actual variants of the target words, it was necessary to check them manually in context and only include those that actually corresponded to the target words. In total, 538 instances of the n-grams were retrieved from the process data. It should be pointed out that a few instances might have been missed in case the forms were too different from the target words to be recognized, either by VariAnt or by the naked eye.

All instances of the n-grams retrieved from the process data were coded using a system adapted from Dahlmann and Adolphs (2007, 2009). The different patterns are listed in Table 1, where P stands for a pause, X for the n-gram (in bold in the examples of Table 1 and in all further examples) and the underscore for anything else. The first pattern, _X_, represents n-grams occurring with no immediately adjacent pauses. PX_ is when the n-gram is immediately preceded but not followed by a pause, and _XP is the opposite. In PXP, the n-gram is immediately preceded and followed by a pause. In effect, this means that the n-gram corresponds to a burst of writing. All these patterns are characterized by the absence of pauses within the n-gram. By contrast, in the last pattern, XPX, the n-gram is interrupted by at least one pause.

Table 1. Pause placement patterns for the coding of n-grams in writing process data (P = pause; X = n-gram; _ = anything else).

Pattern	Example
_X_	{12960}go·fut[BACK]rht[BACK]er·**in·the**·world[RETURN]{19224}
PX_	{13616}[CAPS LOCK]F[CAPS LOCK]**or·example**,·presidents·use·their·power·{17584}
_XP	{4912}counrty[BACK][BACK][BACK]try·**such·as**·{11788}
PXP	{5728}[LSHIFT]**First·of·all**,·{3224}
XPX	{2776}m[BACK]most·**of**·{2200}**the**·really-educated·characters·[Movement]{5265}

Following claims made in the literature (Sections 2 and 3), the PXP pattern, with a pause on both sides, should characterize word sequences that are processed holistically, whereas the XPX pattern, with a pause in the middle, should characterize word sequences that are not processed holistically. If n-grams are processed holistically, the PXP pattern is therefore expected to be more frequent than the XPX pattern. The PX_ and _XP patterns, with a pause at one boundary, could provide some partial evidence for the holistic processing of n-grams. As for the _X_ pattern, it "arguably offers the least reliable information about possible MWE [multiword expression] boundaries" (Dahlmann and Adolphs 2009, p. 137), since the lack of pauses on both sides does not necessarily indicate that the n-gram is not processed holistically. Based on the results of Dahlmann and Adolphs (2007, 2009), however, _X_ might turn out to be a common pattern in the data.

5. Results
5.1. Overall Results

Table 2 shows the distribution of pause placement patterns for all twenty n-grams taken together. It appears that the most common pattern, with 58.74%, is _X_, that is, the n-gram without any immediately adjacent pauses, as illustrated by (6) and (7). The absence of pauses at the boundaries of the n-gram could indicate a lack of holistic processing. However, as noted above, it could be that the n-gram is still processed as a whole but embedded within a larger chunk of language and that it is this larger chunk that is demarcated by pauses. In (7), for example, *kind of* could be processed as a whole, but be uttered as part of a larger chunk made up of the relative clause *which is a kind of basis*, which is preceded and followed by a pause.

Table 2. Distribution of pause placement patterns for all selected n-grams: raw frequencies (n) and percentages (%).

	_X_	PX_	_XP	PXP	XPX	Total
n	316	80	58	48	36	538
%	58.74%	14.87%	10.78%	8.92%	6.69%	100%

(6)　　{3848}is·**th**·**emost**·powerful·{2504}wep[BACK]apon
(7)　　all·rev[BACK]ceived·an·education·{4832}which·is·a·**kind·of**·basis{4607}·and·which·is·part·of·us

The proportions of the other pause placement patterns range between 6% and 15%. The category with the lowest percentage is XPX (n-gram interrupted by a pause), which represents 6.69%. This is to be expected if n-grams are processed as wholes, since their interruption by pauses should then be quite exceptional. It should be noted that, occasionally, the effect of the interruption is mitigated by the presence of pauses on both sides of the n-gram, as exemplified by (8). Although *first of all* is interrupted by a pause (hence the classification as XPX), it is also directly preceded and followed by long pauses, which is arguably a characteristic of holistically stored multiword units. Such cases, however, account for only three of the thirty-six instances of XPX found in the data.

(8)　　{15216}[RETURN][LSHIFT]**First**{7368}·**of·all**,·{12896}

The next category, with 8.92%, is PXP, where the n-gram takes the form of a burst of writing. As predicted, this category is more frequent than XPX (interruption by a pause), although the difference is relatively small (8.92% vs. 6.69%). The low proportion of PXP, on its own, could be said to question the holistic nature of n-grams. On the other hand, it partly confirms findings from Olive and Cislaru (2015), who show that bursts of writing and repeated segments are essentially distinct phenomena, with less than 3% overlap in their data.

The patterns with a pause at only one boundary (PX_ and _XP), finally, represent 14.87% and 10.78%, respectively, and thus together amount to a quarter of the data.

Overall, these results do not provide very strong evidence for the holistic processing of n-grams, since the predominant category (_X_) does not allow for any firm conclusion to be drawn in this respect. The categories that could constitute strong evidence either for (PXP) or against (XPX) holistic processing are both relatively infrequent in the data, and the former is only slightly more common than the latter. In the next section, however, we will see that not all n-grams behave in the same way with regard to pause placement.

5.2. Individual N-Grams

Table 3 presents the distribution of pause placement patterns for each of the twenty n-grams under study, ordered alphabetically. The figures in bold signal the most frequent pattern(s) for each n-gram, and the underlined figures indicate cases where the n-gram is interrupted by a pause with a percentage higher than 10%.

Table 3. Distribution of pause placement patterns for each of the twenty selected n-grams: percentages (and raw frequencies). The bold indicates the most frequent pattern(s) per n-gram; the underline indicates XPX patterns with a percentage over 10%.

	_X_	PX_	_XP	PXP	XPX
a good	**77.78% (28)**	11.11% (4)	0.00% (0)	5.56% (2)	5.56% (2)
a lot of	**47.83%** (11)	34.78% (8)	4.35% (1)	8.70% (2)	4.35% (1)
able to	**80.95%** (17)	9.52% (2)	4.76% (1)	0.00% (0)	4.76% (1)
first of all	7.14% (1)	21.43% (3)	0.00% (0)	**64.29% (9)**	7.14% (1)
for example	**47.06% (8)**	29.41% (5)	11.76% (2)	11.76% (2)	0.00% (0)
for instance	5.56% (1)	27.78% (5)	**33.33% (6)**	**33.33% (6)**	0.00% (0)
I think	**52.63% (10)**	31.58% (6)	5.26% (1)	5.26% (1)	5.26% (1)
in conclusion	36.36% (4)	9.09% (1)	9.09% (1)	**45.45% (5)**	0.00% (0)
in fact	7.14% (1)	35.71% (5)	7.14% (1)	**50.00% (7)**	0.00% (0)
in my opinion	6.67% (1)	**53.33% (8)**	13.33% (2)	26.67% (4)	0.00% (0)
of the	**76.77% (76)**	5.05% (5)	4.04% (4)	0.00% (0)	14.14% (14)
on the other hand	**36.36% (4)**	27.27% (3)	0.00% (0)	18.18% (2)	18.18% (2)
some people	**50.00% (6)**	33.33% (4)	16.67% (2)	0.00% (0)	0.00% (0)
such as	**50.00% (9)**	11.11% (2)	22.22% (4)	16.67% (3)	0.00% (0)
that we	**67.86% (19)**	10.71% (3)	10.71% (3)	0.00% (0)	10.71% (3)
the most	**69.62% (55)**	7.59% (6)	18.99% (15)	1.27% (1)	2.53% (2)
the same	**79.55% (35)**	4.55% (2)	15.91% (7)	0.00% (0)	0.00% (0)
to sum up	0.00% (0)	**40.00% (4)**	30.00% (3)	30.00% (3)	0.00% (0)
to the	**57.14% (16)**	10.71% (3)	7.14% (2)	0.00% (0)	25.00% (7)
us to	**66.67% (14)**	4.76% (1)	14.29% (3)	4.76% (1)	9.52% (2)

Corresponding to the overall results, a majority of the n-grams (14 out of 20) are mostly used with no immediately adjacent pauses (_X_). Among these, there may still be different profiles, though. A comparison of *the same* and *of the*, for instance, reveals that they are predominantly found with no immediately adjacent pauses (_X_) and in similar proportions (79.55% and 76.77%, respectively), but they differ in the extent to which the n-gram is interrupted by a pause (XPX): it is never the case with *the same*, whereas it is the case in over 14% with *of the*. The n-gram *on the other hand* is mostly used in the _X_ pattern too, but with a lower proportion (36.36%). In addition, it is produced within a single burst (PXP) in 18% of its occurrences, which suggests holistic processing, but it is equally often interrupted by a pause (XPX), which suggests a lack of holistic processing.

It can also be noticed from Table 3 that n-grams that do not predominantly occur in the _X_ pattern display different preferences. *First of all* is predominantly found in the PXP pattern (64.29%), for example, whereas *to sum up* favors the PX_ pattern (40%). Interestingly, a few n-grams mostly take the form of a burst (PXP). In addition to *first of all*, this is the case of *for instance* (33.33%, on par with _XP), *in conclusion* (45.45%) and *in fact* (50%). By contrast, none of the n-grams have a preference for the XPX pattern, which corresponds to the interruption of the n-gram by a pause. The highest percentage for this category is 25%, with the n-gram *to the*, and out of the twenty n-grams under study, almost half (nine) are never interrupted by a pause. This difference between the PXP and XPX categories seems to point to possible holistic processing, at least for certain specific n-grams.

Importantly, what Table 3 brings out is the variety of profiles of the n-grams with respect to pause placement patterns. Thus, for some n-grams, evidence for possible holistic processing comes mainly from the high frequency of the PXP pattern (cf. *first of all*), whereas, for other n-grams, several patterns provide cumulative evidence (e.g., high frequency of PX_ and no occurrences of XPX for *in my opinion*). This variety makes it difficult to compare the different n-grams. In the next section, the pause placement patterns are therefore translated into a single score which can be used for comparative purposes.

5.3. The Pause Placement and Processing (PPP) Score

For each of the n-grams, a score has been computed that factors in the different pause placement patterns and their respective proportions. This score, which I have called the Pause Placement and Processing (PPP) score, aims to reflect the potential of the n-gram to be processed holistically, based on the assumptions made in the literature about the link between pause placement and holistic processing (see Sections 2.3 and 3.2). It is computed as follows:

(9) Pause Placement and Processing (PPP) score:
    (% PXP) − (% XPX) + (% PX_/2) + (% _XP/2)

The PPP score assigns a particular weight to each pause placement pattern depending on its capacity to reflect holistic processing. Thus, cases where the n-gram is neither immediately preceded nor immediately followed by a pause (_X_) are neutral in this respect, because a lack of adjacent pauses does not necessarily imply a lack of holistic processing. As noted above, the n-gram could be stored holistically but produced as part of a larger chunk of language. This pause placement pattern is therefore excluded from the computation. Next, the pause placement pattern that corresponds to a burst (PXP) is taken to be strong evidence for the holistic processing of the n-gram. The proportion of PXP for the n-gram under study is therefore counted as a positive value in the PPP score. By contrast, the pause placement pattern where the n-gram is interrupted by a pause (XPX) offers strong evidence against holistic processing. The proportion of XPX is therefore counted as a negative value in the PPP score. In addition, it is assumed that cases where the n-gram is only preceded or only followed by a pause (PX_ and _XP) could indicate holistic processing but are less strong indicators than cases where the n-gram is both preceded and followed by a pause (PXP). The proportions of PX_ and _XP are therefore counted as positive values but divided by two, reflecting the fact that the n-gram is demarcated by a single pause, not two.

An example of the computation of the PPP score can be found in (10) for the n-gram *a lot of*. It is based on the results shown in Table 3 for this n-gram. The proportion of _X_ (47.83%) is disregarded. The percentage of PXP (8.7%) is added as a positive value, while the percentage of XPX (4.35%) is subtracted. The percentages of PX_ (34.78%) and _XP (4.35%) are divided by two and added as positive values. The total score amounts to 23.915.

(10) PPP score of *a lot of*:
    8.7 − 4.35 + (34.78/2) + (4.35/2) = 23.915

The highest possible score, representing a very strong likelihood that the n-gram is processed holistically, would reflect a situation in which all instances of the n-gram are demarcated by pauses on both sides and never interrupted by pauses. In other words, the PXP pattern, corresponding to a burst, would have a proportion of 100%, and the PPP score would therefore amount to 100.

Table 4 lists the n-grams in descending order of their PPP score. The n-gram with the highest PPP score is *in fact*, which has a score of 71.43. The PPP scores then decrease until the n-gram *that we*, which has a score of 0. Two n-grams have a negative score, namely *of the* (−9.60) and *to the* (−16.07). A negative PPP score means that the proportion of XPX (n-gram interrupted by a pause) is higher than the combined proportion of all the other pause placement patterns included in the computation (PXP, PX_ and _XP). For expository purposes, the results will be presented in three groups: n-grams with the highest PPP scores (above 50), n-grams with the lowest scores (below 5) and all other n-grams.

Table 4. Pause Placement and Processing (PPP) score of the twenty selected n-grams.

N-Gram	PPP Score
in fact	71.43
first of all	67.86
to sum up	65.00
for instance	63.89
in my opinion	60.00
in conclusion	54.55
such as	33.33
for example	32.35
some people	25.00
a lot of	23.91
I think	18.42
on the other hand	13.64
the most	12.03
the same	10.23
a good	5.56
us to	4.76
able to	2.38
that we	0.00
of the	−9.60
to the	−16.07

There are six n-grams in the list that have a PPP score higher than 50: *in fact* (71.43), *first of all* (67.86), *to sum up* (65), *for instance* (63.89), *in my opinion* (60) and *in conclusion* (54.55). A comparison of the individual profiles of the six n-grams in Table 3 reveals that none of them have a majority of occurrences with no adjacent pauses (_X_). Instead, they predominantly occur either in the PXP (burst) pattern (for *in fact*, *first of all*, *for instance*[1] and *in conclusion*), as in (11), or in the PX_ pattern (for *to sum up* and *in my opinion*), as in (12).

(11)   {2096}[CAPS LOCK]**I**[CAPS LOCK]**n·fact**,·{4928}edc[BACK][BACK]ducation·{3120}

(12)   {3328}**in·my·opiniou**[BACK]**n**·a·way·to·{2912}

Interestingly, several of these n-grams have been shown to be overused by EFL learners (see Gilquin et al. (2007, pp. IW15, IW23, IW28) on *in my opinion*, *first of all* and *to sum up*, and Paquot (2008, p. 109) on *for instance*), and sometimes specifically by French-speaking EFL learners (see Gilquin and Granger (2015, p. 432) on *in fact*). This suggests high familiarity with—and possibly strong entrenchment of—these multiword units among learners, which could lead to holistic processing. In addition, these multiword units are often used in sentence-initial position in the corpus, as exemplified in (11) for *in fact*. This means that there is potentially a double boundary on the left, one for the beginning of the sentence (a typical location for pauses, especially in L2 writing, see Section 3.2) and another one for the beginning of the multiword unit, which gives writers two good reasons to insert a pause before the n-gram (hence, presumably, the frequent use of the PX_ pattern).

The n-grams with a PPP score lower than 5 include *us to* and *able to* (which have low positive scores of 4.76 and 2.38, respectively), *that we* (which has a score of 0), *of the* and *to the* (which have negative scores of −9.6 and −16.07, respectively). As appears from Table 3, all of these n-grams are predominantly used with no adjacent pauses (_X_), but they are all sometimes interrupted by a pause (XPX), with percentages ranging between about 5% (for *able to*) and 25% (for *to the*), which is the highest percentage of XPX among all twenty n-grams. What is common to these n-grams, with the exception of *able to*, is that they are syntactic chunks which are entirely made up of function words. As illustrated by examples (13) to (16), they are typically embedded within longer stretches of language. In (13), the n-gram *of the* is part of the phrase *most of the time*. In (14), *to the* is embedded in *thanks to the law itself*. In (15), *us to* is embedded in *help us to remember*, with *help* and *us* being initially fused together (*helpus*). In (16), which illustrates the n-gram *that we*, *that* is part of the subordinating conjunction *given that* while *we* is the subject of *can*, which makes the pause between *that* and *we* quite fitting.

(13)   th{2000}ey·regret·them·mod[BACK]st·**of·the**·time[RSHIFT].{2048}

(14)   {3736}th[BACK][BACK]thn[BACK]anks·**to·the**·law·itself{3772}

(15)   {3120}helpus[BACK][BACK]·**us·to**·remeber·{14808}

(16)   Given·**that**·{6896}**we**·can·{6512}speak·about·changing·the·world·{8384}

The low PPP score of these n-grams could be due to the fact that it is the larger chunks within which they are embedded that are processed holistically. This can be illustrated by means of *able to*, which has a low score of 2.38 although it may seem like a good candidate for holistic processing. A closer look at the process data reveals that the infinitive complement (or the infinitive complement clause) appears to be closely linked to *able to*. In example (17), the infinitive clause, which is underlined, follows *able to* within the same burst. In fact, if we take the infinitive complement (or clause) into account to calculate the PPP score, it rises from 2.38 to 21.43. If we add the verb *BE* before, considering a sequence like *are able to do it* in (18) as a unit, the PPP score reaches almost 30. It therefore seems as if a better candidate for holistic processing than *able to* might be [(*BE*) *able to* + Inf], a partially abstract schema in which the last slot is variable.

(17)   {3616}**ab**k[BACK]**le·to**·<u>think·outside·the·bow</u>[BACK]<u>x</u>{2240}

(18)   {11600}<u>are</u>·**able·to**·<u>do·it</u>[LSHIFT]?·{3040}

The same sort of explanation could apply to some of the n-grams that have an intermediate PPP score, lower than 50 but higher than 5. Thus, *such as*, *a lot of*, *the most*, *the same* and *a good* also tend to be followed by a complement within the same burst (underlined in the examples): the adjective *powerful* after *the most* in (19), the noun *politicians* after *a lot of* in (20) and the noun *teacher* after *a good* in (21), for example. This may indicate that these n-grams are processed together with their complements and are stored in the form of lexically filled constructions with an abstract element: [*the most* + Adj], [*a lot of* + N] and [*a good* + N].

(19)   {4881}the·best·job·are·{8031}**the·most**·<u>powerful</u>·{2832}

(20)   {5504}**a·lot·of**·<u>politicians</u>·{2808}

(21)   {2096}[LSHIFT]**A·good**·<u>teacher</u>·{2640}

Among the n-grams with an intermediate PPP score, one may be surprised to find *for example* with a score of 32.35, while *for instance*, a similar phrase, has a much higher score of 63.89. This is all the more surprising since *for example* is more frequent than *for instance* in L1 English (*for example* has a relative frequency of 241.39 per million words in the British National Corpus and *for instance* a relative frequency of 73.94 per million words), and it is also closer to the French equivalent *par exemple*. We would therefore expect *for example* to be more entrenched than *for instance* among French-speaking learners, and hence to show more signs of holistic processing. A possible explanation for this unexpected result is that the potentially better entrenchment of *for example* may ease the combination of multiple chunks within a single burst, as illustrated in examples (22) to (24): in (22), *for example* is combined with a noun phrase, *political ideas*, and in (23) and (24), it is inserted within a full clause (*presidents use their power* in (23) and *internet is a great source of inspiration for many people* in (24)).

(22)   {5456}political·ideas,·**for·example**,·{7232}

(23)   {13616}[CAPS LOCK]**F**[CAPS LOCK]**or·example**,·presidents·use·their·power·{17584}

(24)   i{2116}nternet·[BACK],·**for**,·e[BACK][BACK]·[BACK][BACK]·x[BACK]**example**·is·a·great·sours[BACK]ce·of·
inspiration·for·a[BACK][BACK][BACK]r·a[BACK]many·pep[BACK]ol[BACK]ple·to[BACK][BACK]{2120}

In comparison with other n-grams, the PPP score of *on the other hand* may also seem surprisingly low (13.64), and the percentage of instances interrupted by a pause (XPX) may seem particularly high (18.18%). There are at least two possible explanations for these results, which are not mutually exclusive. The first one is that *on the other hand* is the longest of all the n-grams under study, with 4 words and 14 characters (17 if spaces are included). This leaves more opportunities for pausing and may also be quite long for certain learners to produce in one go. In speech, for example, it has been shown that the average number of words per fluent unit is about six for fluent L1 speakers (Pawley and Syder 2000, p. 195). Another possible explanation is that learners may actually not be so familiar with this chunk, as appears from the many errors that are visible in the process data, with some learners misspelling it (25) and others producing phrases such as *on the first hand* (26) or *in the other hand* (27).

(25)   **On·the·o**<u>h</u>[BACK]**ther·hand**
(26)   **on·the**·<u>first</u>[BACK][BACK][BACK][BACK][BACK]**otherhand**[BACK]
(27)   <u>I</u>n·**the·other·hand**,·[Movement][LEFT Click][LEFT Click][Movement][LEFT Click][Movement][BACK][LSHIFT]**O**
[LEFT Click][Movement]

Carey (2013, pp. 221–22) describes the "approximate forms" of *on the other hand* in a corpus of English as a lingua franca, including, among others, the form *on the other side*. Although the PROCEED sample does not include any instances of *on the other side*, it is a frequent error among French-speaking learners of English,[2] one which might be due to a literal translation of the French equivalent *d'un autre côté* (where *côté* means *side*). The competition of this non-standard form could make the retrieval of the form *on the other hand* even more difficult. It is also interesting to note that *on the other hand* is one of three n-grams (together with *in my opinion* and *I think*) that most often disappears through writing processes: only 55% of its occurrences are kept in the final texts, which might indicate that learners are unsure about its use.

6. Discussion

The literature arguing for the relevance of pauses in establishing the holistic processing of multiword units puts forward two criteria, namely the absence of interrupting pauses and the presence of delineating pauses (see Sections 2.3 and 3.2). The first criterion applies in a majority of cases for the twenty n-grams under study: on average, this represents 93.31% of the data analyzed (XPX accounts for the remaining 6.69%). Moreover, almost half of the n-grams are never interrupted by a pause, and among the other n-grams, interruption never occurs in more than a quarter of the instances. Delineating pauses on both sides of the n-gram (PXP), which correspond to the second criterion, are found in less than 10% of the data on average (8.92%). However, some specific n-grams predominantly take this form, namely *in conclusion* (45.45%), *in fact* (50%) and *first of all* (64.29%). In addition, partial delineation (on one side only of the n-gram: PX_ and _XP) could be said to reinforce this trend, with its overall percentage of 25% and its predominance among some n-grams (61.11% in total for *for instance*, 66.66% for *in my opinion* and 70% for *to sum up*). As for the absence of delineating pauses (_X_), which represents the most frequent pause placement pattern in the data overall (58.74%), it does not provide evidence for holistic processing according to the literature, but it does not really provide evidence against it either. All in all, the data thus reveal a few cases which could suggest a lack of holistic processing (XPX), but they also reveal many cases which seem to confirm, or at least do not contradict, the hypothesis of holistic processing. This is reflected in the PPP score, which is supposed to be indicative of the extent to which a multiword unit could be processed holistically. The PPP score displays positive values except for three n-grams (with one 0 and two negative values) and has an average value close to 27.

Despite these encouraging results, it should be emphasized that, strictly speaking, what the data reveal is some link between n-grams and pause placement, and a tendency for certain n-grams to be produced in one go. This seems to confirm the processing advantage of multiword units (Section 2.2) since they can only be produced fast (i.e., with no interrupting pauses) if they are processed fast. Claims found in the literature (see Sections 2.3 and 3.2) allow us to assume from this processing advantage that (certain) n-grams are processed and stored holistically. However, not everybody agrees that the speed of processing can be equated with holistic processing/storage (cf. Siyanova-Chanturia 2015 and references therein). In other words, because an n-gram is produced in one go does not necessarily mean that it is activated and retrieved as a whole from the mental lexicon. In the present study, it seems quite plausible that an n-gram with a high PPP score such as *in fact* is stored as a whole rather than pieced together from its components, whereas an n-gram with a low PPP score such as *to the* involves the retrieval of two separate elements. Yet, one should be careful not to jump too quickly from the observation of what could be called 'holistic production' to claims about holistic storage.

Moreover, the analysis of the individual n-grams has brought to light factors that seem to account for some of the differences in pause placement and that could point to various effects working in conjunction with processing effects. One of them is the structure of the n-gram. N-grams with a high PPP score tend to be structurally complete (e.g., *first of all*, *in my opinion*), whereas n-grams with a low PPP score tend to be structurally incomplete (e.g., *of the*, *able to*). If the slot of the complement is included in the computation of the PPP score, so as to make the n-gram structurally complete (e.g., [*of the* + N], [(*BE*) *able to* + Inf]), the PPP score of the whole sequence tends to increase, signaling a higher likelihood of being produced as a burst. It should be added that all n-grams that are predominantly found in a pattern with a pause on at least one side (PX_, _XP, PXP) are structurally complete and that all structurally incomplete n-grams are predominantly found in the _X_ pattern, that is, with no delineating pauses. These results are in line with the observation that pauses tend to occur at syntactic boundaries (Section 3.2). They could also be explained by the "one-clause-at-a-time hypothesis" (Pawley and Syder 2000), according to which "speakers plan the lexical content of novel

utterances in chunks no larger than one independent clause at a time" (Pawley and Syder 2000, p. 163), which results in the presence of disfluencies "at or near the boundaries of clauses" (Pawley and Syder 2000, p. 164). As predicted by this hypothesis, n-grams with a high PPP score (*in fact*, *first of all*, *in conclusion*, etc.) are often found in sentence-initial position, a position that could favor the use of a pause before the n-gram. In addition to a processing effect, the presence of delineating pauses could therefore be an effect of syntactic boundaries.

Related to the structurally complete or incomplete nature of the n-grams is their composition. N-grams with a high PPP score are not only structurally complete, but they also tend to include lexical words (e.g., *to sum up*, *for instance*). N-grams with a low PPP score, on the other hand, are not only structurally incomplete, but they also tend to be exclusively made up of function words (e.g., *us to*, *that we*). Biber and Conrad (1999, pp. 183–84) underline the lack of salience of n-grams, which they explain by the predominance of function words in n-grams and the fact that function words typically go unnoticed. If an n-gram like *us to* is less noticeable and less easily recognizable as a chunk than one like *for instance*, it is also less likely to be processed as a whole. This is especially true for EFL learners, for whom teaching effects might be at work too. In particular, structurally complete n-grams including lexical words may be brought to their attention and may even be part of vocabulary lists which learners are required to memorize. Structurally incomplete n-grams made up of function words, by contrast, are unlikely to be explicitly taught in the classroom. Given their low salience, they are also unlikely to be picked up by learners from the (limited) language input that they are exposed to.

Another factor that seems relevant to pause placement is writers' degree of familiarity with the n-gram. While all n-grams investigated here have been selected because they are frequent in the data analyzed, those that have a very high PPP score (*in fact*, *first of all*, etc.) have been shown to be generally overused by learners, to the extent that they could be described as "phraseological teddy bears" (Hasselgård 2019), being highly familiar to learners and always providing them with a safe option to rely on. Conversely, a phrase such as *on the other hand*, which has a lower PPP score (13.64) and a relatively high proportion of interruption by pauses (18.18%), appears to be less familiar to learners, as testified by the errors and revisions found in the process data and as predicted by the possible competition with the non-standard form *on the other side*. The contrast between *for instance* (PPP score: 63.89) and *for example* (PPP score: 32.35) suggests that there might be an intermediate level of familiarity that favors the production of a multiword unit as a burst. If the multiword unit is not familiar to the writer, it is expected to be produced word by word (XPX). If it is familiar to the writer, it is expected to be produced in one go (PXP). If it is very familiar to the writer and fully automatized, it may be so easy for the writer to produce that it is likely to be embedded within a larger chunk of language (_X_). Learners' knowledge of formulaic language might therefore have an impact on their capacity to process multiword units as wholes.

Formulaicity may be involved in other, less direct ways, based on findings from research on speech fluency. Several scholars, starting with Pawley and Syder (1983), have argued that the use of formulaic sequences in speech enhances fluency. The reasoning behind this is that by using ready-made expressions, one can free up cognitive resources that can be exploited to better plan longer stretches of language, which results in a higher degree of fluency. The lack of interruption of n-grams, therefore, is not necessarily (and not exclusively) a sign that the n-gram is stored holistically, but it could be a consequence of the effective use of n-grams. Put differently, learners who rely more on multiword units might be more fluent (generally, not just when producing n-grams), and this may not say anything about how multiword units are stored.[3] Formulaic language and fluency are also related to each other through their common association with language proficiency: more proficient learners tend to be more fluent (see, e.g., Götz 2019 on filled pauses) and they also tend to make better use of multiword units (see, e.g., Garner et al. 2019). It might be, therefore, that (certain)[4] n-grams tend to be found in texts produced by more proficient learners and that these texts include fewer pauses because the writers are more fluent. Such explanations based on the link between formulaic language and fluency, however, can only account for the lack of pauses within n-grams; they do not predict the presence of pauses at the boundaries of n-grams.

The length of the n-gram could also have some impact on pause placement, and hence on the PPP score: longer n-grams entail more opportunities for pausing (see Rajtar 2016 for speech), which should lead to a lower PPP score. However, this impact turns out to be quite marginal in the data studied here. The n-grams with a PPP score higher than 50 include as many bigrams as trigrams, and the n-grams with the lowest PPP score (under 5) are all made up of two words only. It is possible that n-grams need to be longer than three words to cause a decrease in the PPP score, as is the case for *on the other hand*. The average burst length in the PROCEED sample under study amounts to 21 characters, which means that the learners can produce up to four or five words on average without pausing. With its 17 characters (spaces included), *on the other hand* may be just about manageable for some learners (if they are familiar with the chunk—see above), but longer n-grams may be incompatible with a single-burst production.

The fact that the study has been conducted on learner language brings its own potential effects too, including teaching effects and language proficiency effects, as noted above. The heavy cognitive load that is placed on writers using an L2 (see, e.g., Schoonen et al. 2009) also means that fewer cognitive resources might be available for certain aspects of text production, possibly including the fast retrieval and fluent production of multiword units. This effect is expected to be even stronger for argumentative writing, the type of writing found in PROCEED, as it has been shown to be cognitively more demanding than other genres such as narrative texts (e.g., Matsuhashi 1981). Possible transfer effects cannot be excluded either. It has been demonstrated that learners tend to transfer writing skills and language features from their L1 (e.g., Odlin 1989; Berman 1994). The existence of a congruent multiword unit in learners' L1 could have a facilitative effect on the storage, retrieval and production of the corresponding n-gram in L2. This could explain the high PPP score of *in fact*, which has a direct (and frequent) equivalent in French, namely *en fait*. By contrast, n-grams with no direct equivalent in learners' L1 could be less likely to be stored holistically and more difficult to produce fluently, as suggested earlier for *on the other hand*, which has no literal translation in French. Transfer could also have an impact on pause placement patterns. In the same way as L2 speech prosody can show traces of transfer (e.g., Ueyama 2012), it could be that the 'writing prosody' of learners' L1, including where pauses typically fall, influences the prosody of their L2 writing.

With respect to the process/product interface and the link between processing and frequency, this study suggests that n-grams, which are selected based on their recurrence, are good candidates for fluent production and possible holistic processing. However, it also shows that not all n-grams are equally likely to be produced fluently and processed holistically. Interestingly, higher frequency does not seem to lead to enhanced fluency and hence a better chance of holistic processing. Actually, the eight most frequent n-grams in the PROCEED sample (*of the*, *the most*, *the same*, *to the*, *a good*, *that we*, *us to* and *able to*) correspond to the n-grams with the lowest PPP scores. The n-grams that are less frequent, on the other hand, tend to have high PPP scores. *In my opinion*, for example, is the second least frequent n-gram out of the twenty under study, but it has a high PPP score of 60. This could confirm the role played by other factors such as those described above (structure of the n-gram, composition, length, degree of familiarity, etc.). More generally, these findings support the view that corpus-derived quantitative measures should be interpreted cautiously in matters of mental processing (see Section 2.2).

7. Conclusions

This study of pause placement around n-grams in writing process data from the Process Corpus of English in Education (PROCEED) has not provided strong evidence for across-the-board holistic processing of multiword units. However, by relying on the Pause Placement and Processing (PPP) score, it has highlighted certain n-grams whose holistic processing seems plausible (e.g., *in fact*, *first of all*), as well as n-grams which one would not expect to be stored holistically (e.g., *to the*, *that we*). This suggests that the use of pauses as indicators of holistic processing, initially applied to speech, might be transposable to writing through the study of keylogging data.

These results should be interpreted with two main caveats in mind. The first one is that although only the most frequent n-grams have been chosen for further investigation, some of the pause placement patterns correspond to a small number of instances. More data should be analyzed to confirm the findings obtained in this study. The second caveat is that, as illustrated in Section 6, the noticeable effects are not necessarily (only) processing effects. They could be fluency effects or effects of grammatical boundaries, for example. It could also be that some of the pauses considered in this analysis are not indicators of holistic processing, since pauses can have many different functions (Olive 2010). In order to go beyond the observation of holistic production and confirm the existence of holistic processing, it would therefore be necessary to examine other types of data and seek converging evidence.

In future research, the duration of pauses should be given more attention. Here, all pauses above the threshold of 2000 ms have been treated in the same way. However, using another threshold would most probably yield different results. The impact of this choice could be determined by comparing different thresholds, as Van Waes and Leijten (2015) do for the study of writing fluency. The actual duration of pauses could also be taken into account in the study of holistic processing. Although Schilperoord (2001, p. 76) makes it clear that pausing time does not equal processing time, he points out that "longer pauses reflect cognitive processes that are *relatively* more effortful compared to processes reflected by shorter pauses" (Schilperoord 2001, p. 77; emphasis original). A longer pause before an n-gram could thus indicate that the n-gram is more difficult to retrieve from the 'phrasicon' (i.e., the mental inventory of phraseological units), perhaps because it is less familiar to the writer, or it could even indicate that the n-gram has not been retrieved from the phrasicon (meaning that it is not stored as a phraseological unit), but has been pieced together on the spot. The duration of pauses should also ideally be interpreted in relation to each individual's average pausing time and typing speed (see Wengelin 2006).

In addition, it would be desirable to compare the present results with similar results from L1 English data, so as to distinguish what is specific to L2 data from what applies to writing process data in general. A comparison with the L1 French data included in PROCEED would also be interesting, as it might reveal similarities in the L1 and L2 production of n-grams, and hence point to possible cases of transfer. The comparison of individual writers would be useful too, as not all writers may process multiword units in the same way (see Howarth 1998). Other avenues for research include the possible link between pause placement patterns and association measures (here, multiword units have exclusively been selected on the basis of their frequency), the comparison of PPP scores for n-grams and non-n-grams (here, only n-grams have been investigated) and the application of this methodology to other types of phraseological phenomena (here, only multiword units of the n-gram type have been analyzed). The processing of formulaic language, as a cognitive phenomenon, potentially involves many influencing factors. This is what makes its study challenging, but also fascinating.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The PROCEED data are not publicly available at this stage. The data extracted from the corpus and presented in this study are available on request from the author.

Conflicts of Interest: The author declares no conflicts of interest.

Notes

1. *For instance* shows equal proportions of PXP and _XP.
2. Several examples can be found in the French component of the International Corpus of Learner English, e.g., *Answering 'yes' would be unfair to the people or organizations which spend their time struggling for the protection of nature, **on the other side** answering 'no' would be a rather naïve reaction.*
3. Pawley and Syder (1983) still suggest that these formulaic sequences are retrieved as wholes, which indicates that the two explanations are not mutually exclusive.

4. The presence of n-grams, as such, is not a marker of proficiency. It is only certain n-grams (those that are used accurately, those that are typical of the register at hand, etc.) that have been shown to correlate with proficiency.

References

Abdel Latif, Muhammad M. Mahmoud. 2013. What do we mean by writing fluency and how can it be validly measured? *Applied Linguistics* 34: 99–105.

Anthony, Laurence. 2017. *VariAnt (Version 1.1.0) [Computer Software]*. Tokyo: Waseda University. Available online: https://www.laurence-anthony.net/software (accessed on 4 January 2024).

Baaijen, Veerle M., David Galbraith, and Kees de Glopper. 2012. Keystroke analysis: Reflections on procedures and measures. *Written Communication* 29: 246–77.

Bailey, Jim, and OBS Studio Contributors. 2012. OBS Studio. Available online: https://obsproject.com (accessed on 4 January 2024).

Berman, Robert. 1994. Learners' transfer of writing skills between languages. *TESL Canada Journal* 12: 29–46.

Biber, Douglas, and Susan Conrad. 1999. Lexical bundles in conversation and academic prose. In *Out of Corpora: Studies in Honour of Stig Johansson*. Edited by Hilde Hasselgård and Signe Oksefjell. Amsterdam: Rodopi, pp. 181–90.

Biber, Douglas, Susan Conrad, and Viviana Cortes. 2004. *If you look at...*: Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25: 371–405.

Bloom, Melanie. 2008. Second language composition in independent settings: Supporting the writing process with cognitive strategies. In *Language Learning Strategies in Independent Settings*. Edited by Stella Hurd and Tim Lewis. Bristol: Multilingual Matters, pp. 103–18.

Boomer, Donald S. 1965. Hesitation and grammatical encoding. *Language and Speech* 8: 148–58.

Breuer, Esther Odilia. 2017. Revision processes in first language and foreign language writing: Differences and similarities in the success of revision processes. *Journal of Academic Writing* 7: 27–42.

Breuer, Esther Odilia. 2019. Fluency in L1 and FL writing: An analysis of planning, essay writing and final revision. In *Observing Writing: Insights from Keystroke Logging and Handwriting*. Edited by Eva Lindgren and Kirk P. H. Sullivan. Leiden: Brill, pp. 190–211.

Brown, Gillian, Karen L. Currie, and Joanne Kenworthy. 1980. *Questions of Intonation*. London: Croom Helm.

Carey, Ray. 2013. On the other side: Formulaic organizing chunks in spoken and written academic ELF. *Journal of English as a Lingua Franca* 2: 207–28.

Chenoweth, N. Ann, and John R. Hayes. 2001. Fluency in writing: Generating text in L1 and L2. *Written Communication* 18: 80–98.

Chenoweth, N. Ann, and John R. Hayes. 2003. The inner voice in writing. *Written Communication* 20: 99–118.

Cieślicka, Anna. 2006. Literal salience in on-line processing of idiomatic expressions by second language learners. *Second Language Research* 22: 115–44.

Cislaru, Georgeta, ed. 2015. *Writing(s) at the Crossroads: The Process/Product Interface*. Amsterdam: John Benjamins.

Cislaru, Georgeta, and Thierry Olive. 2017. Segments répétés, jets textuels et autres routines. Quel niveau de pré-construction ? *Corpus* 17: 1–21.

Cislaru, Georgeta, and Thierry Olive. 2018. *Le processus de textualisation. Analyse des unités linguistiques de performance écrite*. Louvain-la-Neuve: De Boeck Supérieur.

Conklin, Kathy, and Norbert Schmitt. 2008. Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics* 29: 72–89.

Conklin, Kathy, and Norbert Schmitt. 2012. The processing of formulaic language. *Annual Review of Applied Linguistics* 32: 45–61.

Dahlmann, Irina, and Svenja Adolphs. 2007. Pauses as an indicator of psycholinguistically valid multi-word expressions (MWEs)? In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*. Prague: Association for Computational Linguistics, pp. 49–56.

Dahlmann, Irina, and Svenja Adolphs. 2009. Spoken corpus analysis: Multimodal approaches to language description. In *Contemporary Corpus Linguistics*. Edited by Paul Baker. London: Continuum, pp. 125–39.

Durrant, Philip, and Julie Mathews-Aydınlı. 2011. A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes* 30: 58–72.

Ellis, Nick C., and Rita Simpson-Vlach. 2009. Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory* 5: 61–78.

Ellis, Nick C., Rita Simpson-Vlach, and Carson Maynard. 2008. Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly* 42: 375–96.

Erman, Britt. 2007. Cognitive processes as evidence of the idiom principle. *International Journal of Corpus Linguistics* 12: 25–53.

Flinn, Andrea. 2023. How often do pauses occur in lexical bundles in spoken native English speech? *Corpus Pragmatics* 7: 303–22.

Flower, Linda, and John R. Hayes. 1981a. A cognitive process theory of writing. *College Composition and Communication* 32: 365–87.

Flower, Linda, and John R. Hayes. 1981b. The pregnant pause: An inquiry into the nature of planning. *Research in the Teaching of English* 15: 229–43.

Gánem-Gutiérrez, Gabriela Adela, and Alexander Gilmore. 2018. Tracking the real-time evolution of a writing event: Second language writers at different proficiency levels. *Language Learning* 68: 469–506.

Garner, James, Scott Crossley, and Kristopher Kyle. 2019. N-gram measures and L2 writing proficiency. *System* 80: 176–87.

Gilquin, Gaëtanelle. 2022. The *Process Corpus of English in Education*: Going beyond the written text. *Research in Corpus Linguistics* 10: 31–44.

Gilquin, Gaëtanelle, and Sylviane Granger. 2015. Learner language. In *The Cambridge Handbook of English Corpus Linguistics*. Edited by Douglas Biber and Randi Reppen. Cambridge: Cambridge University Press, pp. 418–35.

Gilquin, Gaëtanelle, Sylviane Granger, and Magali Paquot. 2007. Improve your writing skills (Writing sections). In *Macmillan English Dictionary for Advanced Learners, Second Edition*. Edited by Michael Rundell. Oxford: Macmillan Education, pp. IW1–IW29.

Goldberg, Adele E. 2003. Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences* 7: 219–24.

Götz, Sandra. 2019. Filled pauses across proficiency levels, L1s and learning context variables: A multivariate exploration of the *Trinity Lancaster Corpus Sample*. *International Journal of Learner Corpus Research* 5: 159–80.

Grabe, William, and Robert B. Kaplan. 1996. *Theory and Practice of Writing: An Applied Linguistic Perspective*. London: Routledge.

Graves, Donald H. 1975. An examination of the writing processes of seven year old children. *Research in the Teaching of English* 9: 227–41.

Guinet, Eric, and Sonia Kandel. 2010. Ductus: A software package for the study of handwriting production. *Behavior Research Methods* 42: 326–32.

Hairston, Maxine. 1982. The winds of change: Thomas Kuhn and the revolution in the teaching of writing. *College Composition and Communication* 33: 76–88.

Hasselgård, Hilde. 2019. Phraseological teddy bears: Frequent lexical bundles in academic writing by Norwegian learners and native speakers of English. In *Corpus Linguistics, Context and Culture*. Edited by Viola Wiegand and Michaela Mahlberg. Berlin: De Gruyter, pp. 339–62.

Hayes, John R., and N. Ann Chenoweth. 2006. Is working memory involved in the transcribing and editing of texts? *Written Communication* 23: 135–49.

Hayes, John R., and Linda S. Flower. 1983. Uncovering cognitive processes in writing: An introduction to protocol analysis. In *Research On Writing: Principles and Methods*. Edited by Peter Mosenthal, Lynne Tamor and Sean A. Walmsley. New York: Longman, pp. 207–20.

Howarth, Peter. 1998. Phraseology and second language proficiency. *Applied Linguistics* 19: 24–44.

Jiang, Nan, and Tatiana M. Nekrasova. 2007. The processing of formulaic sequences by second language speakers. *The Modern Language Journal* 91: 433–45.

Kaufer, David S., John R. Hayes, and Linda Flower. 1986. Composing written sentences. *Research in the Teaching of English* 20: 121–40.

Kim, Soo Hyon, and Ji Hyon Kim. 2012. Frequency effects in L2 multiword unit processing: Evidence from self-paced reading. *TESOL Quarterly* 46: 831–41.

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar. Vol. I. Theoretical Prerequisites*. Stanford: Stanford University Press.

Leijten, Mariëlle, and Luuk Van Waes. 2013. Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication* 30: 358–92.

Lin, Phoebe M. S. 2018. *The Prosody of Formulaic Sequences: A Corpus and Discourse Approach*. London: Bloomsbury.

Lindgren, Eva, and Kirk P. H. Sullivan. 2019. *Observing Writing: Insights from Keystroke Logging and Handwriting*. Leiden: Brill.

Matsuhashi, Ann. 1981. Pausing and planning: The tempo of written discourse production. *Research in the Teaching of English* 15: 113–34.

Medimorec, Srdan, and Evan F. Risko. 2017. Pauses in written composition: On the importance of where writers pause. *Reading and Writing* 30: 1267–85.

Odlin, Terence. 1989. *Language Transfer: Cross-linguistic Influence in Language Learning*. Cambridge: Cambridge University Press.

Olive, Thierry. 2010. Methods, techniques, and tools for the on-line study of the writing process. In *Writing: Processes, Tools and Techniques*. Edited by Nathan L. Mertens. New York: Nova Science Publishers, pp. 1–18.

Olive, Thierry, and Georgeta Cislaru. 2015. Linguistic forms at the process-product interface: Analysing the linguistic content of bursts of production. In *Writing(s) at the Crossroads: The Process/Product Interface*. Edited by Georgeta Cislaru. Amsterdam: John Benjamins, pp. 99–123.

Olynyk, Marian, Alison d'Anglejan, and David Sankoff. 1987. A quantitative and qualitative analysis of speech markers in the native and second language speech of bilinguals. *Applied Psycholinguistics* 8: 121–36.

Palmer, Harold E. 1933. *Second Interim Report on English Collocations*. Tokyo: Kaitakusha.

Paquot, Magali. 2008. Exemplification in learner writing: A cross-linguistic perspective. In *Phraseology in Foreign Language Learning and Teaching*. Edited by Fanny Meunier and Sylviane Granger. Amsterdam: John Benjamins, pp. 101–19.

Pawley, Andrew. 1986. Lexicalization. In *Language and Linguistics: The Interdependence of Theory, Data, and Application*. Edited by Deborah Tannen and James E. Alatis. Washington, DC: Georgetown University Press, pp. 98–120.

Pawley, Andrew, and Frances Hodgetts Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In *Language and Communication*. Edited by Jack C. Richards and Richard W. Schmidt. London: Longman, pp. 191–226.

Pawley, Andrew, and Frances Hodgetts Syder. 2000. The one-clause-at-a-time hypothesis. In *Perspectives on Fluency*. Edited by Heidi Riggenbach. Ann Arbor: The University of Michigan Press, pp. 163–99.

Peters, Ann M. 1983. *The Units of Language Acquisition*. Cambridge: Cambridge University Press.

Pianko, Sharon. 1979. A description of the composing processes of college freshman writers. *Research in the Teaching of English* 13: 5–22.

Prior, Paul. 2004. Tracing process: How texts come into being. In *What Writing Does and How It Does It: An Introduction to Analyzing Texts and Textual Practices*. Edited by Charles Bazerman and Paul Prior. Mahwah: Lawrence Erlbaum, pp. 167–200.

Rajtar, Wojciech. 2016. Formulaic language in native and learner English—A corpus-based study of silent pauses. In *Variability in English across Time and Space*. Edited by Ewa Waniek-Klimczak and Anna Cichosz. Łódź: Wydawnictwo Uniwersytetu Łódzkiego, pp. 77–91.

Raupach, Manfred. 1984. Formulae in second language speech production. In *Second Language Productions*. Edited by Hans W. Dechert, Dorothea Möhle and Manfred Raupach. Tübingen: Narr, pp. 114–37.

Révész, Andrea, and Marije Michel. 2019. Introduction. *Studies in Second Language Acquisition* 41: 491–501.

Rosenberg, Sheldon. 1977. Semantic constraints on sentence production: An experimental approach. In *Sentence Production: Developments in Research and Theory*. Edited by Sheldon Rosenberg. Hillsdale: Erlbaum, pp. 195–228.

Sabourin, Laura. 2009. Neuroimaging and research into second language acquisition. *Second Language Research* 25: 5–11.

Schilperoord, Joost. 2001. On the cognitive status of pauses in discourse production. In *Contemporary Tools and Techniques for Studying Writing*. Edited by Thierry Olive and C. Michael Levy. Dordrecht: Kluwer Academic Publishers, pp. 61–87.

Schmitt, Norbert, Sarah Grandage, and Svenja Adolphs. 2004. Are corpus-derived recurrent clusters psycholinguistically valid? In *Formulaic Sequences. Acquisition, Processing and Use*. Edited by Norbert Schmitt. Amsterdam: John Benjamins, pp. 127–51.

Schneider, Ulrike. 2014. Frequency, Chunks and Hesitations. A Usage-Based Analysis of Chunking in English. Ph.D. dissertation, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany.

Schoonen, Rob, Patrick Snellings, Marie Stevenson, and Amos van Gelderen. 2009. Towards a blueprint of the foreign language writer: The linguistic and cognitive demands of foreign language writing. In *Writing in Foreign Language Contexts: Learning, Teaching, and Research*. Edited by Rosa M. Manchón. Bristol: Multilingual Matters, pp. 77–101.

Scott, Mike. 2008. *WordSmith Tools (Version 5) [Computer Software]*. Liverpool: Lexical Analysis Software.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Siyanova-Chanturia, Anna. 2013. Eye-tracking and ERPs in multi-word expression research: A state-of-the-art review of the method and findings. *The Mental Lexicon* 8: 245–68.

Siyanova-Chanturia, Anna. 2015. On the 'holistic' nature of formulaic language. *Corpus Linguistics and Linguistic Theory* 11: 285–301.

Siyanova-Chanturia, Anna, and Ron Martinez. 2015. The idiom principle revisited. *Applied Linguistics* 36: 549–69.

Siyanova-Chanturia, Anna, Kathy Conklin, and Norbert Schmitt. 2011a. Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research* 27: 251–72.

Siyanova-Chanturia, Anna, Kathy Conklin, and Walter van Heuven. 2011b. Seeing a phrase 'time and again' matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Language, Memory, and Cognition* 37: 776–84.

Sosa, Anna Vogel, and James MacFarlane. 2002. Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word *of*. *Brain and Language* 83: 227–36.

Spelman Miller, Kristyan. 2000. Academic writers on-line: Investigating pausing in the production of text. *Language Teaching Research* 4: 123–48.

Stolarek, Elizabeth A. 1994. Prose modeling and metacognition: The effect of modeling on developing a metacognitive stance toward writing. *Research in the Teaching of English* 28: 154–74.

Sullivan, Kirk P. H., and Eva Lindgren. 2006. *Computer Keystroke Logging and Writing: Methods and Applications.* Oxford: Elsevier.

Takayoshi, Pamela. 2018. Writing in social worlds: An argument for researching composing processes. *College Composition and Communication* 69: 550–80.

Tremblay, Antoine, and Harald Baayen. 2010. Holistic processing of regular four-word sequences: A behavioural and ERP study of the effects of structure, frequency, and probability on immediate free recall. In *Perspectives on Formulaic Language: Acquisition and Communication*. Edited by David Wood. London: Continuum, pp. 151–73.

Ueyama, Motoko. 2012. *Prosodic Transfer: An Acoustic Study of L2 English and L2 Japanese*. Bologna: Bologna University Press.

Van Lancker, Diana, Gerald J. Canter, and Dale Terbeek. 1981. Disambiguation of ditropic sentences: Acoustic and phonetic cues. *Journal of Speech and Hearing Research* 24: 330–35.

Van Waes, Luuk, and Mariëlle Leijten. 2015. Fluency in writing: A multidimensional perspective on writing fluency applied to L1 and L2. *Computers and Composition* 38: 79–95.

Wahl, Alexander. 2015. Intonation unit boundaries and the storage of bigrams: Evidence from bidirectional and directional association measures. *Review of Cognitive Linguistics* 13: 191–219.

Wengelin, Åsa. 2006. Examining pauses in writing: Theories, methods and empirical data. In *Computer Keystroke Logging and Writing: Methods and Applications*. Edited by Kirk P. H. Sullivan and Eva Lindgren. Oxford: Elsevier, pp. 107–30.

Wray, Alison. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.