Automatic processing of real-time recorded writing: pausal segmentation versus chunking

Georgeta Cislaru, Iris Eshkol-Taravella, Sarah Almeida-Barreto

Georgeta Cislaru*, Iris Eshkol-Taravella*, Sarah Almeida-Barreto**
*MoDyCo, CNRS – Université Paris Nanterre
**Université Sorbonne nouvelle

Automatic processing of real-time recorded writing: pausal segmentation versus chunking

Abstract
Thanks to keystroke-logging software, real-time recording of the writing process has become a valuable resource for psycholinguistics, linguistics and NLP, allowing for a better understanding of writing as a technology and as a socio-cognitive practice. While psycholinguistics is interested in the behavioral dimension related to cognitive functioning and linguistics seeks to understand the linguistic principles underlying writing processes, NLP approaches are confronted with a series of methodological questions related to the automatic processing of logging data. In this paper we apply a chunking tool on POS and burst-type annotated process data, based on a corpus of short texts produced by university students. Our main results show that i) burst segmentation coincides with chunking in 75% of cases; ii) some chunks and POS are more likely to attract pauses; iii) some chunks and POS are more sensitive to revision processes.

Key-words: writing process, chunking, bursts, pausal segmentation

1. Introduction

The writing process may be described as a flow of linguistic data interspersed with pauses. The aim of our study was to test the possibility of describing the writing process in terms of chunking, drawing inspiration from applications on spoken data. The corpus used comprises written data recorded in real time during the production process using Inputlog, a keystroke tracking software (Leijten and Van Waes, 2013). These recordings provide all the temporal data and writing events such as production and revisions. The writing process is structured around alternating pauses and periods of production in a proportion of around 50/50, comparable in this respect to oral production. The temporally linear alternation between pauses and production is regularly interrupted – in around 20% of the sequences - by disfluencies which take the form of revisions and mark the spatial non-linearity of the process, manifested by backtracking and modifications to the text already produced.

The complexity of the data, combining temporal and linguistic data, temporal linearity, spatial non-linearity and disfluencies, traditionally addressed within the scope of psycholinguistics, presents a series of challenges for both NLP and linguistics. The pause functions as a spontaneous segmenter of the language flow. Since periods of production result in textual sequences, the question arises of how to define and describe units segmented by pauses.

The linguistic material produced between two pauses is defined as *bursts of writing*: [pause] *a cousin who* [pause] *agrees to take her in* [pause] *the* [pause] *w* [pause] *eek* [pause] - [pause] *end* [pause] (Cislaru and Olive, 2018; Chenoweth and Hayes, 1981). Kaufer et al. (1986) found, through verbal protocols, that bursts constitute basic units of written production (see also Hayes 2009). Several studies have sought to describe and categorize the writing process, particularly from a behavioral point of view, by observing, for example, the duration of pauses according to several levels of analysis: words, groups, sentences, paragraphs (van Hell et al., 2008; Medimorec and Risko, 2017). Linguistic studies are still scarce; they mainly explore two avenues: that of a potential overlap between discursive or prefabricated routines and bursts (Cislaru and Olive, 2017; Gilquin, 2020) and that of a correspondence between grammatical constituents and bursts (Cislaru and Olive, 2018). The results obtained so far with these two approaches remain inconclusive, however: correspondences with grammatical constituents, for example, which correspond in Cislaru and Olive's (2018) terminology to saturated bursts, are estimated to represent less than 50% of productions. As various studies have shown, language performance units are not fully accountable for by traditional syntactic theories (Gee and Grosjean, 1983; Brazil, 1995 on speech; Doquet, 2011; Cislaru and Olive, 2018 on writing).

The challenge of the linguistic characterization of bursts therefore remains, and we propose to take it up by using the chunking method (Tellier et al., 2012; Eshkol-Taravella et al., 2020), which will be described below. Building on the hypothesis of the units segmenting the information flow (Chafe,

1992; Sinclair and Mauranen, 2006), we hypothesize that chunks are good candidates to describe the writing process, both as i) cognitive tools for processing information (Johnson, 1970); and as ii) units of linguistic segmentation (Abney, 1991).

Our working hypotheses are as follows:

1) Bursts of writing can be defined as linguistically relevant structures, as the result of a spontaneous segmentation of the flow of information by writers (in the tradition of Chafe, 1992);
2) The spontaneous segmentation of textual drafts could correspond to segmentation into chunks, given the proximity of the management of the process to oral production;
3) Revisions, however, are less likely to correspond to chunks, given their disfluent and fragmentary nature.

In the following sections we will briefly present the notion of chunking, the data analyzed, the methodology adopted and the results of the analyses. We conclude with a discussion of the contributions and prospects of the work.

2. Data and Methodology

The analysis corpus consists of 165 argumentative texts on societal themes such as the use of cannabis or cigarettes, produced by 83 Psychology undergraduates. The corpus was first pre-processed using a Python script to transform the .idfx Inputlog output files into a .csv spreadsheet presenting, column by column, information such as the identification of the writer (anonymized), the type of text produced, the duration of pauses, bursts, burst types, the production duration of a cycle [pause+burst], keyboard events (two were taken into account: character production and deletion), etc.

The categories considered in the present study are bursts, pauses and burst types. We distinguish three types of bursts:

- Production bursts (P), where the writer produces text in both temporal and spatial linearity, with production regularly interspersed with pauses;
- Revision bursts (R), where the writer pauses to return to the text already produced and modify it: in this case there is spatial non-linearity;
- Edge revision (RB) bursts, where the writer modifies the immediately produced segment after a pause, e.g.: *we have just seek* [pause] ~~*k*~~ → *n*.

Pauses vary in length and not all were considered relevant to the study. We set a threshold of 2 seconds, with calculation of individual variations for normalization, in order to retain only those pauses considered in the literature to reflect cognitive activity, with pauses below this threshold being assimilated to mechanical pauses such as searching for a key on the keyboard, double-tap, etc.

The processing of writing data is also a challenge for NLP, particularly in terms of the pre-processing of processual data and the annotation of non-homogeneous corpora presenting disfluencies. To analyze these data, we concatenated neighboring bursts and inserted a separating pause symbol in between. We obtained nine different configurations:

P-burst + P-burst
P-burst + R-burst
R-burst + P-burst
R-burst + R-burst
R-burst + RB-burst
RB-burst + P-burst
P-burst + RB-burst
RB-burst + R-burst
RB-burst + RB-burst

We used automatic annotation tools to obtain more precise information about the potential relationship between chunking, pausing and parts of speech. The automatic annotation for chunking was done using the online version of SEM (Segmenteur-Étiqueteur Markovien developed by Tellier et al., 2012). The following types of chunks were identified (adapted from Eshkol-Taravella et al., 2020):

adjectival chunk (AP): adjective head after the verb (*it is too pretty*);
adverbial chunk (AdP): adverb head (*perhaps*);
nominal chunk (NP): noun phrases including adjectives placed before and after the noun and

non-clitic pronouns (*your beautiful shoes*);
prepositional chunk (PP): phrase introduced by a preposition (*by far*);
verbal chunk (VP): phrases organized around a verbal head, associated with its clitics (*we hear you* – in French, *nous vous entendons*);
punctuation (SENT): typographical marks such as strong punctuation.

SEM also allows POS annotation. Based on the announced results for SEM, we should expect an f1-score ranging from 70.3 % to 87.0 %. Due to this rather low performance, we decided to use Stanza's POS annotation to obtain more precise information, especially about morphological features (feats). Stanza was created in 2020 by the Stanford NLP Group. It is a multilingual collection of tools which allow many different types of NLP annotation such as, in our case, part of speech and feats. We obtained the following results: chunking precision: 92%; chunking recall: 98%; F-measure: 95%.

After annotating using SEM and Stanza, three different files with different annotations were obtained: types of bursts/pauses, POS, and chunks. In order to align these data, we tokenized the burst/pauses annotation file with Stanza: POS (and feats) annotation for each token was then added to the tokens before and after the pause.

For the alignment with chunking, SEM and Stanza data needed to be normalized. For French, this involved changing "du" into "de le", "des" into "de les", etc. Some manual corrections were also necessary. Some errors were not easily predictable and solvable; they were counted as noise, and amounted to 9.88 % of all burst combination contexts.

3. Results

Once the three annotations were aligned, several measures were carried out. The raw results contain:

- The number of pauses for each pause type
- The number of parts of speech of each type
- The number of part-of-speech bigrams
- The number of chunks of each type
- The number of chunk bigrams
- The number of part-of-speech bigrams according to chunks
- The contexts before, after and around pauses for parts of speech
- Chunks, pauses and POS combined.

Based on these results, we calculated the relative frequency of each pattern according to its combinatory type among the nine configurations listed above: parts of speech, chunks or parts of speech in the context of chunks and the average number of characters, words and chunks before the pause.

3.1. The configuration Chunk+Pause+Chunk

The first configuration we were interested in was the one where the burst boundary corresponds to a chunk boundary. This does not mean that a burst equals a chunk, but only that pause segmentation respects chunk boundaries – one burst can include several chunks. This correspondence was attested for 75% of bursts. Table 1 summarizes the results in this category from two angles (the pause follows a chunk vs the pause precedes a chunk) by retaining the frequencies according to the types of chunks. Nominal and Prepositional chunks in particular attract pausal segmentation, either before (33% and 18% of the pauses, respectively) or after (25% and 20% of the pauses).

	Total	**Chunk + Pause**		**Pause + Chunk**	
		Chunk frequency	**Pause frequency**	**Chunk frequency**	**Pause frequency**
NP	6083	9 %	25 %	11 %	33 %
PP	4815	9 %	20 %	8 %	18 %
VN	3962	7 %	13 %	7 %	13 %
AP	595	10 %	3 %	6 %	2 %

AdP	2048	7 %	7 %	9 %	9 %

Table 1. Chunk+Pause+Chunk sequences.

We then checked pause segmentation related to Nominal and Prepositional chunks depending on the nature of the bursts and the nine types of combinations. The results are listed in Table 2 and show some sensitivity to the combinatorics. For instance, when revision processes are involved (R+RB and R+P), Nominal chunks attract a higher percentage of breaks after (in bold) and a significantly lower percentage before (R+RB, in italics). Prepositional chunks attract a lower percentage of pauses both before and after in a revision context (R+RB, in italics, but see also R+P and R+R).

	P+P	**P+R**	**P+RB**	**R+P**	**R+R**	**R+RB**	**RB+P**	**RB+R**	**RB+RB**	**Total**
NP+Pause	16%	13%	18%	**29**%	24%	**37**%	19%	14%	24%	25%
Pause+NP	26%	35%	19%	23%	19%	*7*%	28%	20%	20%	33%
PP+Pause	17%	8%	21%	*6*%	10%	*4*%	12%	10%	12%	20%
Pause+PP	14%	11%	10%	14%	*9*%	*7*%	16%	10%	11%	18%

Table 2. Chunk+Pause+Chunk sequences involving NP and PP chunks according to the nine types of burst combinations.

3.2. The configuration POS+Pause+POS

Based on the configuration described in 3.1., we then exploited the POS annotations and analyzed the morphosyntactic nature of the linguistic units (words or combinations of words) immediately preceding or following a pause (see Table 3). We wanted to check the probability that some POS are preferred boundaries, due to their frequency or to their position in a chunk. 15% of the lowest frequencies were excluded from the study (this includes Verbs, Conjunctions, Prepositions). It can be seen that Nouns (25%) and strong punctuation (14%) attract pausal segmentation after, whereas Adjectives (15%) and Definite articles (11%) attract pauses before. Whether considering pauses before or after, strong punctuation and weak punctuation are found more often than other categories in the immediate vicinity of a pause.

	Total	**POS + Pause**		**Pause + POS**	
		POS freq	**Pause freq**	**POS freq**	**Pause freq**
N	7334	10 %	25 %	3 %	7 %
DET (def)	3500	2 %	2 %	9 %	11 %
ADP	5276	3 %	6 %	8 %	15 %
ADV	2769	6 %	6 %	8 %	7 %
PRON	2427	4 %	4 %	9 %	8 %
S-Punctuation	1364	30 %	14 %	16 %	7 %
W-Punctuation	1850	11 %	7 %	12 %	8 %

Table 3. Chunk $^{(POS)}$+Pause+$^{(POS)}$ Chunk sequences: integrating POS annotation.

3.3. The case of chunks broken by a pause

Our results showed that 25% of bursts did not correspond to chunks, which means that some chunks are broken apart by the occurring pauses. We examined chunks interrupted by pauses, paying attention to the morphosyntactic nature (POS) of the boundaries before and after the pause. Nouns attracted 22% of pauses, either preceding or following them. Adjectives attracted 21% of pauses following them.

	Total	POS + Pause		Pause + POS	
		POS freq	Pause freq	POS freq	Pause freq
N	7334	2 %	22 %	2 %	22 %
DET (def)	3500	2 %	9 %	2 %	8 %
ADP	5276	3 %	21 %	1 %	7 %
ADV	2769	1 %	5 %	1 %	4 %
PRON	2427	1 %	2 %	1 %	3 %
S-Punct	1364	0 %	1 %	1 %	1 %
W-Punct	1850	0 %	1 %	1 %	3 %

Table 4. Pause boundaries (POS) in the configuration of chunks broken by a pause.

Looking at behavior by type of burst concatenation, we found that the P+R configuration favors breaking a chunk after a Noun (20%) or a strong punctuation mark (31%) and before a definite article (11%) or an adjective (16%). RB+R (14%) and R+RB (11%) configurations follow in favoring a break after a noun.

4. Discussion and conclusions

Our results show that 75% of pauses occur between two chunks, i.e. 75% of bursts correspond to chunks, and this global percentage varies little between burst types. Nouns (POS) and Nominal chunks seem to be the units that favor a higher proportion of pausal breaks than the other categories, although Prepositional chunks and Adjectives (POS) also attract pauses. Strong punctuation may constitute separate chunks segmented by pauses, which needs to be studied further.

We noticed however that these categories are handled in different ways depending on the combinatorics of bursts: thus, in a revision context (R+RB and R+P), Nominal chunks attract more pauses after the chunk than the average. Conversely, Prepositional chunks attract fewer pauses both before and after in the same combinations.

In conclusion, auto-segmentation in writing exhibits regularities that bring bursts closer to chunks. Nouns are the most salient markers in chunking and pause segmentation, but they play an ambivalent role, as they can both promote segmentation into chunks and cause breakage within chunks. Some types of chunks and POS are more sensitive to segmentation according to the type of writing operation (production, revision, immediate revision). Further analysis and predictive statistics need to be developed in order to obtain a more fine-grained picture of chunking and segmentation specificities according to chunk types, POS and burst combinatorics.

References

Abney S. 1991. Parsing by chunks. In R. Berwick, R. Abney and C. Tenny (ed.) *Principle based Parsing*. Kluwer Academic Publisher, pp. 257-278.

Brazil, D.. 1995. *A Grammar of Speech*. Oxford: Oxford University Press.

Chafe, W. (1992) 'Information flow in speaking and writing', in P. Downing, S. D. Lima and M. Noonan (ed.) *The Linguistics of Literacy*. Amsterdam – Philadelphia: John Benjamins, pp. 17-29.

Chenoweth, N.A. and Hayes, J.R. (2001) 'Fluency in Writing: generating text in L1 and L2', *Written Communication*, 18(1), pp. 80‑98.

Cislaru, G. and Olive, T. (2017) 'Segments répétés, jets textuels et autres routines. Quel niveau de pré-construction ?', *Corpus*, 17, pp. 61-89, https://doi.org/10.4000/corpus.2846

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

Doquet, C. (2011) *L'écriture débutante. Pratiques scripturales à l'école élémentaire*. Rennes: PUR.

Eshkol-Taravella I., Maarouf M., Badin F., Skrovec M. and Tellier I. 2020. 'Chunk Different Kind of Spoken Discourse: Challenges for Machine Learning', *Language Resources and Evaluation Conference*, May 2020, Marseille, France. pp.5164-5168.

Gee, J. P. and Grosjean, F. (1983) ‘Performance structures: a psycholinguistic and linguistic appraisal’, *Cognitive Psychology*, 15, pp. 411-458.
Gilquin, G. (2020) ‘In search of constructions in writing process data’, *Belgian Journal of Linguistics*, 34, pp. 99-109. DOI: 10.1075/bjl.00038.gil
Hayes, J. R. (2009) ‘Chapter 4: From Idea to Text’, in D. Myhill (ed.), *The Sage Handbook of Writing Development*. London: SAGE Publications, pp. 65-79.
van Hell, J., Verhoeven, L. and van Beijsterveldt, L. (2008) ‘Pause time patterns in writing narrative and expository texts by children and adults’, *Discourse Processes*, 45, pp. 406–427. DOI:10.1080/01638530802070080
Johnson, N. F. (1970) ‘The role of chunking and organization in the process of recall’, *Psychology of Learning and Motivation*, 4. pp. 171–247.
Kaufer, D., Hayes, J. R. and Flower, L. (1986) ‘Composing written sentences’, *Research in the Teaching of English*, 20, pp. 121-140.
Leijten, M. and Van Waes, L. (2013) ‘Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes’, *Written Communication*, 30(3), pp. 358-392. DOI : https://doi.org/10.1177/0741088313491692
Medimorec, S. and Risko, E. F. (2017) ‘Pauses in written composition: on the importance of where writers pause’, *Reading and Writing: an Interdisciplinary Journal*, 30, pp. 1267–1285. DOI: 10.1007/s11145-017-9723-7
Sinclair, J. and Mauranen, A. (2006) *Linear Unit Grammar: integrating speech and writing*. Amsterdam – Philadelphia: John Benjamins.
Tellier, I., Dupont, Y. and Courmet, A. (2012) ‘Un segmenteur-étiqueteur et un chunker pour le français’, *Traitement Automatique des Langues Naturelles (TALN 2012)*, Demo session, Grenoble.