



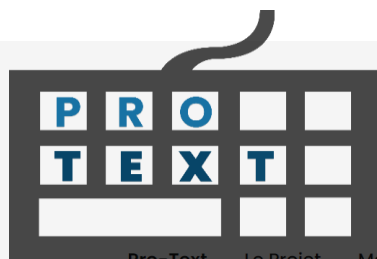
VISUALISATION FOR A 3-SIDED BEHAVIORAL/LINGUISTIC/MACHINE LEARNING ANALYSIS OF KEY-LOG DATA

Georgeta Cislaru

Nistor Grozavu

Maxime Olivié

(Pro-TEXT project)



Pro-TEXT

Les processus de Textualisation

[Pro-Text](#) [Le Projet](#) [Membres](#) [Publications](#) [Activités](#) [Annonces](#) [Ressources](#) [Documents \(privé\)](#)

Pro-Text

ANR Pro-TEXT – Les processus de textualisation: modélisations linguistiques, psycholinguistiques et d'apprentissage automatique

Processes of Textualization: Linguistic, Psycholinguistic, and Machine Learning Modeling

CONTACT

[Georgeta Cislaru](#)

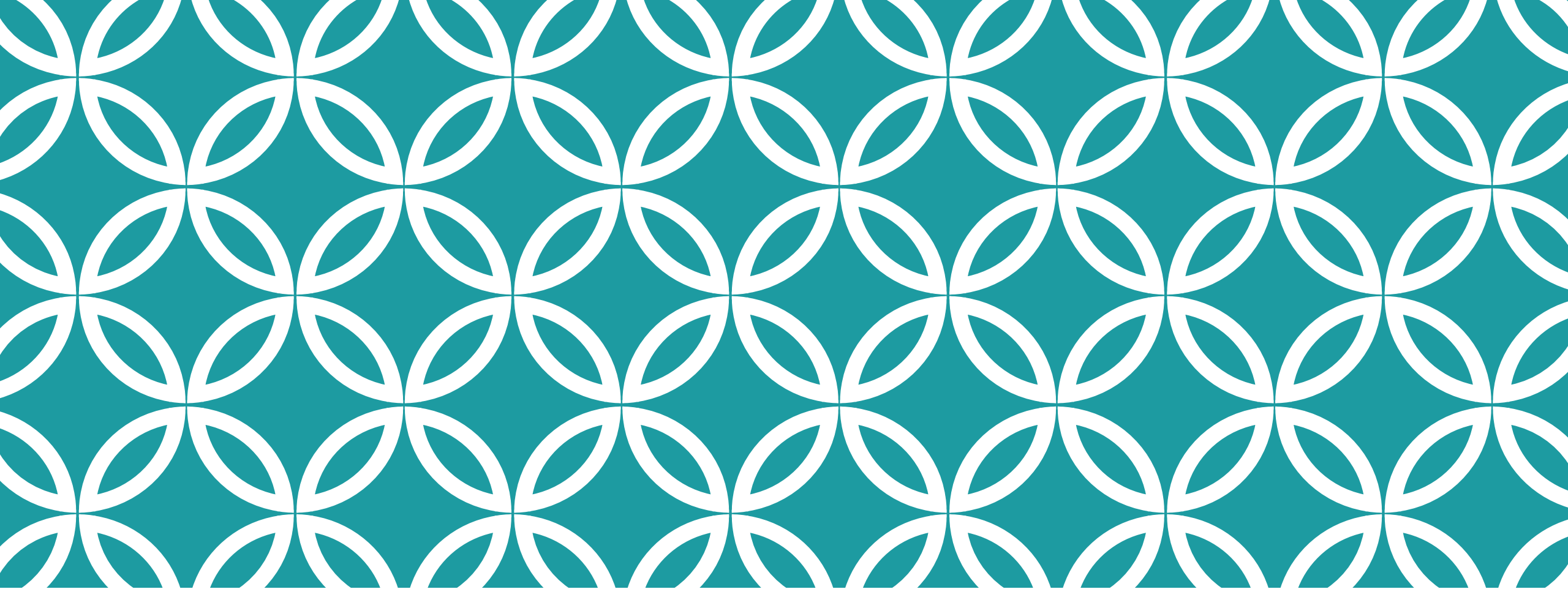
ANR



SUMMARY

- I. Writing and visualization: liminar comments
- II. 3 ways to articulate linearity and delinearity
- III. 3 ways to represent delinearity: revisions (rewriting)
- IV. Articulating hierarchical and (de)linear viewpoints
- V. Articulating and filtering linguistic and behavioral data
- VI. Machine learning conceptualizations: visualizing clusters

<https://nlp.maximeolivie.fr/visualization/null>



**WRITING AND VISUALIZATION:
LIMINAR COMMENTS**



WRITING AS AN OBJECT OF REPRESENTATION

Writing as a complex task aiming to produce a complex object: the text.

- Text as a construct of heterogeneous units.
- Rewriting as a multi-layered process.

Difficulty to accurately relate writing and rewriting to a chronology

Difficulty to grasp and represent the multiple layers

REPRESENTING THE WRITING PROCESS

Writing processes recorded through keystroke-logging software show at least two dynamic dimensions prior to the final text:

- The first is incremental, cumulative and spatially oriented (ex. from left to right and top to bottom). It leads to a constant progression of the number of characters, positions and pauses in the text.
- The second is non-linear, cumulative (insertion) or non-cumulative (deletion), spatially non-oriented.

Consequences:

- a variation of the number of characters and positions
- a redistribution of the association of a character with a position
- the number of pauses of different length is still in constant progression

All of this temporal and spatial information on the process constitutes multi-dimensional data articulating behavioral and linguistic data.

REPRESENTING THE WRITING PROCESS

The visual representation of the writing process asks for a distinct level of conceptualization, where temporal dynamics would find a static representation based on seriality and duration measures.

The point of interest for us lies in the fact that writing manipulates language, and language production is even the primary objective of the process.

But not only is the production of language structures part of the temporal dynamic, it can also be assumed to influence the course of the process.

A VIEWPOINT ON LANGUAGE

Using maps, schemas or diagrams to represent language data is common in linguistics (Legallois 2021, Mazziotta et al. 2023); they are about language families or dialects, semantic layers and relations, syntactic structure and relations, discourse structure.

From the standpoint of linguistics, visualizing writing and the writing process involves identifying the relevant linguistic forms and relationships to be highlighted. From the perspective of the study of writing, the identification of these linguistic objects is intimately linked to the dynamics of the process.

VISUALIZATION: DEFINITION, AIMS, ISSUES

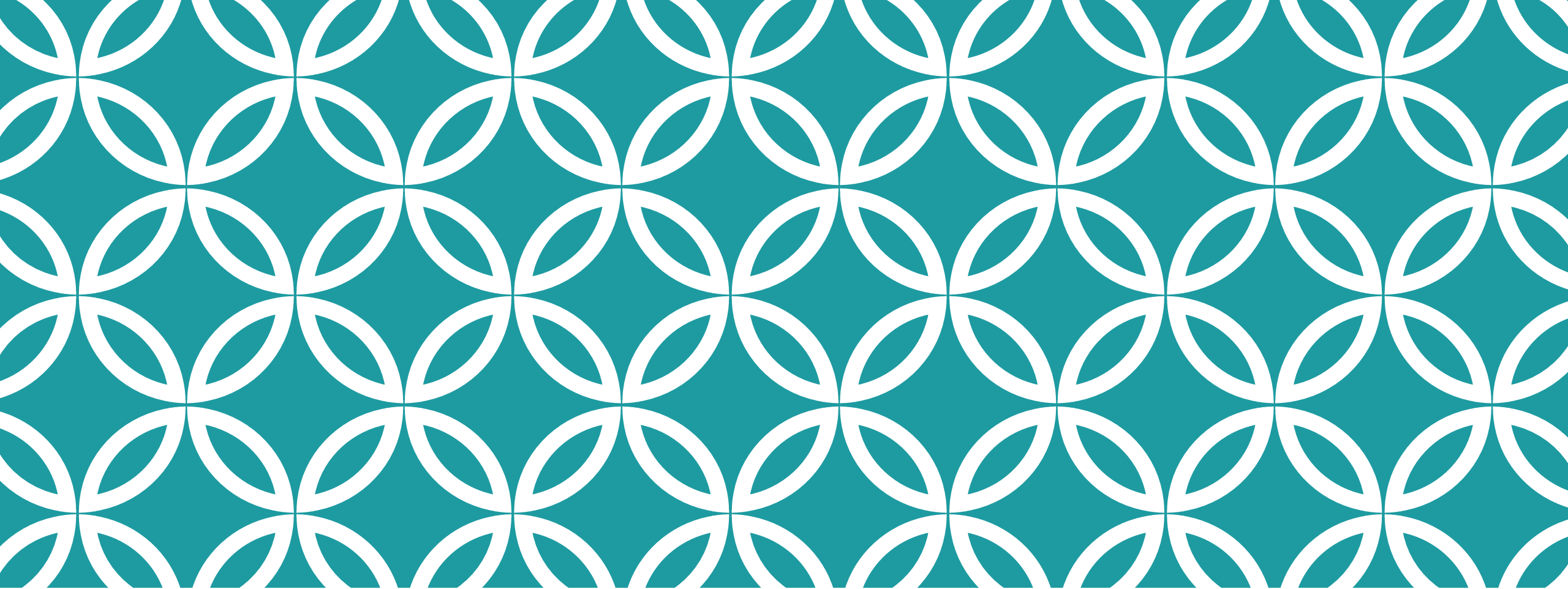
In the context of writing, ***visualization can be defined as a type of conceptualization which amounts to knowledge construction and representation.***

A conceptualization of the conceptualization: what for?

- To represent the writing process, or some specific operations?
- To focus on behavioral aspects, temporal dynamics, or language units?

What is the final end of the visual conceptualization?

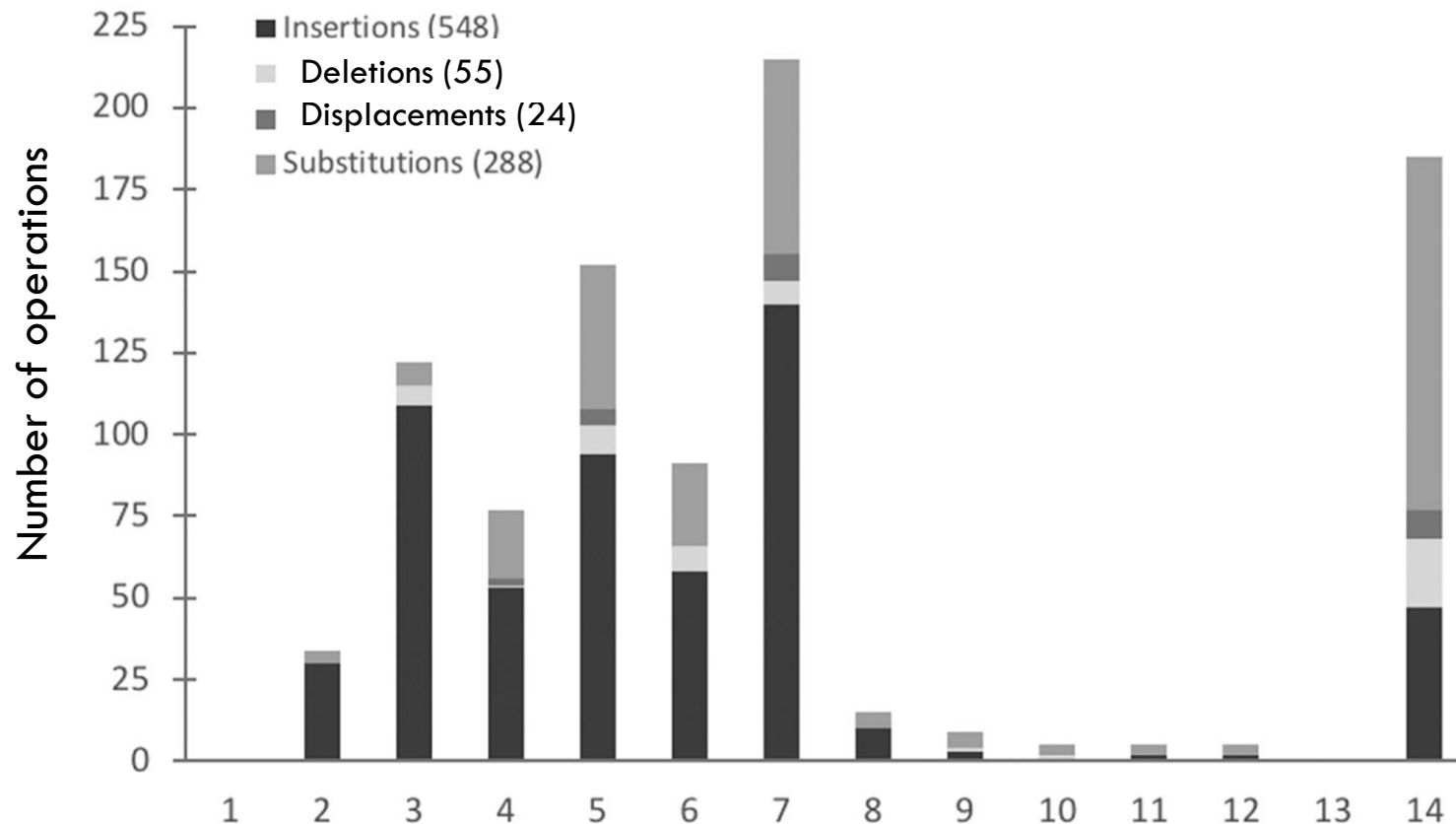
- To help writers to structure their process, to better understand writers' strategies?
- To analyze the writing process, its mechanisms and dynamics?



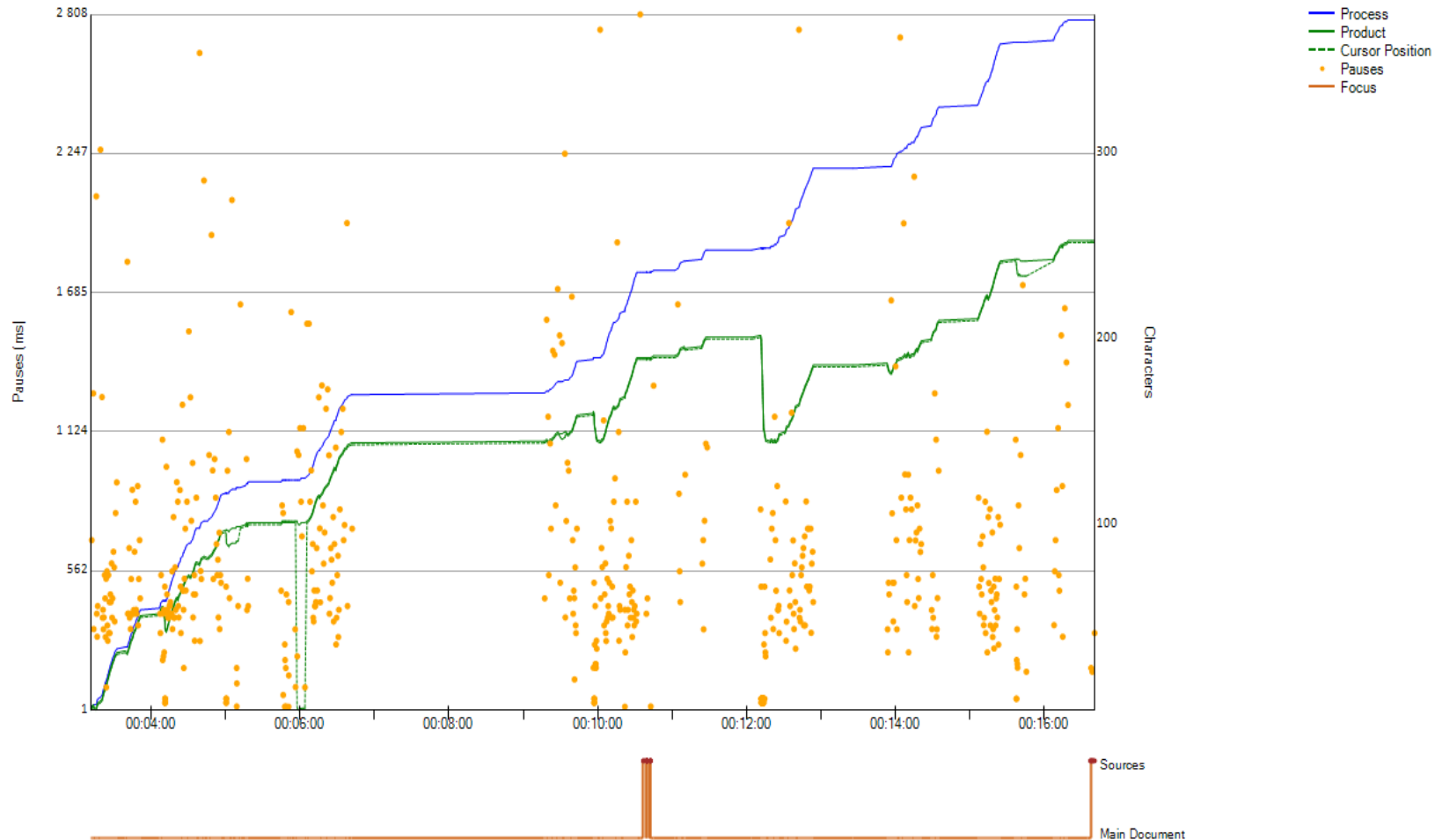
3 WAYS TO ARTICULATE LINEARITY AND DELINEARITY



REWRITING OPERATIONS FOR A 14-SESSION WRITING PROCESS (ALLONGOS TOOL, LARDILLEUX ET AL. 2013)



KEYLOG-DATA (INPUTLOT): A QUASI-LINEAR REPRESENTATION



CASCADING REPRESENTATION COMPILING BURSTS AND PAUSES

C'était [pause 2.371]

It was

en décem [pause 3.183]

in decem

bre [pause 6.099]

ber

je connaître ce garçon depuis la marenelle il e [pause 3.729]

I know [error] this guy since pre-school he w

tait tr [pause 2.044]

ere [tr]

CASCADING REPRESENTATION FOR A PARADIGMATIC VIEW

1

Rappelons

2

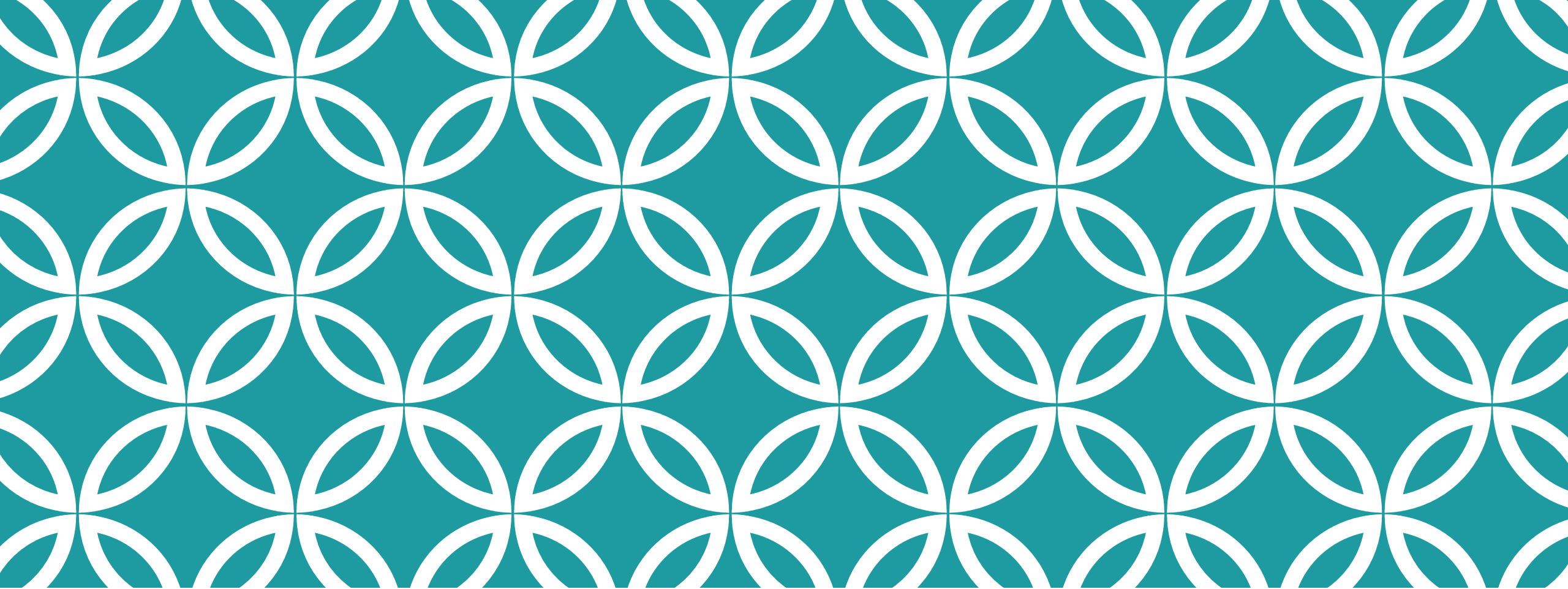
simplement **la**

3

la citation

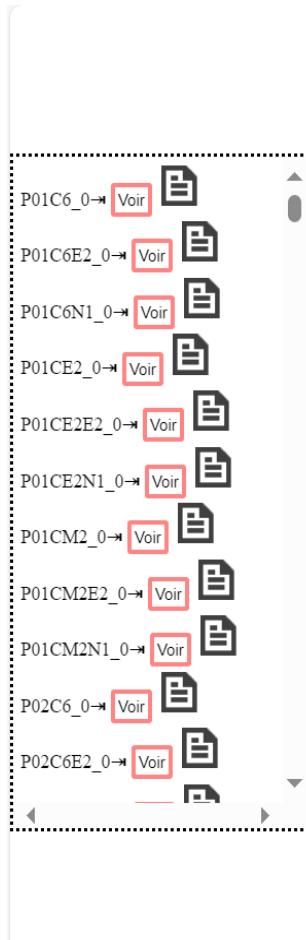
4

a définition que donne Maingueneau de la fable, citée



3 WAYS TO REPRESENT DELINEARITY: REVISIONS (REWRITING)

DELETIONS AS SPATIAL DELINARITY



[wordcount:64]

```
~FOCUS::Wordlog - Word::2127214::2127214~~0fCf2127260f0~1fCf212775f2127791~2fBACKf2129554f2129632~1fBACKf2131909f2131972~0fCf2134936f0~1fCf2135435f0~2fCf2135482f0~3fCf2135513f0~4fCf2135560f0~5fCf2135606f0~6fCf2135638f0~7fCf2135684f0~8fCf2135716f0~9fCf2135767f0~10fCf2135800f0~11fCf2135840f0~12f
```

clean Texte final

C'était la dernière récréation une personne était au gool puis une personne lui demande je peu aller au gool non j'y était avant toi mais ils ton mis un but même je reste au gool non ci non ci et il se bousculer et il se taper puis un instit vu Anthoni et Isaacs se bagarer il leur demande pourquoi vous vous bagarer.

Interrompu

```
0:<cc>:1  
0:<cccccccccccc>:1  
0:<e>:1  
0:<eeee>:1  
0:<e>:1  
0:<Ca se passer>:5  
4:<sa>:2  
1:<ete>:5  
4:<:>:5  
4:<ait la dernie>:14
```


REVISION DATA VISUALIZATION

DELETED & MODIFIED SEQUENCES HTML VISUALIZATION

C'etait_en_décembre_je_connaiss^{xex} ^{xr} *a**i**s*_ce_garçon_depuis_la_marenelle_il_et_e_a
t_a_i_t_{très_gentil}^{xm}a

^{xax} ^{xix} ^{xsx} ^{**x} ^{xq} *Q*uand_nous

⊗ ⊗ | | |
⊗ ⊗ ⊗ ⊗ | ⊗ |
| | ⊗ |
| | |
↑ ↑ ↑

⊗ ⊗ | | |
↑ ↑ ↑

DATA RECOVERY

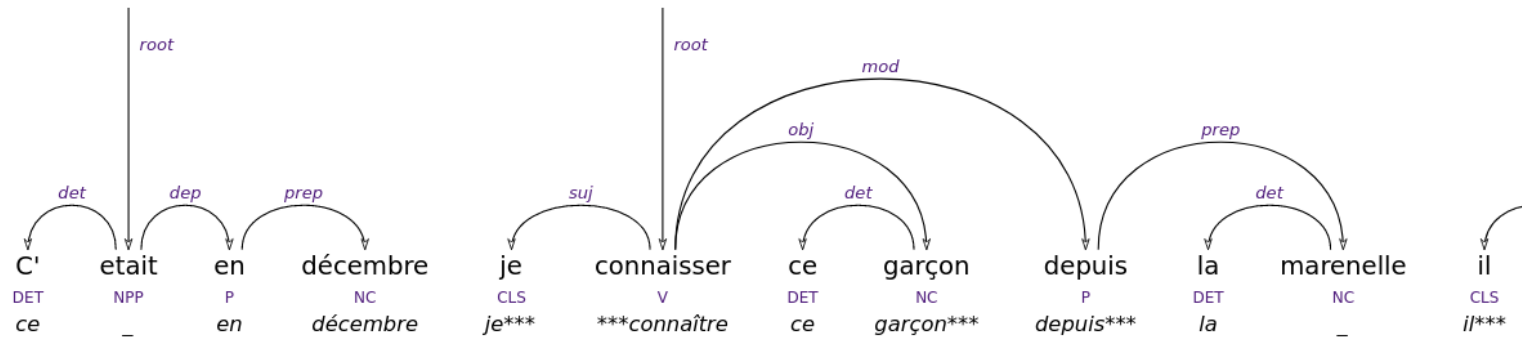
TEXT INTERMEDIATE VERSIONS (SEE ALSO MAHLOW ET AL. 2022)

```
1 <0>
2 C'etait en décembre je connaisser ce garçon depuis la marenelle il etea
3 <1>
4 C'etait en décembre je connaisser ce garçon depuis la marenelle il e
5 <2>
6 C'etait en décembre je connaisser ce garçon depuis la marenelle il etait très gen-
  til ma
7 <3>
8 C'etait en décembre je connaisser ce garçon depuis la marenelle il etait très gen-
  til m
9 <4>
10 C'etait en décembre je connaisser ce garçon depuis la marenelle il etait très gen-
  til mais quand nous sommes rentrée en sixième il trainaient avec des q
11 <5>
12 C'etait en décembre je connaisser ce garçon depuis la marenelle il etait très gen-
  til quand nous sommes rentrée en sixième il trainaient avec des q
13 <6>
14 C'etait en décembre je connaisser ce garçon depuis la marenelle il etait très gen-
  til uand nous sommes rentrée en sixième il trainaient avec des q
```




**ARTICULATING HIERARCHICAL AND
(DE)LINEAR VIEWPOINTS**

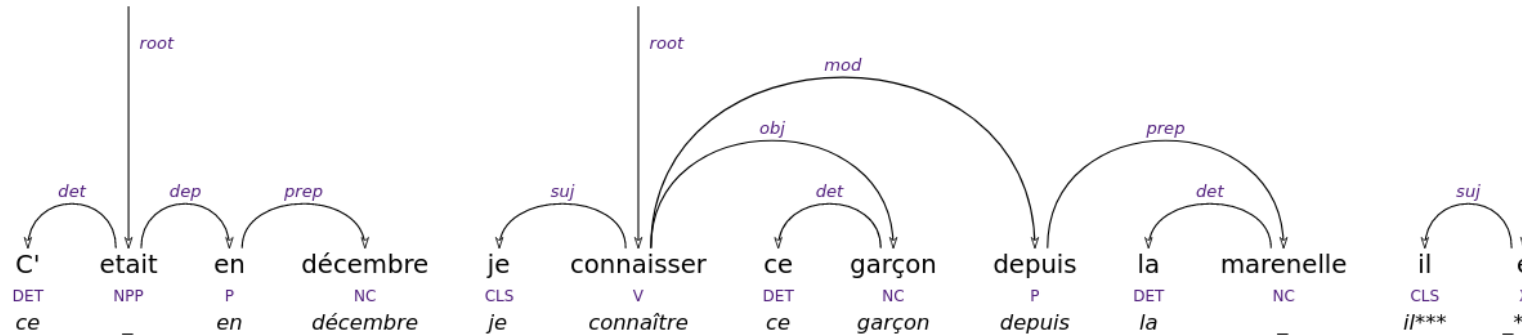




text_en

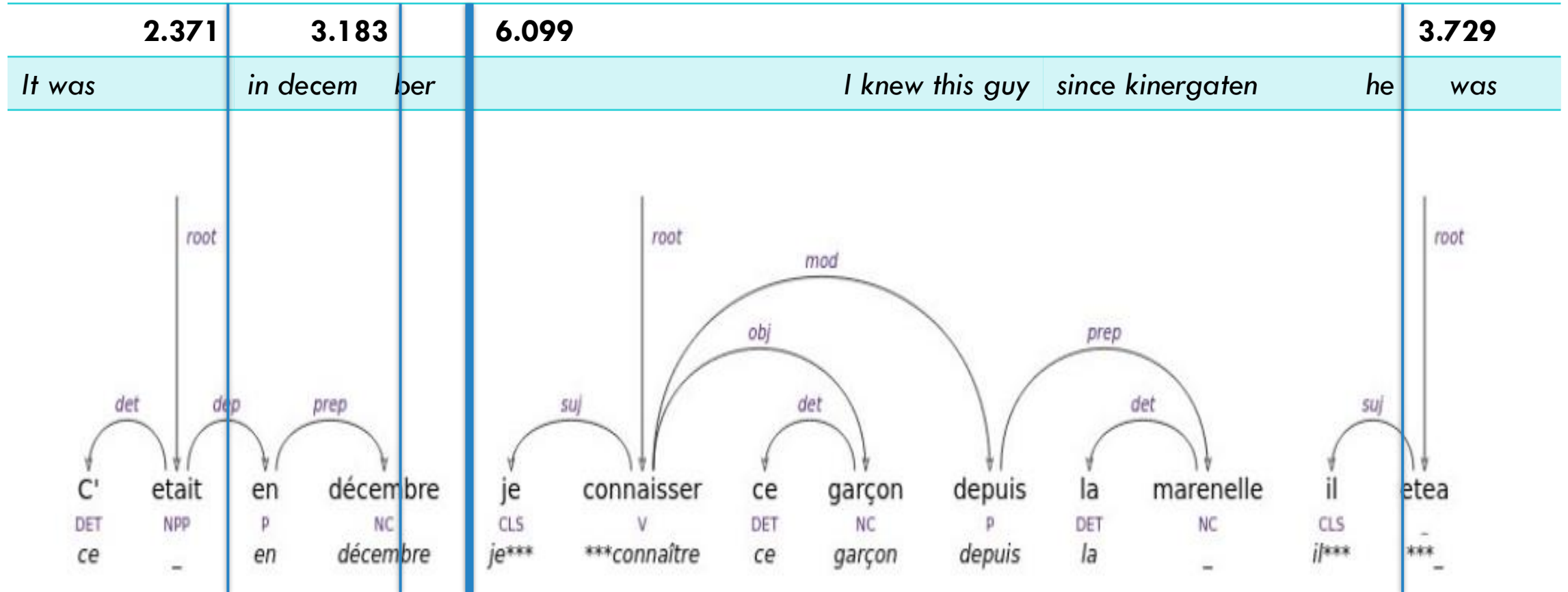
2 1647254080.840160-809515_00002 C' etait en décembre je connaisser ce garçon depuis la marenelle il e

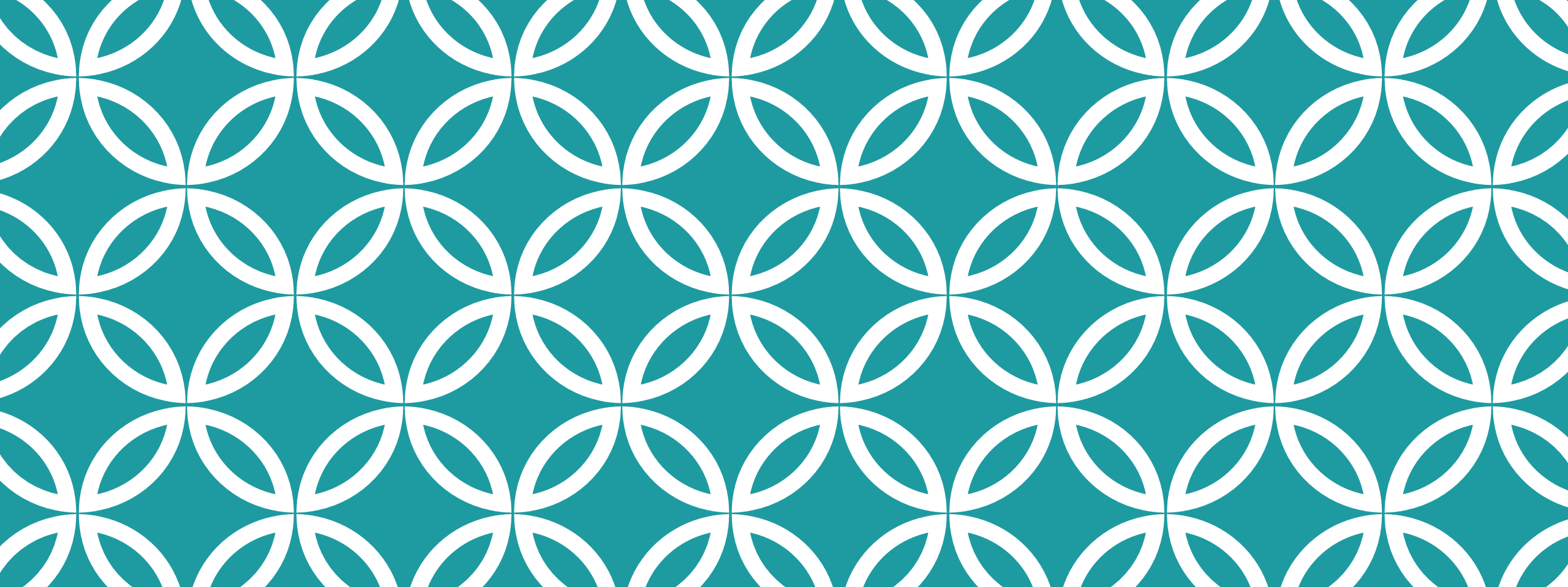
 [aleksandra.haddad](#)



POS & DEPENDENCY ANNOTATIONS

SEGMENTING DEPENDENCIES





**ARTICULATING & FILTERING
LINGUISTIC AND BEHAVIORAL DATA** |

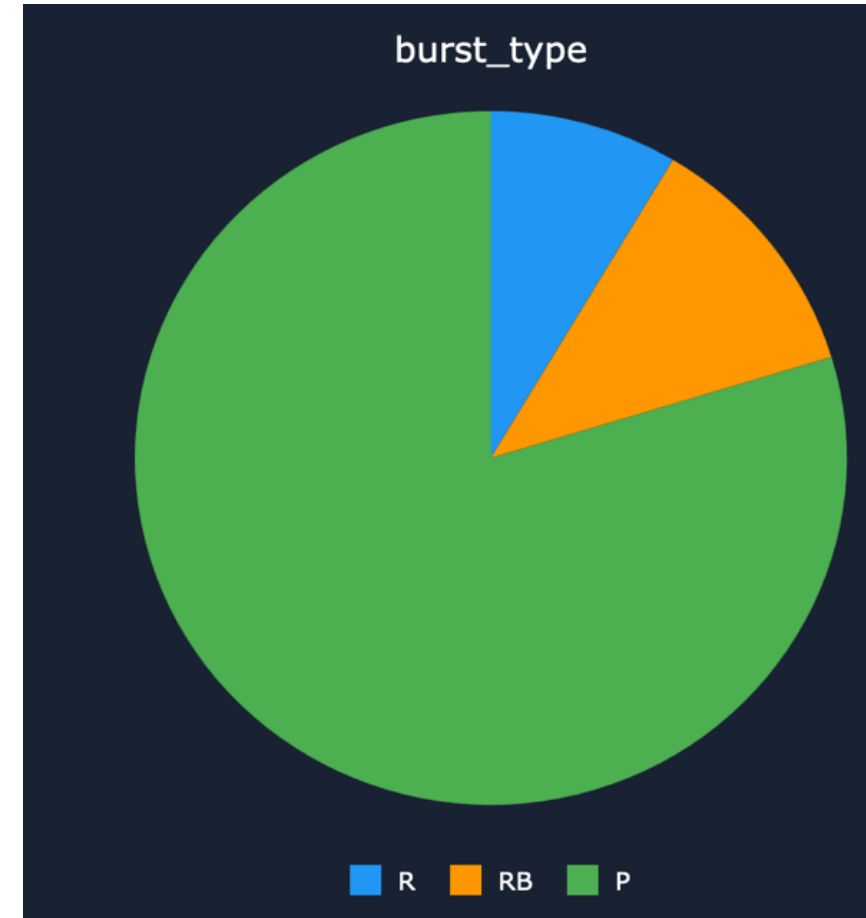
1D: PIE CHART

Display of a single QUALITATIVE attribute:

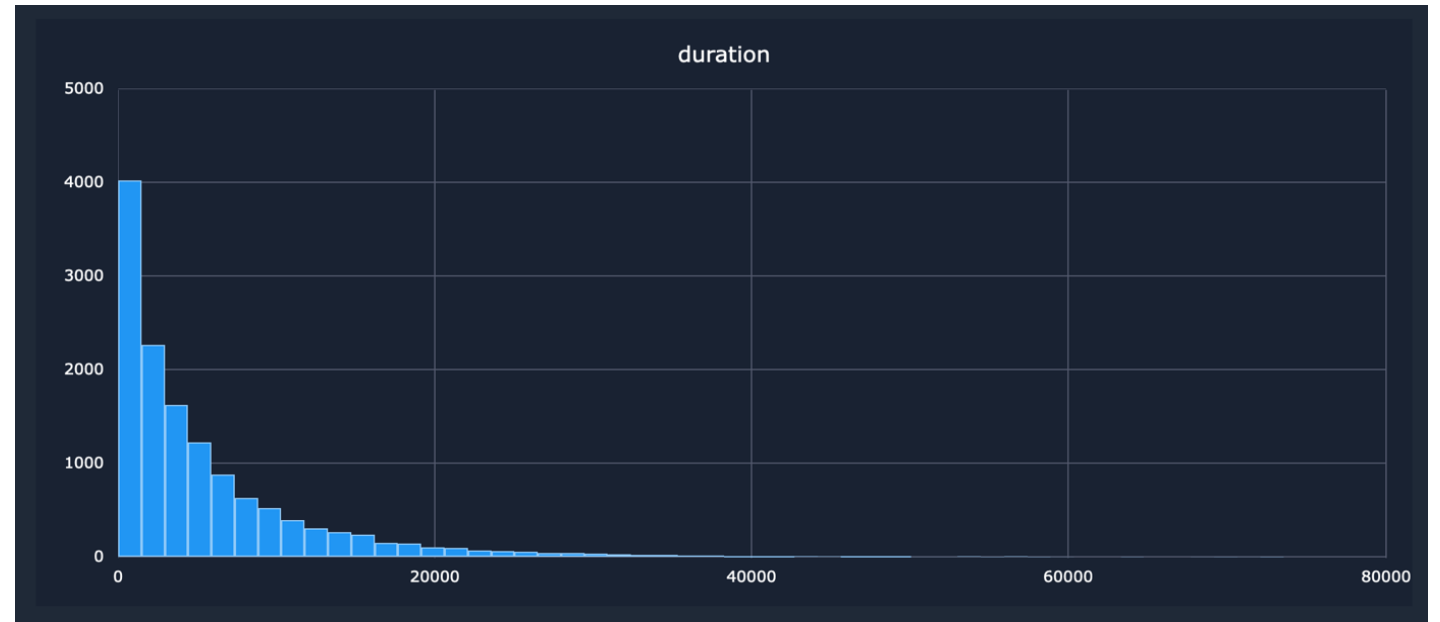
- Session, burst type, cluster, session, writer...

Selected attribute: `burst_type` :

- Distribution of bursts of type P (production), R (revision) and RB (edge revision) throughout a corpus.



1D: HISTOGRAM



Display of a single QUANTITATIVE attribute

- Duration, pct_pause, treshold, doc_len, burst_len ...

Chosen attribute: duration

- Each bar represents the number of bursts in the corpus with the duration attribute value between the bar's limit values a and b .

2D: SCATTER PLOT (X AND Y QUANTITATIVE ATTRIBUTES)

Relates two QUANTITATIVE attributes in the form of a scatter plot

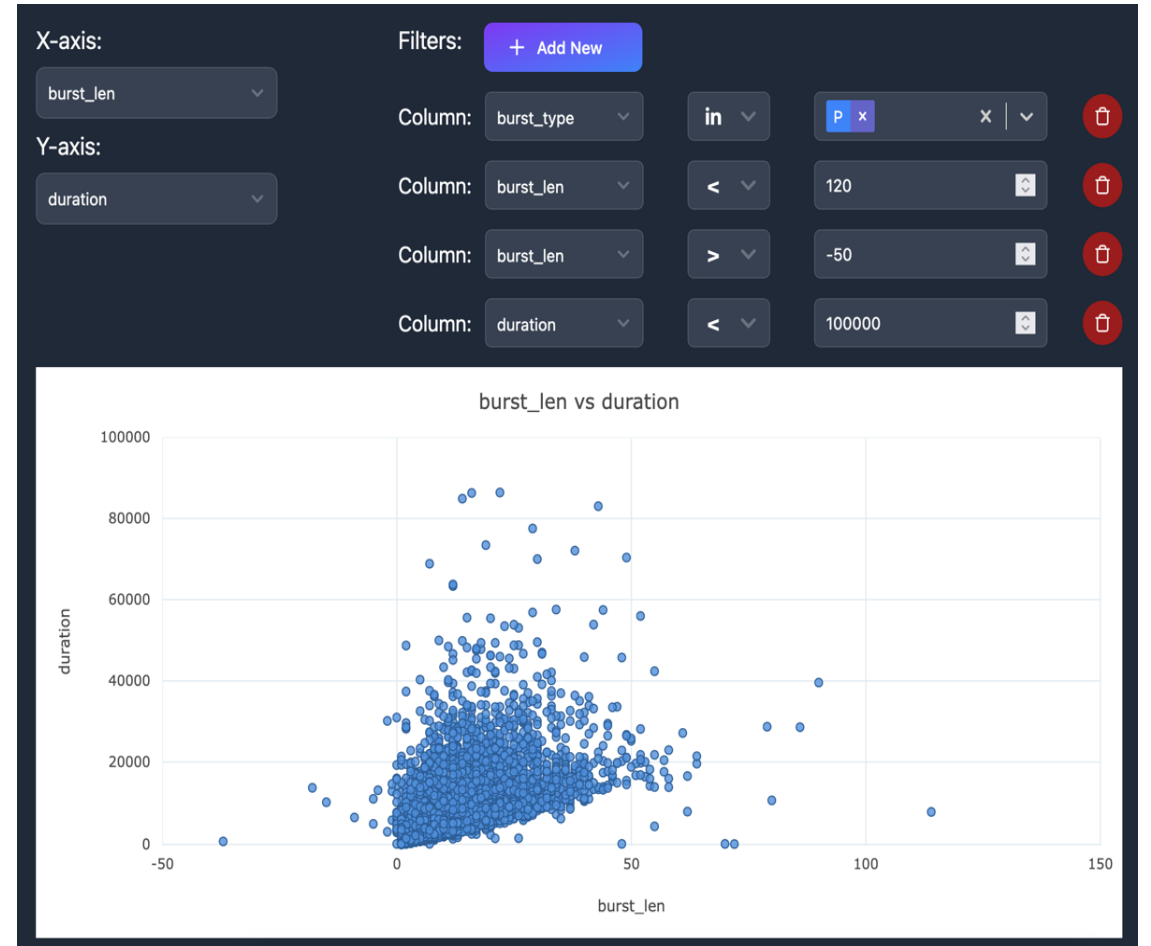
Highlights trends / potential correlations

In this example:

- X-axis attribute = burst_len
- Y-axis attribute: duration

4 filters to eliminate outliers that may affect the quality/readability of the representation:

- burst_type = P (production)
- burst_len < 120 and > -50
- duration < 100000



2D: SCATTER PLOT (QUALITATIVE X AND QUANTITATIVE Y)

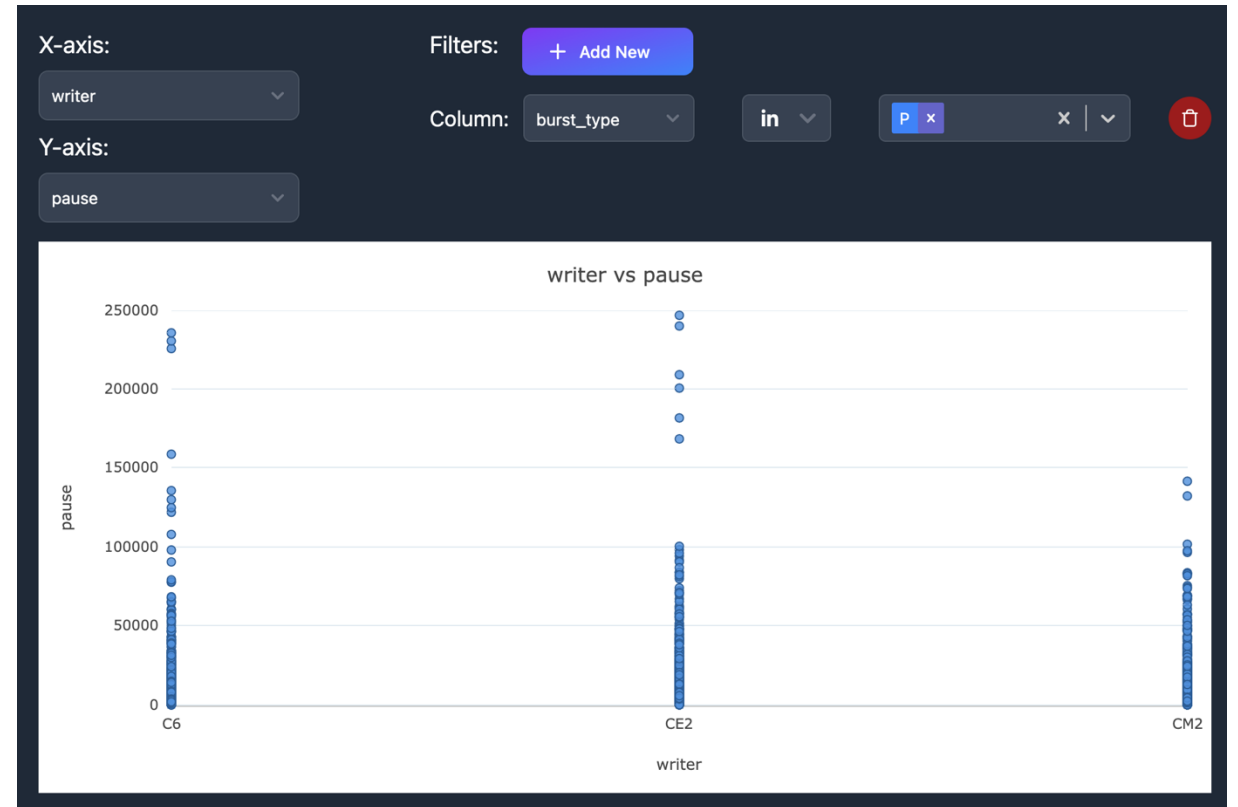
Relates a QUALITATIVE attribute on the X axis to a QUANTITATIVE attribute on the Y axis (or vice versa).

Highlights trends / potential correlations

In this example:

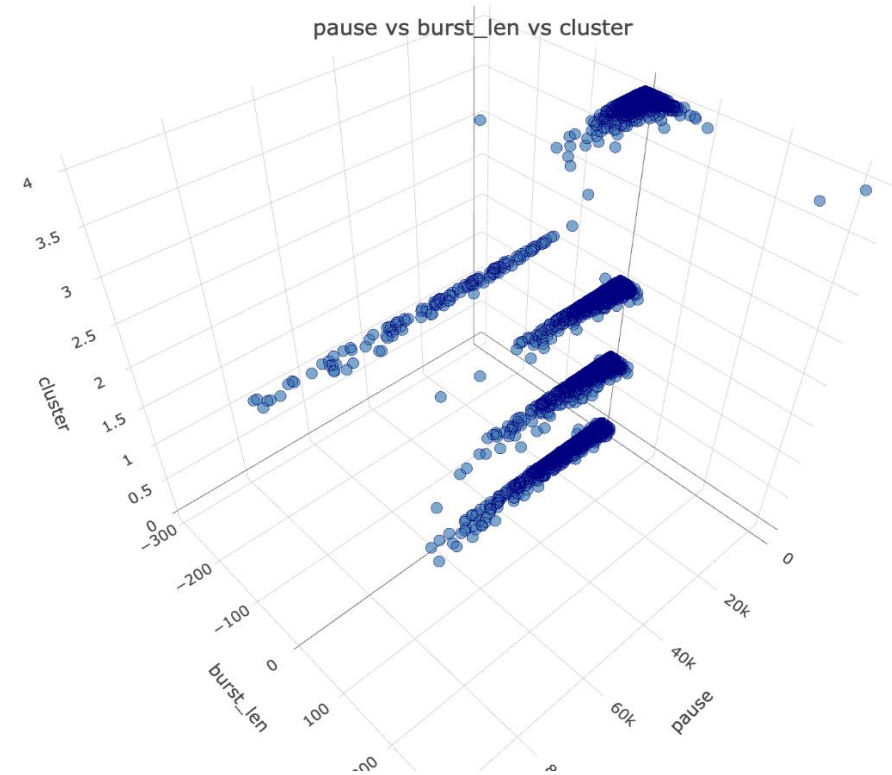
- X-axis attribute = writer
- Y-axis attribute = pause
- Writer metadata: here, the school level

1 filter to represent only P-type bursts
(production)

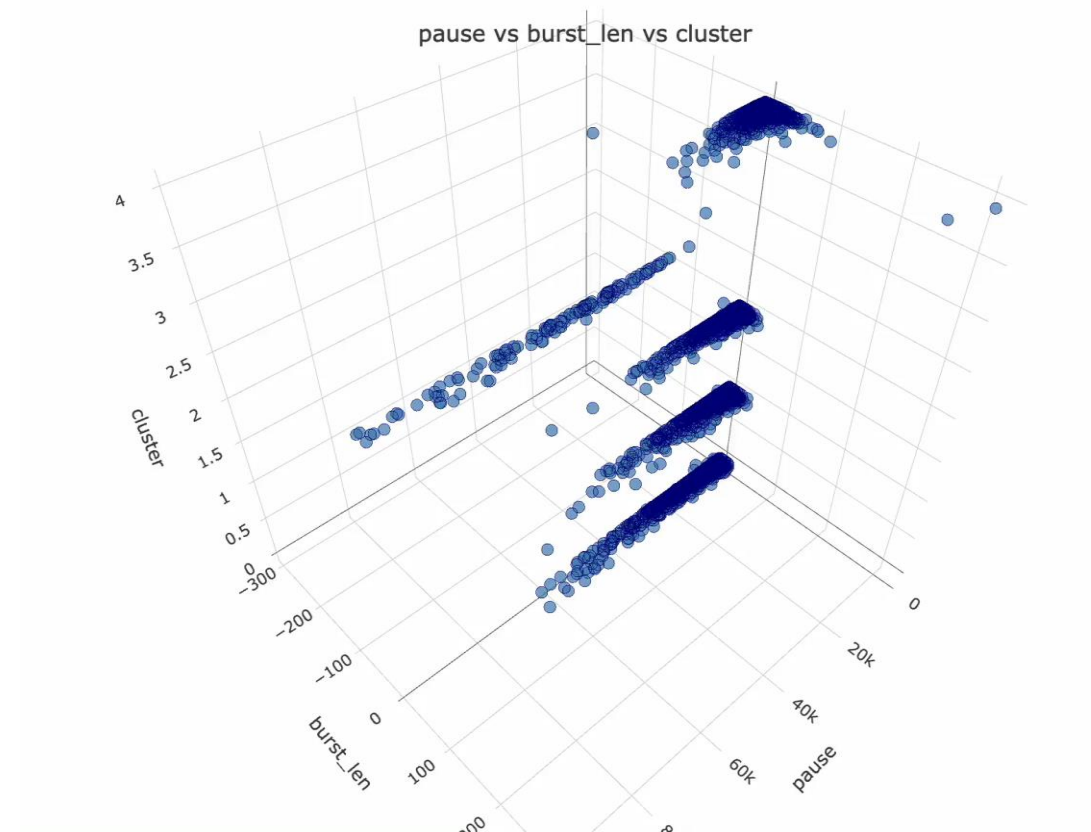


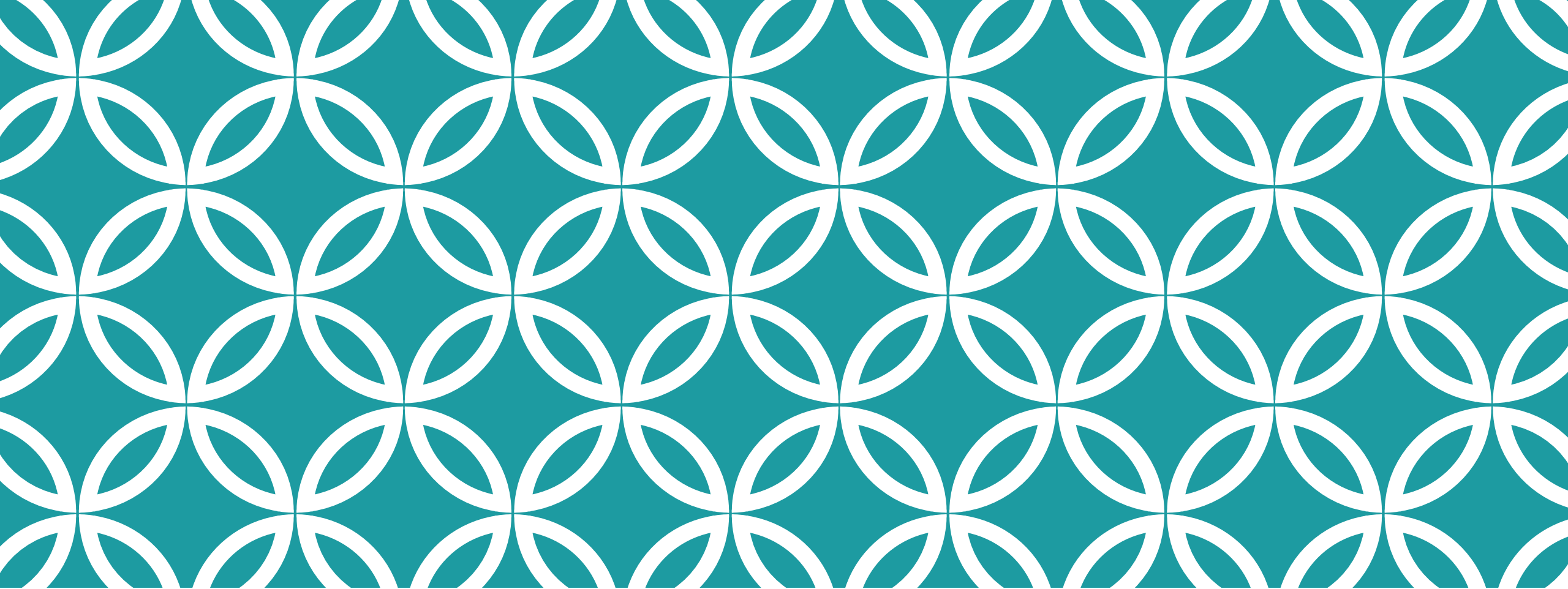
3D: SCATTER PLOT (QUANTITATIVE OR QUALITATIVE X,Y,Z)

- Linking three selected attributes in 3D space
- In the example :
 - X axis: pause in milliseconds
 - Y axis: burst_len (number of characters)
 - Z axis: cluster
- 2 filters:
 - burst_type = P (production)
 - pause \leq 100000 ms



3D VISUALIZATION MANIPULATION





**MACHINE LEARNING
CONCEPTUALIZATIONS: VISUALIZING
CLUSTERS**



BURSTS OVERVIEW

Displays bursts in the order in which they were produced

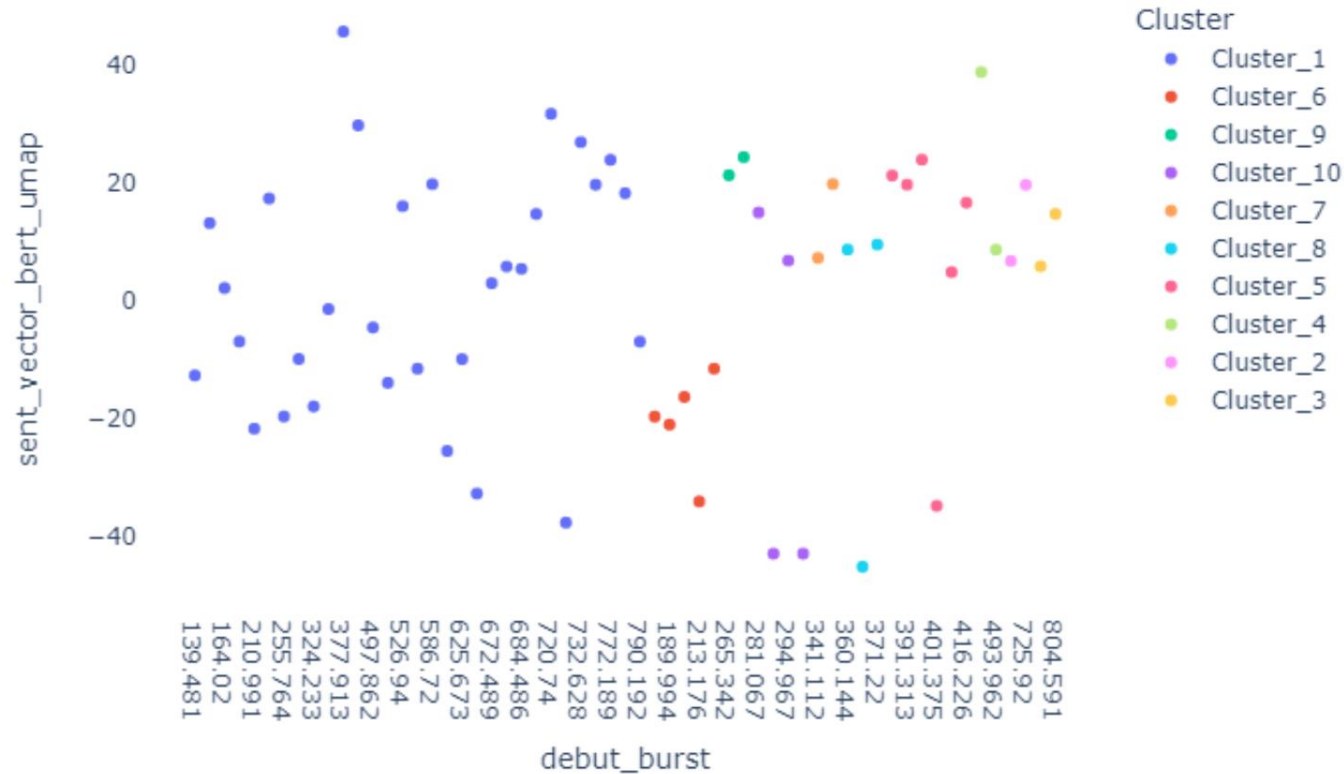
Burst color = cluster type

The width of the light-blue 'bubble' is proportional to the duration of the pause.

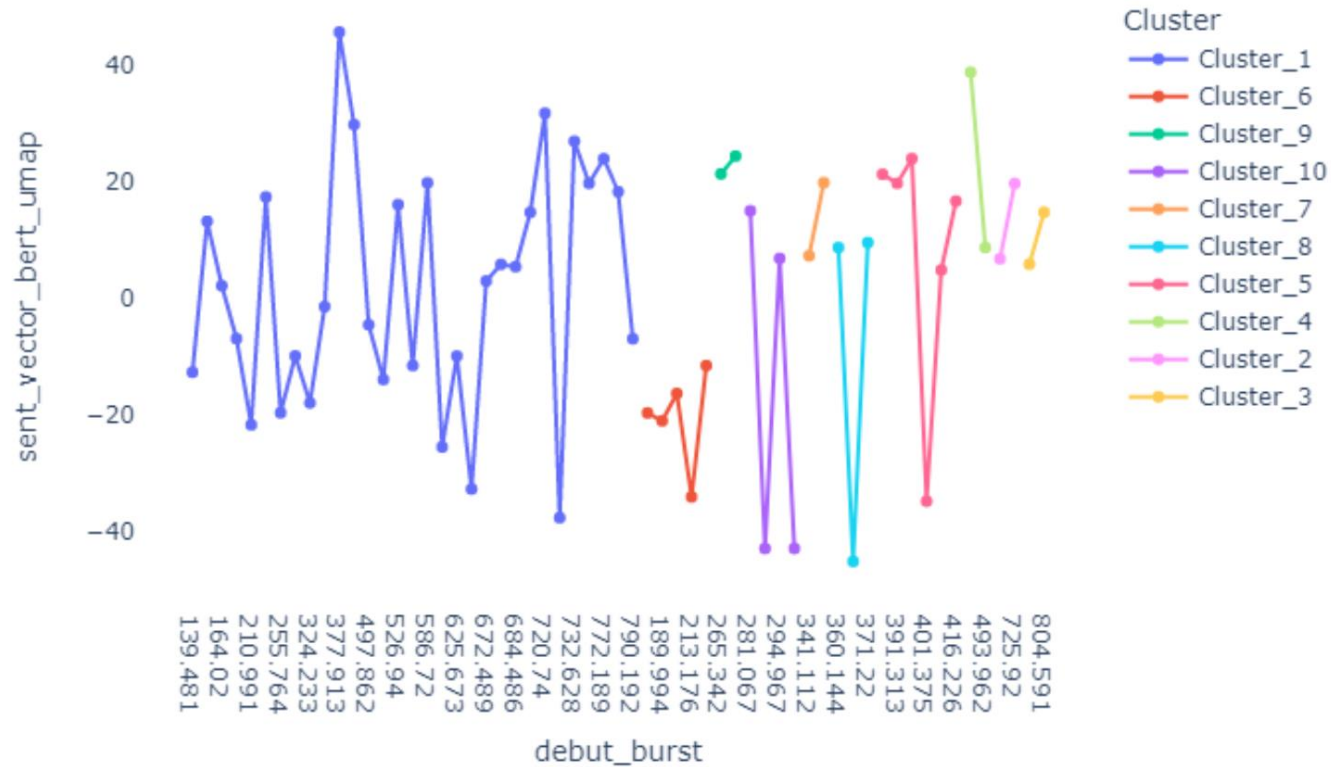
Only P-type bursts are displayed. R and RB bursts can be added.



BURSTS CLUSTERING VISUALIZATION BASED ON ON BERT-SENTENCE TRANSFORMER MODEL



DYNAMIC BEHAVIOR OF CLUSTERS BASED ON ON BERT-SENTENCE TRANSFORMER MODEL



CONCLUSIONS & PERSPECTIVES

Two specific constraints:

- the perception of the writing process as dynamic, (non)linear and resistant to a monochrome two-dimensional visual representation;
- the identification of the research questions that determine the identity of the objects and features to be observed, as well as the links between them.

The necessity of prior analyses:

- “visualization should be precisely a perspective that modulates its [the mass of information] internal possibilities and then stabilizes itself where it encounters an interpretive question that can be associated with a body of objects emerging from the bottom of the information archive. And it is left to artificial intelligence and its deep learning to offer *salient* visualizations.” (Basso Fossali et al. 2022: 8-9)

Graphic features like color, pattern, shape, position, orientation Bertin (1967) to be attributed a **value**, digitally exploited and operationalized – a new literacy?