Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Automatic detection of production-sustaining linguistic units

Quentin Feltgen[1] Gaëtanelle Gilquin[1]
(1) CECL, PLIN, ILC, Université catholique de Louvain, Louvain-la-Neuve, Belgique
quentin.feltgen@gmail.com, gaetanelle.gilquin@uclouvain.be

RÉSUMÉ

La production textuelle est segmentée par des pauses en jets textuels de longueur variable, interprétés comme manifestant une certaine cohérence cognitive dans la rédaction. Pour favoriser la fluence de ce processus, les scripteurs peuvent avoir recours à des unités linguistiques qui permettent de maintenir le flux de la production. L'objectif de cette contribution est de proposer une méthode de TAL pour détecter automatiquement ces unités. Nous l'appliquons à un corpus d'apprenants écrit en anglais L2 et montrons, d'une part, que les unités de structuration du texte (connecteurs, etc.) opèrent souvent de manière isolée, et d'autre part que la production peut être maintenue par le recours à des stratégies de complémentation (syntagme prépositionnel, proposition infinitive) qui permettent d'élaborer à partir d'un contenu déjà en place.

ABSTRACT

Automatic detection of production-sustaining linguistic units

Writing production unfolds in ebbs and flows, segmented by pauses into bursts of writing that are believed to match an underlying cognitive consistency of the writing process. To strengthen their writing fluency, writers may rely on linguistic units that sustain the production flow. This contribution aims to offer an NLP method to automatically detect these production-sustaining units. We apply it to a learner corpus of English L2 writing and show that text-structuring markers (connectors and the like) work in an isolated way. On the other hand, this process can be sustained by complementation strategies, such as the reliance on prepositional phrases or infinitive clauses, which elaborate on the existing content.

MOTS-CLÉS : données de keylogging; corpus d'apprenants; unités phraséologiques; jets textuels; fluence.

KEYWORDS: keystroke logging; learner corpus; phraseological units; bursts of writing; fluency.

Actes de l'atelier Traitement de données langagières dynamiques
par les outils et méthodes du TAL 2025 (DYN-TAL), pages 17–22.

1 Introduction

Writing production is a fundamentally heterogeneous process, interspersed with pauses that hint at ongoing cognitive processes (Schilperoord, 2002). However, these disfluencies in the writing process may be alleviated to some extent thanks to the recourse to linguistic devices that sustain the production and prolong the burst of writing between two pauses (Cislaru & Olive, 2018). For example, the intraphrasal occurrence of the conjunction *et* ('and') in French has been shown to associate with longer bursts (Feltgen *et al.*, 2022). So far, however, there has been no systematic investigation of which linguistic units may play such a role. Our contribution is intended to fill this gap.

2 Materials & Methods

2.1 Corpus data

The keylogging data has been recorded with the Inputlog software (Leijten & Van Waes, 2013) as part of the PROCEED corpus (Gilquin, 2022). The corpus is made up of essays in English from French-speaking students in English Linguistics and Literature, which they had to write in 45 minutes based on a thought-provoking quote. In total, we recorded 501 texts, totaling nearly 150,000 words.

2.2 Bursts extraction

To extract the bursts from each individual text, we proceeded in two steps. First, we identified 'production sequences', that is, sequences of keyboard events locally contributing to the elaboration of the text. These sequences correspond to an uninterrupted succession of either production events (letters, spacing, punctuation) or revision events (deletion of characters immediately adjacent to the current insertion point cursor position). All move events (e.g. using the keyboard arrows) and all click events end such a sequence. These sequences are then further segmented into so-called bursts of writing (Alves *et al.*, 2008) by using a pause-defining threshold: every time the inter-key interval between two keyboard events is longer than the threshold, we consider it a pause. All produced material (including deletion events) between two pauses or between a pause and the border of a production sequence is considered a burst. The threshold has been fixed to 2s, in accordance with the reference value that prevails in the literature (Wengelin, 2006; Barkaoui, 2019).

2.3 Preprocessing

The bursts are minimally cleaned through a number of steps. First, all caps-production events (pressing of the shift key or the caps-lock key) are removed from the bursts. Then, the within-word revision events are accounted for in the following way: if the sequence of deleted characters is entirely made up of actual letters (no spacing, no punctuation), both the deleted characters and the revision marks are removed from the bursts (e.g. `re«drem«eam` was replaced with `dream`).

Words are formed by merging uninterrupted sequences of alphabetic characters (no spacing, no punctuation, no apostrophe). We then define n-grams as a succession of n words in the sequence of words within a burst, irrespective of the events interspersed between them (spacing, punctuation, etc.).

2.4 Long bursts identification

We then partitioned the corpus to single out longer bursts, indicative of a sustained textual production. The length of the bursts is defined here based on the number of words, following the operative definition of words outlined in the preceding section. Alternative definitions of the burst length could, for instance, rely on the number of events within each burst.

As displayed in Figure 1, even though bursts of a greater length (e.g. 7 words or more) correspond to a small percentage of all bursts (about 10%), their contribution to the corpus size remains important (roughly 50%). Here we decided to only keep the upper 5% of bursts.

One important methodological decision is to choose between applying that 5% threshold at the group level or the individual level. Both would make sense: applying it at the group level emphasizes that participants who may rely the most on burst-sustaining language units would be able to produce longer bursts and therefore contribute more to the overall 'long bursts' sub-corpus. On the other hand, we may regard the use of these linguistic devices as a generally available tool allowing each individual to produce longer bursts compared to what they usually do, while the greater fluency of some participants would stem from a higher writing skill overall. We opt for the latter in the following. With this choice, the upper 5% of bursts correspond to 39% of the total number of words.

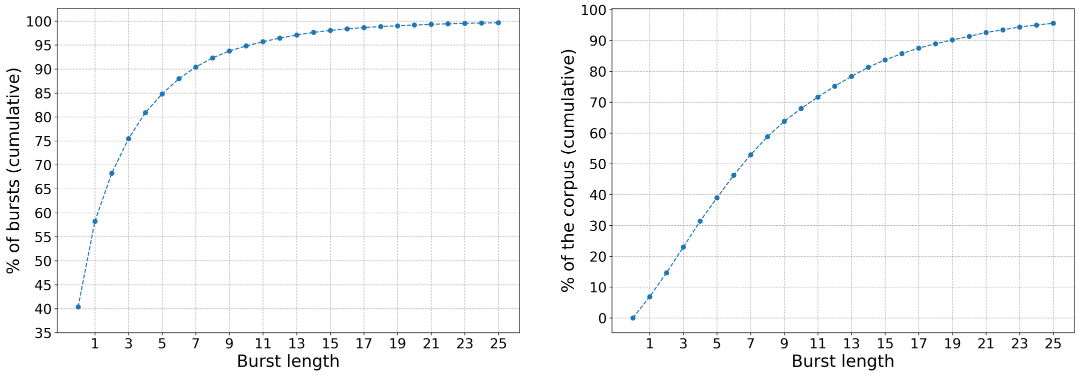

Figure 1: Cumulative sum of the number of bursts of a given length and of their corresponding contribution to the total corpus size (in number of words).

2.5 Statistical test

At this point, we considered, for each possible n-gram in our corpus, how many tokens fell in the reduced 'long bursts' sub-corpus. Indeed, if a linguistic unit helps sustain text production, then it should preferentially be found in the sub-corpus of longer bursts. To provide a statistically meaningful score quantifying this attraction, we relied on the hypergeometric distribution (Lafon, 1980). The hypergeometric distribution indicates the probability of finding k tokens of a form in a subset of size N of the corpus, knowing that in the full corpus of size M, there are n tokens of that form. This probability distribution therefore gives an exact p-value for our observations, under the null hypothesis that the linguistic units are not sensitive to the bursts' length.

Noting $F(k)$ the cumulative distribution of the hypergeometric distribution, we then selected the forms that were significantly frequent in the long bursts sub-corpus ($F(k) > 0.975$, meaning that there is a 97.5% chance to have fewer tokens than observed in the sub-corpus of longer bursts) and the forms that were significantly infrequent in that sub-corpus ($F(k) < 0.025$: there is less than 2.5% chance to observe fewer tokens of that form in the sub-corpus of longer bursts). We performed this analysis for all the 3-grams separately. The 2-grams are indeed too fragmented to capture phraseological patterns, while the 4-grams are seldom repeated. Furthermore, we restricted ourselves to 3-grams with a frequency greater than or equal to 10 in the whole corpus.

3 Results

3.1 The issue of quotes

One major confounder in our study was the typing of the quote at the beginning of the student's essay. Although this could be partly remedied by discarding the first few bursts (quotes may also reoccur later in the text), it leads to an interesting observation, in that not all the quotes are processed equally easily. Some are consistently copied within a few, long bursts (e.g. *Education is the most powerful weapon which you can use to change the world.*), so that nearly all the n-grams that compose them are found among the ones most attracted to longer bursts, while other quotes are not consistently copied in large chunks (e.g. *I always lie in the present, the future I don't know, the past I no longer have*), mostly because a long sentence is probably easier to chunk than several shorter ones.

In the following, we removed from the results all n-grams belonging to these quotes. We thus discarded 104 out of the 147 selected 3-grams. We also performed an alternative analysis by first discarding the opening ten bursts of each text to exclude the quotes, leading to largely similar results.

3.2 Linguistic units attracted to the shorter bursts

The following 3-grams are significantly infrequent in the long bursts sub-corpus and therefore more typical of shorter bursts (they are sorted by order of decreasing significance): *first of all*, *in my opinion*, *on the other*, *the other hand*, *more and more*, *some of them*, *and so on*, *as a conclusion*, *I believe that*, *is not a*, *all around the*, *I agree with*, *in my view*, *we have to*, *our everyday life*, *the lack of*, *I think that*, *this is not*, *this essay will*, *the age of*, *agree with the*.

Among these n-grams, we find numerous text structuring markers, falling into the category of what Dostie (2004) calls "textual connectors" or what Traugott (2022) refers to as "1DSM", largely monofunctional Discourse Structuring Markers, to which we also attach marks of epistemic stance. These may often be produced in isolation. 3-grams like *this essay will* are to be understood in the same way. Other 3-grams that do not serve a clear discursive function, like *all around the* (which is complemented by either *world* or *globe*), *is not a*, *the age of*, *the lack of*, *our everyday life*, and *some of them*, are more opaque. The 3-gram *is not a*, for instance, is seldom found in isolation, but often ends bursts, signaling an intention to introduce a nuance/a reversal of view, without necessarily a clear idea of what should come next. Similarly, *the age of* and *some of them* are often burst-final, suggesting that the writer wants to add some precision, but the information is not readily available.

3.3 Linguistic units attracted to the longer bursts

The 3-grams significantly attracted to the longer bursts are the following, sorted by decreasing order of significance: *to be able*, *who we are*, *wouldn t be*, *will always be*, *they are not*, *to learn from*, *it is in*, *not have the*, *they don t*, *it s not*, *we have been*, *what we have*, *seen as a*, *to be in*, *people who have*, *can lead to*, *to make the*, *be able to*, *to deal with*, *what we want*, *to do it*, *we do not*.

This includes phraseological units introducing prepositional phrases (*seen as a*, *it is in*, *can lead to*) and infinitive clauses (*to deal with*, *to be able to*, *to be in*, *to make the*, *to do it*, and also *to learn from*, which can be complemented by *mistakes*, *the past*, or a reference to other people). A few 3-grams also serve to introduce a subordinate clause (*what we have*, *people who have*, *what we want*).

4 Discussion

We acknowledge several limitations in our work. First, the frequency threshold (10 hits) is somehow problematic in that it does not distinguish a form used very often in one specific idiolect from a form that is widespread among the writers. Second, revisions may artificially increase the bursts' length, in that they split the content of words over several successive sequences, even though we tried to partly alleviate this issue by applying the effects of within-word revisions in the pre-processing step. Third, we only considered 3-grams, but a combination of several n-grams would likely be preferable; for instance, the text structuring 2-gram *in conclusion* is also typical of shorter bursts.

Our results nonetheless point to an intriguing fact. The longer bursts are characterized by units tied to generic and 'universal truths' statements, like *we do not* (e.g. *we do not pay attention to our history lessons*), *will always be*, or *people who have*, which refers to a loosely identified group (e.g. *people who have back problems*). It seems that writers are therefore fluent when asserting general-truth statements. Conversely, when they introduce nuances (*some of them*), specifications (*the age of*), or discuss category membership (*is not a*), they tend to stop and be more disfluent.

A more general question pertains to how these units relate to phraseological 'teddy bears' (Hasselgård, 2019), to the gradual acquisition and functional diversification of lexical bundles throughout the learning process (Lenko-Szymanska, 2014; Ädel & Erman, 2012), and to the transfers from L1 in the use and frequency of these lexical bundles (Paquot, 2013). However, these studies focus on the frequencies of use of these bundles and not on the contrasted role they play with respect to fluency, which may explain why most established lexical bundles do not play a role in our analysis.

5 Conclusion

In this paper, we automatically extracted the 3-grams most characteristic of longer bursts, hypothesizing that they help sustain the writing production. We showed that text structuring markers are associated with disfluencies (they are often produced in isolation or at the end of a burst), while phraseological units introducing prepositional phrases – a device largely adopted by learners of English (Gilquin, 2018) – or infinitive clauses help sustain the writing production in allowing the possibility of elaborating on an existing theme.

References

ÄDEL A. & ERMAN B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of english: A lexical bundles approach. *English for specific purposes*, **31**(2), 81–92.

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

BARKAOUI K. (2019). What can l2 writers'pausing behavior tell us about their l2 writing processes? *Studies in Second Language Acquisition*, **41**(3), 529–554.

CISLARU G. & OLIVE T. (2018). *Le processus de textualisation: analyse des unités linguistiques de performance écrite*. De Boeck Supérieur.

DOSTIE G. (2004). *Pragmaticalisation et marqueurs discursifs*. De Boek et Duculot.

FELTGEN Q., CISLARU G. & BENZITOUN C. (2022). Étude linguistique et statistique des unités de performance écrite: Le cas de et. In *SHS Web of Conferences*, volume 138, p. 10001: EDP Sciences.

GILQUIN G. (2018). Exploring the spoken learner english constructicon: A corpus-driven approach. In R. ALONSO ALONSO, Ed., *Speaking in a second language*, p. 127–152. John Benjamins.

GILQUIN G. (2022). The process corpus of english in education: Going beyond the written text. *Research in Corpus Linguistics*, **10**(1), 31–44.

HASSELGÅRD H. (2019). Phraseological teddy bears. *Corpus linguistics, context and culture*, **15**, 339–362.

LAFON P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique*, **1**(1), 127–165.

LEIJTEN M. & VAN WAES L. (2013). Keystroke logging in writing research: Using inputlog to analyze and visualize writing processes. *Written Communication*, **30**(3), 358–392.

LENKO-SZYMANSKA A. (2014). The acquisition of formulaic language by efl learners: A cross-sectional and cross-linguistic perspective. *International Journal of Corpus Linguistics*, **19**(2), 225–251.

PAQUOT M. (2013). Lexical bundles and l1 transfer effects. *International Journal of Corpus Linguistics*, **18**(3), 391–417.

SCHILPEROORD J. (2002). On the cognitive status of pauses in discourse production. In *Contemporary tools and techniques for studying writing*, p. 61–87. Springer.

TRAUGOTT E. C. (2022). *Discourse structuring markers in English*. John Benjamins Publishing Company.

WENGELIN Å. (2006). Examining pauses in writing: Theory, methods and empirical data. In K. SULLIVAN & E. LINDGREN, Eds., *Computer key-stroke logging and writing*, p. 107–130. Brill.