Analysing the Dynamics of Textualisation through Writing Bursts and Pauses in French

Kehina Manseri, Ioana-Madalina Silai, Iris Eshkol-Taravella, Georgeta Cislaru

Motivation

- Part of the ANR-ProText project, focused on **textualisation** and **writing processes**.
- Understand how **cognitive processes** shape writing behaviour by examining the **real-time** unfolding of linguistic events.

Objectives

- Investigate **key factors** contributing to **pausal segmentation** in writing.
- Leverage these insights to develop **automated prediction models** for pauses in text production.
- **Main hypothesis**: pausal segmentation linked to morpho-syntactical boundaries.

Dataset

- **Experiments** focused on participants writing on various topics (e.g., student fees, smoking, pollution) with different levels of constraint.
- The corpus includes **56 files**, of which **33 fully analysed**.

Data collection, processing and linguistic annotation

Collection

1

- **Keystrokes and behaviours** were logged using **InputLog software**, generating **IDFX files**.
- Text production was segmented into **bursts triggered by significant pauses.**

```
<event type="keyboard" id="12">
   <part type="wordlog">
     <position>0</position>

<documentLength>1</documentLength>
     <replay>True</replay>
   </part>
   <part type="winlog">

<startTime>3945148</startTime>
     <endTime>3945244</endTime>
     <key>VK_L</key>
     <value>L</value>
     <keyboardstate>
       <key>VK_LSHIFT</key>
     </keyboardstate>
   </part>
 </event>
```

Extract of IDFX file

Processing

2

The resulting data, along with additional extracted information, was organized into a **CSV file** for analysis.

ID	n	burst Dur	pause Dur	total Actions	total Chars	final Chars	total Del	inner Del	pos Start	pos End	categ	char Burst
P+S4	1	6.27	319.23	27	27	27	0	0	0	27	P	La médecine traditionnelle
	2	10.84	46.73	41	39	37	2	2	27	54	P	peut être un atout dans not
	2	10.84	46.73	41	39	37	2	2	54	53	P	⌫
	2	10.84	46.73	41	39	37	2	2	53	63	P	s société
	2	10.84	46.73	41	39	37	2	2	63	62	P	⌫
	2	10.84	46.73	41	39	37	2	2	62	64	P	s
	3	2.88	46.98	21	21	21	0	0	64	85	P	du fait qu'elle soit
	4	8.57	54.76	36	36	36	0	0	85	121	P	exercée de génération en génération
	5	1.07	8.99	3	2	2	1	0	121	120	ER	⌫
	5	1.07	8.99	3	2	2	1	0	120	122	ER	,
	6	3.76	32.46	20	20	20	0	0	122	142	P	selon les cultures.
	7	2.47	61.14	13	11	9	2	2	142	153	P	Cela dit qi
	7	2.47	61.14	13	11	9	2	2	153	152	P	⌫
	7	2.47	61.14	13	11	9	2	2	152	151	P	⌫
	8	4.88	38.38	27	27	27	0	0	151	178	ER	la médecine traditionnelle
	9	4.67	30.38	26	26	26	0	0	178	204	P	peut freiner la recherche
	11	7.66	47.55	37	28	19	9	9	203	212	R	de nouve

The non-linear nature of the text meant that bursts were **not spatially adjacent** but could be **temporally connected**.

Annotation

3

Two-level annotation

<u>Text Reconstruction</u>
Pauses and Behaviours

```
|La médecine traditionnelle
|peut être un atout dans
no~s société~s |du fait
qu'elle soit |exercée de
génération en génération~,
|selon les cultures<~>.
|Cela dit ~|la médecine
traditionnelle |peut
<~>{être un frein dans}<~>
la recherche{, }{|dans
l'évolution}~{| d}~~{|e
nouveaux traitements par
exemple.  }
```

<u>Chunking</u>
Pauses, Behaviours, Chunks

```
"P+S4.txt":
[["|","PAUSE"],
["La médecine
traditionnelle","NP"],
["|","PAUSE"],
["peut","VN"],
["être","VN"],
["un atout","NP"],
["dans no~s société~s","PP"],
["|","PAUSE"],
```

Actions are represented by specific symbols (~, <>, |, {})

Results

- **43%** overlap between pause and chunk boundaries (though further testing is needed to identify the significance of this result).
- Some chunks show **greater discrepancy** in the frequency of pauses **before versus after them** :

  This suggests different cognitive planning processes.

Before Pause	Noun = 10.8%	Adj = 19%	Conj = 5.7%
After Pause	Noun = 16.3%	Adj = 1%	Conj = 10.1%

- The character most often added in revision bursts was **"s"**.

Predictive modelling of pauses

Dataset Enrichment

Dataset enriched with additional features:
- relative word frequency in individual texts;
- absolute occurrences in the French language (*Hermit, 2016*);
- part-of-speech tags using Spacy;
- pauses between each character typed within bursts.

Vectorisation using Word2Vec (*Mikolov et al., 2013*).

ID	n	char Burst	POS	Freq in French	Pauses	Relative freq in text
P+S4	1	La médecine traditionnelle	DET,NOUN,ADJ	6524,785	[0, 0.143, 0.032, 0.528, ...]	0.024,0.048,0.048
	2	peut être un atout dans not	VERB,AUX,DET,ADV,ADP	337435,636481,3752833,1757,1293334	[0.104, 0.096, ...]	0.048,0.048,0.048,0.024,0.073
	2	⌫			[0.104, 0.096,...]	
	2	s société			[0.104, 0.096,...]	
	2	⌫			[0.104, 0.096,...]	
	2	s			[0.104, 0.096,...]	
	3	du fait qu'elle soit	ADP,NOUN,AUX	1148060,904031,154251	[0.08, 0, 0.001,...]	0.024,0.024,0.024
	4	exercée de génération en génération	VERB,ADP,NOUN,ADP,NOUN	108,7225478,3814,2509883,3814	[0.312, 0.448, ...]	0.024,0.048,0.024,0.024,0.024
	5	⌫			[0.72, 0]	
	5	,			[0.72, 0]	
	6	selon les cultures.	ADP,DET	23610,2606014	[0.192, 0.056,...]	0.024,0.024
	7	Cela dit qi	PRON,VERB	605091	[0, 0.16, 0.152,...]	0.024,0.024
	7	⌫			[0, 0.16, 0.152,...]	
	7	⌫			[0, 0.16, 0.152,...]	
	8	la médecine traditionnelle	DET,NOUN,ADJ	4134008,6524,785	[0.024, 0.112,...]	0.048,0.048,0.048
	9	peut freiner la recherche	VERB,DET	337435,4134008	[0.121, 0.112, ...]	0.048,0.048
	11	de nouve	ADP	7225478	[0.384, 0.128, ...]	0.048

Results

- Limited **RNN** effectiveness due to the **hierarchical nature** of the data and writing variability.
- Turning the prediction problem into a **classification task** by discretizing pause lengths into **five categories** improved accuracy to **0.25** using a sequential model.
- More accurate predictions for **pauses before bursts** than for **pauses after**, suggesting pauses are influenced more by following content.
- Best results achieved **with binary classification** using **CamemBERT** *(Martin et al., 2020)* and **reconstructed text inputs**, with an accuracy of **0.82**.

Conclusion

- Ongoing work on **refining prediction models** and analysing **chunk-pause correlations**.
- Future goals: improve **predictive accuracy**, explore new strategies, and expand findings to **different writing contexts**.

References:
DUPONT Y. & PLANCQ C. (2017). Un étiqueteur en ligne du français. In *24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, p. 15–16.
HERMIT D. (2016). Frequencywords.
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.
MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE , SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* : Association for Computational Linguistics.
MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space.
SCHILPEROORD J. (2002). On the cognitive status of pauses in discourse production.