The contributions of lexical analysis in the study of learning situations in STEM

ALEXANDRE BOOMS, FABIEN EMPRIN

INSPÉ de l'académie de Reims, CEREP (UR 4692)
University of Reims Champagne-Ardenne
France
alexandre.booms@univ-reims.fr
fabien.emprin@univ-reims.fr

ABSTRACT
Didactics research methods often rely on verbal exchanges to analyze classroom situations, resulting in large text corpora that can be difficult to process. This paper presents a method that partially automates this type of corpus analysis, thereby facilitating data access and enhancing its reliability for researchers. We describe a procedure based on the concept of didactic episodes, followed by data processing using the Reinert method. We test this approach on a corpus from various STEM disciplines in a sixth-grade level. We show that each didactic episode typically contains a specific "lexical world" determined by the Reinert method, distinct from others within the same learning sequence. These results lead us to introduce an analytical concept: didactic robustness. This concept enables researchers to evaluate how well a didactic episode reflects the intended learning objectives and allows comparisons within STEM didactics.

KEYWORDS
Didactics, methodology, content analysis, qualitative analysis, statistical analysis

RÉSUMÉ
Les méthodes de recherche en didactiques s'appuient souvent sur des échanges verbaux pour analyser des situations de classes. Cela aboutit à des corpus volumineux dont le traitement peut être complexe. Dans cette contribution, nous présentons une méthode d'analyse permettant un traitement automatique partiel de ce type de corpus afin de faciliter l'accès aux données pour le chercheur. Nous présentons une technique de préparation d'un corpus basés sur le concept d'épisodes didactiques, puis le traitement de ces données par une méthode informatique : la méthode Reinert. Nous testons ensuite cette démarche sur un corpus issu de différentes disciplines scientifiques et technologiques dans une classe de sixième. Nous montrons que les épisodes didactiques sont généralement le siège de « mondes lexicaux » spécifiques déterminés par la méthode Reinert, différents les uns des autres au sein d'une même séquence d'apprentissage. Ces résultats, nous permettent de développer un nouveau concept analytique, la robustesse didactique. Ce concept permet aux chercheurs d'évaluer dans quelle mesure un épisode didactique reflète les objectifs d'apprentissage visés et autorise des comparaisons au sein de la didactique des STIM.

MOTS CLÉS
Didactique, méthodologie, analyse de contenu, analyse qualitative, analyse statistique

INTRODUCTION

In didactic research, classroom session analysis often involves recording verbal interactions. Theyare then analyzed to gain insights into student activity. Chesnais and Coulange (2022) note that analyzing language—spoken or written—is a common way to access the 'black box' of student learning. Typically, researchers code transcripts manually, sometimes using digital tools (Derobertmasure et al., 2022). To ensure precision, double-blind coding is recommended (Van der Maren, 1996). However, large corpora from classroom recordings or interviews are time-consuming to process (Emprin, 2018). Lexicometric tools offer another approach, allowing for discourse analysis and automatic classification (Bart, 2011).

This raises the following research question: How can lexicometric tools assist researchers in analyzing large textual corpora from classroom transcripts in didactic research? One lexicometric method stands out—the Reinert method—which identifies 'lexical worlds' through co-occurrence analysis (Reinert, 1993). Emprin (2018), analysing teachers' discourse and official texts in particular, shows that its relevant for analyzing of written texts with a didactic purpose. We propose extending its application to verbal exchanges in STEM teaching. The contribution that follows thus describes how we selected this approach and applied it to a didactics research project (Booms, 2022). We first defined and identified didactic episodes manually based on Margolinas's (2004) criteria structuring an episode: question / student's response / conclusion' pattern. We then hypothesized that automating corpus processing would improve the efficiency and objectivity of this manual analysis through a computerized statistical approach. Although they use different techniques, lexicometric and manual analyses tend to converge, we propose going beyond lexicometry as a reliability tool to consider its use as a support for researchers. We thus propose developing the concept of didactic robustness, based on the combination of quantifiable criteria from statistical transcript analysis and qualitative data from manual processing.

We present our research context and theoretical framework, explain our choice of software, describe our corpus and methodology, and discuss how lexical analysis can enhance didactic research, especially in STEM disciplines.

CONTEXT

Our study focuses on the teaching of four teachers in the same sixth-grade class, each teaching a different STEM subject during the same period. This controlled setup minimizes confounding variables (Van der Maren, 1996). We focus on STEM disciplines due to their shared didactic features (McDonald, 2016). At the chosen school, science and technology are taught in an integrated curriculum that combines instruction in Life and Earth Sciences, physical sciences, technology, and mathematics, all of them focusing on viticulture, a key aspect of the local economy. Our sample consists of four teachers who were filmed during a teaching sequence: a three-session sequence in technology, a two-session sequence in physics, a four-session sequence in biology, and a six-session sequence in mathematics.

THEORETICAL FRAMEWORK

This study is part of a broader project aimed at analyzing the impact on teaching of the presence of a disabled student equipped with a computer. We use a theoretical framework aimed at analyzing student learning in this context. We adopt the propositions of Margolinas (2004),

based on Brousseau's Theory of Didactic Situations (TDS) (1997). Regarding mediation, Margolinas suggests breaking down the tasks given to students according to the structure: question / student's response / conclusion in order to support the analysis of language interaction protocols in class. "These sequences (question/conclusion) define episodes, which vary in length and degree of nesting" (Margolinas, 2004, pp. 26-27).

We refer to these episodes as didactic episodes. The three components of these episodes, namely, the question phase, the student's response, and the conclusion, are considered as episodes phases. Each episode is labeled as *EP_n*, and the phases are labeled as follows:

- *EP_n-1*: question phase
- *EP_n-2*: student response
- *EP_n-3*: Conclusion – The teacher concludes the task by giving the expected answer.
- *EP_n-4*: Students copy the conclusion

We hypothesize that didactic episodes, which involve specific verbal exchanges related to the task and the targeted knowledge, can be highlighted through lexical analysis.

LEXICAL ANALYSIS AND DIDACTIC EPISODES

The Reinert Method

The lexicometric analysis method proposed by Reinert (1986, 2007) reveals a tendency for vocabulary to distribute non-randomly across certain parts of a text corpus (Reinert, 2008). These vocabulary clusters in specific areas of the text form what Reinert refers to as 'lexical worlds'. According to him, these provide insight into "the content as being, first and foremost, the content of an activity" (Reinert, 1990, p. 982) and thus offer access to the speaker's mental organization (Bart, 2011). Although the corpus analyzed in this study is in French, the Reinert method can be fully applied to other languages(for example Baillat et al., 2018).

The mathematical and algorithmic processing involved in this method has been thoroughly described in several of Reinert's publications (1983, 1986, 1990). At the end of the computerized analysis "it is not the most frequent co-occurrences that are highlighted, but rather those that deviate most from a random distribution" (Emprin, 2018, p. 186), thanks to a technique called descending hierarchical classification (DHC).

The Reinert method results in the formation of stable classes, whose number is determined by the automatic analysis. These classes, according to Reinert's hypothesis (1990), represent the lexical worlds activated by the speaker. They act as traces that can be considered as indicators of the subject's discursive activity.

These resulting classes are then interpreted by the researcher, who evaluates their relevance to the research objective. To do so, the researcher cross-references the obtained classes with various variables assigned to them. This step does not alter the original classification provided by the software. This method is independent of the language of the corpus.

Choice of software

There are two main software tools based on the Reinert method: the commercial program Alceste and our choice, IRaMuTeQ, which is free and open-source. While cost is a key consideration, IRaMuTeQ also offers specific features that influenced our decision.

Firstly, IRaMuTeQ performs descending hierarchical classification (DHC) using an optimized computation process (Ratinaud, 2018). This allows for improved precision and speed when analyzing large corpora. Additionally, users can modify the default lemmatization dictionaries, providing flexibility in how text is processed.

The software also gives access to analysis matrices, calculation tables, and raw results, enabling deeper data exploration beyond what is shown in the graphical output interface.

Corpus preparation and analysis with the software

The base corpus consists of the complete verbatim transcriptions of interactions between the teacher and the students. We marked each section of the corpus with context-specific variables to guide the analysis. To achieve this, we manually segmented the data based on the "question/ student's response /conclusion" three-phase structure. This allows us to exclude exchanges with no direct learning objectives, such as roll-call or homework collection.

Sequences, sessions, episodes, and phases are identified as variables in the transcription using starred lines. For example, the line **** *discipline_math EP_3 sousep_3 seance_2 defines* phase 3 (conclusion without copying) of episode 3 from the second session of the mathematics sequence. This preparation allow us to identify text origin but as no influence in 'lexical word' computation.

The results depend on the size of the grouped text segments (RTS). We therefore adjusted the size of these RTS to maximize the percentage of analyzed text segments.

Detailed description of the corpus

We analyzed the following:

- Three episodes from the physical sciences sequence;
- Four episodes from the life and earth sciences sequence;
- Nine episodes from the mathematics sequence, three of which are incomplete due to unexpected schedule changes;
- Three episodes from a technology sequence.

TABLE 1

List of retained didactic episodes

Session	Episode
Physical Sciences (2 hours)	
1 + 2	EP 1 – "What is the Earth's direction of rotation?"
1 + 2	EP 2 – "What is the duration of a day?"
1 + 2	EP 3 – "Why is it colder in winter?"
Life and Earth Sciences (4 hours)	
1	EP 1 – "What happens to leaves?"
1 + 2	EP 2 – "What is in soil?"
2 + 3	EP 3 – "How does organic matter turn into mineral matter in soil?"
3	EP 4 – "Analyzing an experiment"
4	EP 5 – "Food chains"
Mathematics (6 hours)*	
1	EP 1 – "YES/NO Figures"
1 + 2	EP 2 – "Draw an axis of symmetry without folding"
2	EP 3 – "Drawing symmetry using a set square"
3* + 4	EP 4 – "Drawing symmetry using a compass"
3* + 4	EP 5 – "Drawing the symmetric of simple figures"
5	EP 6 – "Drawing the symmetric of a point with GeoGebra"
5 + 8	EP 7 – "A Pretty GeoGebra Frieze"
6* + 7	EP 8 – "Conservation of lengths"
7 + 8 + 9*	EP 9 – "Multiple axes of symmetry"
Technology (3 hours)**	
1	EP 3 – "How is glass bottle made?"

2	EP 2 – "What materials are used on a bicycle?"
4	EP 4 – "Origin of materials"
Note: * Episodes 3, 6, and 9 in mathematics were only partially filmed. ** In technology, only the rotation of the 3rd group is reported.	

DATA PRODUCED BY IRAMUTEQ

IRaMuTeQ performs a variety of automatic data analyses. Here, we focus on two types of result presentations derived from the Reinert method: profiles/anti-profiles and dendrograms.

Profiles and anti-profiles

The relationship between a class and a didactic episode can be established through the variables, namely, the didactic episodes and their phases, as defined earlier, with a chi-square (χ^2) test. An episode is considered significantly present in a class if the χ^2 is greater than 3.84 at a 95% confidence level ($p<0.05$). The relationship between the 'didactic episode' and its significant presence in the class constitutes a profile. A statistically significant absence – conversely represented by a negative value – in a class is referred to as an anti-profile in IRaMuTeQ.

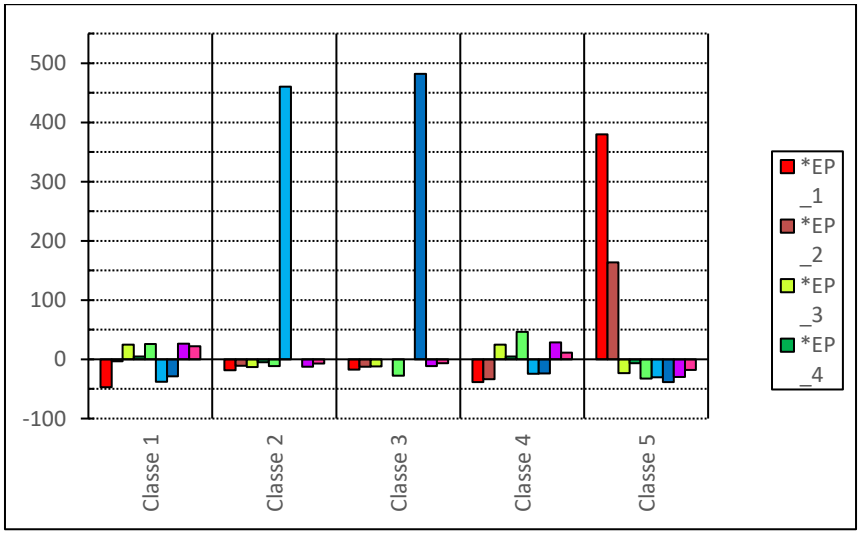
For example, IRaMuTeQ provides the following results in class 2 of the mathematics **S**ession:

TABLE 2

χ^2 values of didactic episodes (EP_n) in class 2 of the mathematics sequence

Class	Variable within the class's profile	χ^2	p	Variable within the class's anti-profile	χ^2	p
2	*EP_6	460,45	3,827 683e-102	*EP_1 *EP_3 *EP_8 *EP_5 *EP_2 *EP_9 *EP_4	-18,58 -13,04 -12,36 -11,17 -10,79 -7,32 -4,73	1,627 617e-05 3,047 523e-04 4,386 522e-04 8,332 032e-04 1,018 999e-03 6,811 522e-03 2,967 527e-02

FIGURE 1



χ^2 values of didactic episodes by class in the mathematics sequence

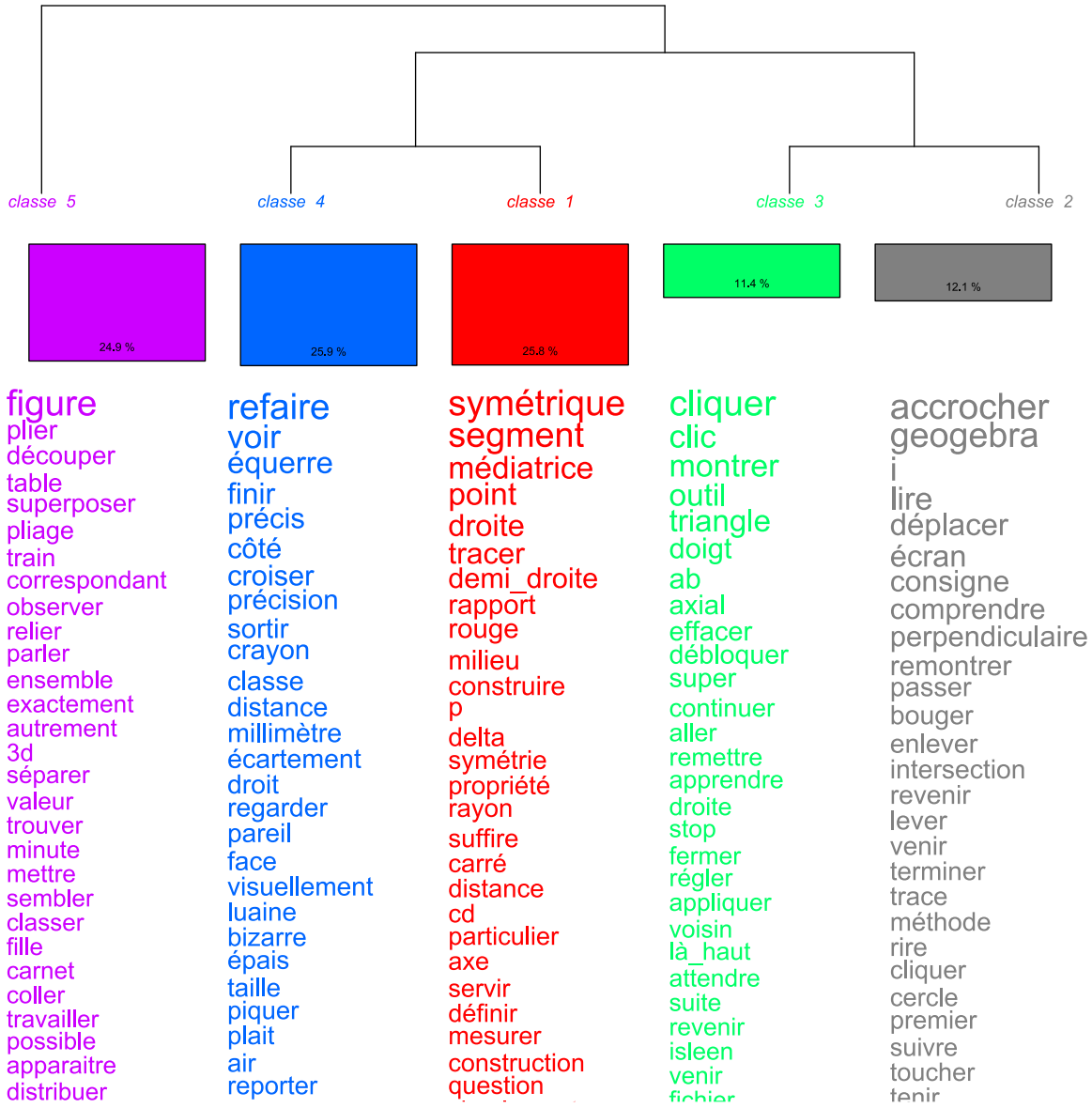
With a χ^2 of 460.5, episode 6 is significantly present in class 2. Episodes 1, 2, 3, 4, 5, 8, and 9 are significantly absent, all with χ^2 values below -3.84. Episode 7 does not appear in this table, indicating it is neither significantly present nor absent in class 2. For easier interpretation, these data can be visualized graphically, which helps in associating classes with episodes.

Although class 2 is clearly related to episode 4 (χ^2>450), due to scale issues in the graph, it is necessary to refer to the numerical tables to confirm significant absence of other episodes.

The dendrogram

IRaMuTeQ also generates a dendrogram from the corpus, which graphically represents the output of the descending hierarchical classification (DHC). It identifies various classes and their lexical content. These data are also available in table form, but the dendrogram makes the complex information more digestible. Words that appear on the branches are those significantly present in each class, ordered by decreasing χ^2, facilitating interpretation of the lexical worlds.

FIGURE 2



Dendrogram of classes obtained by the Reinert method for the mathematics sequence

The dendrogram shows five classes obtained through DHC. Since the vocabulary in each class represents distinct lexical worlds, we simplify by equating classes with lexical worlds. The distribution of these classes is tied to specific parts of the corpus, i.e., moments in the learning sequence.

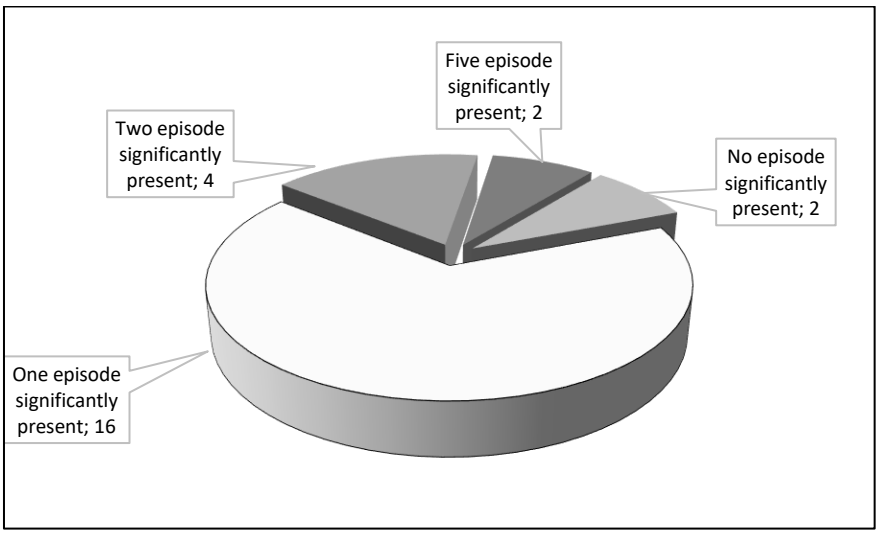
For example, the vocabulary in class 5, associated with hands-on activities (folding, cutting, layering, observing, etc.), appears at different times than the vocabulary in class 3, which can be linked to digital manipulation (clicking, click, showing, erasing, unlocking, etc.). The size of the lemma in the graph corresponds to its χ^2 value in the class, and therefore to the significance of its presence within that class. The connections between classes shown in the upper part of the dendrogram are also indicators of their proximity. Thus, class 3 is closer to class 2 than to class 5 in the transcription. Finally, another indicator in the dendrogram is the percentage of segments classified in each class, which gives an idea of its relative weight in the corpus.

ANALYSIS

Episodes generally contain a unique lexical world

Our working hypothesis was that specific lexical worlds exist within didactic episodes and that these can be revealed through lexical analysis. These lexical worlds could then characterize the knowledge at play and the learning modalities. We began with counting the number of episodes that are significantly present in each class identified through DHC.

FIGURE 3

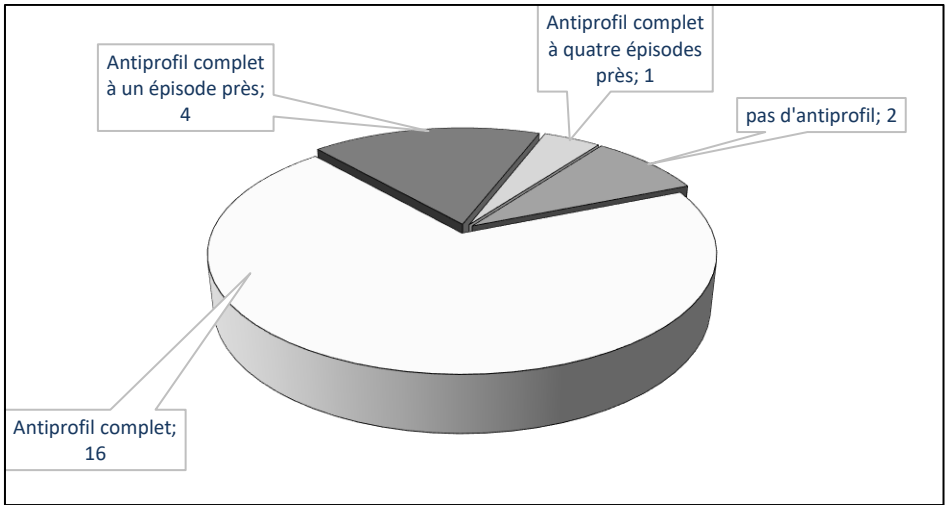


Relationship between IRaMuTeQ-identified classes and the number of significantly present episodes

The graph above shows that sixteen episodes are significantly present in only one class. Four classes are characterized by the significant presence of two episodes. Two classes are shared by five episodes, and two classes cannot be associated with any specific didactic episode.

We complemented this result with an analysis of significant absence. We defined a complete anti-profile as a class from which all other episodes are significantly absent. By extension, n-episode anti-profiles are those where all but n episodes are significantly absent.

FIGURE 4



Relationship between classes and number of episodes significantly absent from them

The graph above shows that sixteen classes display complete anti-profiles, four of them lack only one episode. The class that could not be assigned a profile in Figure 3 is not part of an anti-profile either.

The two graphs above, when used together, clarify the relationship between the lexical worlds constructed by IRaMuTeQ and the didactic episodes. Most classes are significantly associated with one or, at most, two didactic episodes, and are characterized by the total or partial exclusion of the others. There is thus a strong relationship between specific episodes and distinct lexical worlds.

An episode can contain two lexical worlds

A single episode may be associated with two different classes. This indicates that during a didactic episode, the teacher might use vocabulary corresponding to multiple lexical worlds.

TABLE 3

χ^2 values for classes shared by one episode across phases or sessions, depending on the discipline

Discipline	**Episode**	**Shared Class(es)**	**Presence in phase**		**Presence in Session**	
			phase	χ^2	Session	χ^2
Physical Sciences	EP 2	2	2-1	90,09	1	7,21
		5	None			
	EP 3	4	3-4 (NS)	2,57	2 (NS)	2,57
		6	3-2 (NS)	3,4		
Life and Earth Sciences	EP 2	1	2-3	54,11	2	63,34
		5	2-2	156,06	1	171,12
Technology	EP 4	2	4-4 (NS)	2,68	1	12,69
		4	4-1	214,14		
(NS) = Not significant						

The table shows which classes are shared by a single episode and how they appear across phases or teaching sessions.

In physical sciences, episode 2 has a specific lexicon in phase 2-1, represented by class 2. Class 5 seems to span the entire episode and both sessions, as it is not significantly present in any individual phase. Episode 3 in the same sequence is unevenly distributed across phases 3-2 and 3-4, but not in a statistically significant way.

In life and earth sciences, class 1 is associated with phase 2-3. Class 5 is associated with phase 2-2

In technology, class 4 is significantly present in phase 4-1. Class 2 is present in phase 4-4, but not significantly.

These results demonstrate that, sometime, different lexical worlds may emerge at different moments within the same didactic episode. In summary, didactic episodes indeed host identifiable and distinct lexical worlds, which can be revealed through lexical analysis.

The relationship between lexical world and the task at hand

Even when classes are identified and significantly associated with a didactic episode, the nature of the discourse within those episodes remains an open question. Just because a lexical world exists in an episode doesn't guarantee that it directly relates to the learning content. As Emprin reminds us: "The software 'cannot read'; it processes words without reference to their meaning" (Emprin, 2018, p. 189). This highlights a limitation of automated processing: while it supports corpus analysis, establishing a connection between identified lexical worlds and actual educational content remains the researcher's responsibility. To illustrate this point, we examine a document-based activity in technology on material families.

TABLE 4

Association between episodes, significantly associated classes, and vocabulary within those classes

Episode	χ^2 within the class	Class	Vocabulary from the class where $\chi^2 > 10$
EP 4 Origin of materials	62,20	2	aimer, ranger, arrêter, critérium, asseoir, k, Ardghal*, répondre, dernier, finir, citer, [vêtement], allumer, question, refaire, sac, fond, crayon, terminer, donner, place, Artegal*, attendre, réponse, raison, plait, suite, ordinateur
	270,67	4	chrome, Google, marcher, Firefox, flash, Technoflash, contrôle, Adobe, autoriser, get, lancer, taper, ouvrir, recherche, cliquer, fonctionner, copier, tac, explorer, navigateur, animation, activer, fermer, ordi, internet, clique, adresse, moteur, entrée, techno, grand, essayer, clavier, pc, appuyer, rechercher, aller, jour, fille
* Pseudonymized student name			

In this episode, class 2 reflects classroom management vocabulary and class 4 reflects instrumental support (Abboud-Blanchard et al., 2013) related to the use of computers. In this case, neither lexical world relates directly to materials science—the intended knowledge domain. Thus, even though automated processing identifies lexical worlds associated with episodes, these worlds may not always correspond clearly to the knowledge being taught. However, in this technology episode, the vocabulary does offer insights into class operations and resource use.

Lexical worlds shared between multiple didactic episodes

We have seen that the same lexical world can be shared by several didactic episodes with the same teacher. In our corpus, we observed two different manifestations of this phenomenon: the

first concerns a didactic link between two episodes that may be temporally close, and the second indicates a form of content transversalities across multiple episodes.

A didactic link between two episodes

Such articulation is common. It appears in technology, mathematics, and life and earth sciences. To illustrate this point, we will examine the overlap of two episodes in life and earth sciences.

Episodes 3 and 4 are temporally intertwined. Episode 3 is about "How does organic matter become mineral matter?" and Episode 4 is about “An experimental analysis task”.

TABLE 5

Association between episodes and classes in the life and Earth sciences sequence

	Class			
Episode	**2**	**3**	**4**	**7**
	χ^2	χ^2	χ^2	χ^2
3	**41.47**	-8.45	**11.78**	**23.09**
4	-13.80	**148.51**	**41.18**	-9.16

The table shows the existence of a shared lexicon (class 4) between episodes 3 and 4, as well as the presence of lexical worlds specific to episode 3 (classes 2 and 7) and to episode 4 (class 3).

The student response phase is shared between both, but the conclusions are distinct: one focuses on the experiment, and the other one on the roles of decomposers.

TABLE 6

Significant vocabulary from Class 4 in life and earth sciences

Verbs	**Nouns**	**Adjectives**	**Adverb**
passer, retirer, analyser, condition, tester, comparer, mélanger, réunir, enlever, laisser, décomposer, réaliser, apprendre, réaliser, tirer, apprendre	témoin, test, expérience, manipulation, résultat, alcool, importance, sucre, nature, étape, alcootest, définition, levure, élément, forêt, objectif, fois, litière, couleur	rose, important, vert, dernier	naturellement

The shared vocabulary (Class 4) shows how the teacher bridges the upcoming analysis on the role of decomposers (forest, to decompose, forest floor litter) with a previous analysis (alcohol, breathalyzer, rose) concerning the transformation of sugar into alcohol. She takes the opportunity to revisit the elements of the experimental approach (experiment, test, control).

Shared lexical worlds appear to be indicators of didactic proximity between episodes. In our corpus, this kind of didactic articulation also appeared in a similar way in mathematics and technology.

A form of transversality

We have seen that a single lexical world can span across two didactic episodes. Now we look at cases where more than two episodes are significantly associated with the same class. This phenomenon suggests a form of didactic transversality.

To illustrate this phenomenon, we observe the mathematics sequence. Several episodes (3, 4, 5, 8, and 9) appear in classes 1 and 4. Table 7 below shows the lexicon of these two

classes. For better data readability, we only retained values with a chi-square (χ^2) greater than 10.

TABLE 7

Vocabulary from Classes 1 and 4 in mathematics ($\chi^2 > 10$)

Verbs	Nouns	Adjectives	Adverb
Classe 1			
tracer, construire, suffire, servir, définir, mesurer, placer, conserver	segment, médiatrice, point, droite, demi-droite, rapport, milieu, symétrie, propriété, rayon, carré, distance, axe, construction, question, centre, codage, besoin, extrémité, dimension, côté, cercle, exercice, mesure	symétrique, rouge, particulier, deuxième, égal	Simplement, effectivement
Classe 4			
refaire, voir, finir, croiser, sortir, regarder	équerre, côté, précision, crayon, classe, distance, millimètre, écartement, droit, face	précis, pareil	visuellement

Vocabulary related to geometric objects and their properties is present in class 1, while terms related to mastering figure construction are found in class 4. Moreover, the pronoun 'you' (*tu*) is significantly present in class 4 and significantly absent from class 1, whereas the opposite is true for the pronoun 'we' (*on*).

TABLE 8

χ^2 values of tu (you) and on (we) in classes 1 and 4

	χ^2 in class 1	**χ^2 in class 4**
Tu	-61.78	75.66
On	13.99	-40.33

When the teacher addresses individual students, the vocabulary from class 1 is used, whereas the vocabulary from class 4 corresponds to moments of collective communication with the class.

Episodes 3 and 4 focus on basic techniques for plotting points using a set square and compass, while episode 5 deals with constructing the symmetrical counterparts of simple figures—an aspect that may, at first glance, explain the lexical proximity observed. In contrast, episodes 8 ('Conservation of Lengths') and 9 ('Multiple Axes of Symmetry') are designed to explore various properties of axial symmetry. The analysis of the transcripts from episodes 8 and 9 highlights that the exchanges focus primarily on the construction of figures, rather than on the knowledge initially intended. All these episodes, while formally distinct, share concerns about how well students master basic drawing techniques, which is a priority even when more advanced content is intended.

Conclusion of this section

When many episodes belong to the same class, it may indicate a loss of focus on disciplinary knowledge. Instead, generic task-related discourse (e.g., instructions, document handling) dominates. In our corpus, this kind of didactic articulation also appeared in a similar way in technology. However, it seems reasonable to frame the relationship between the increased

number of episodes involved and the shift toward technical procedures as a hypothesis for future research, rather than as a reliable result.

RESULTS AND DISCUSSION

The objective of this study was to examine whether the didactic episodes identified by a researcher (Margolinas, 2004) could be systematically associated with specific lexical worlds, as defined by Reinert's method. The analysis presented here indicates that, in most cases, each episode is predominantly linked to a single lexical world and corresponds to a complete anti-profile. When instances episodes are associated with multiple lexical worlds, it may call for a reassessment of the episode's granularity, suggest the presence of dual instructional objectives, or reflect shifts in discursive positioning—whether addressing the entire class or individual students. Overall, the findings support a clear relationship between didactic episodes and lexical worlds within the framework of our corpus. The dialogue between automated analysis and researcher-led analysis thus enhances the level of information that would be accessible to the researcher alone by refining the analysis of classroom discourse.

Our study demonstrates that the structuring of didactic activity into episodes remains consistent, whether it is performed manually or generated automatically using the Reinert method. This confirmation, based on statistical co-occurrence analysis, leads us to introduce the notion of didactic robustness of episodes. We extend this notion from the concept of task robustness as defined by Robert (2007, p. 303): "A robust task gives rise to possible activities—if not minimally varied from those anticipated in the a priori analysis—regardless of the teacher's interventions. Robustness thus corresponds to a potential for 'non-variability' in the activities linked to a given task formulation".

If a robust task resists variation caused by teachers' choices, a didactic episode, in our view, can be defined as one in which the targeted knowledge and learning objectives are clearly expressed in the lexicon—something that can be revealed through automated text analysis. These lexical elements should be present exclusively in this episode and, consequently, absent from others within the same sequence. Lexical analysis thus allows us to define three criteria for establishing this didactic robustness:

1. There is a lexical class that can be significantly associated with the episode.
2. The vocabulary significantly present in this class directly reflects the targeted learning content.
3. This lexical class is significantly absent from the other episodes in the learning sequence.
4. The values of chi-square (χ^2) corresponding to these three criteria provide both quantitative indicators (criteria 1 and 3) and a qualitative indicator (criterion 2), based on the lexical content, for assessing didactic robustness.

Since these values are not absolute but depend on the analyzed data, the notion of robustness is only meaningful within the same sequence and for a single teacher. It allows us to evaluate how precisely a teacher frames the learning objectives within their instruction, particularly through the most robust episodes. However, a lack of robustness may also be informative. In the case of mathematics, for example, it can reveal complex learning phenomena that students do not fully master, or reflect a lesson structure composed of multiple question-and-answer sequences with overlapping objectives.

If this finding is confirmed in other corpora, one possible implication would be the use of automated pre-processing of large datasets using the Reinert method. This would enable

researchers to rapidly identify robust didactic episodes and to compare instructional practices across teachers implementing the same lesson. The ability to quickly process large volumes of data could thus serve as a complementary tool for didactics researchers. At the same time, our analysis shows that human interpretation remains essential for understanding the lexical worlds identified by the software.

CONCLUSION

This study highlights the relevance of combining statistical co-occurrence analysis and human analysis to enhance and objectify the study of didactic episodes in classroom sessions within the STEM field.

Beyond its role in validation, we show that lexical analysis enables a new understanding of teaching and learning processes in action.

In 2018, Emprin suggested that lexical analysis could help make sense of large textual corpora generated from classroom transcripts, thus serving didactics research. In this contribution, we demonstrate that this statistical analysis method can extract meaningful data from such corpora, particularly when the data are organized into didactic episodes—a structuring concept from mathematics didactics.

This structuring appears to be applicable across STEM disciplines, which are didactically close to mathematics. It provides a way to objectify certain characteristics of the corpus, making it more usable for researchers.

That said, at times, this method produces objects that resist simple analysis. In our study, while the statistical treatment helped simplify complex situations, final understanding still depended on human interpretation. When such interpretation was applied, the results appeared coherent and justified, whether in terms of learning progression or student support.

We conclude that the continued development of automated lexical analysis offers promising opportunities to deepen our understanding of classroom dynamics. We believe that pursuing this line of research could help reveal recurring patterns that may serve as models for future applications of AI in didactic analysis.

REFERENCES

Abboud-Blanchard, M., Cazes, C., & Vandebrouck, F. (2013). Théorie de l'activité et double approche : Genèses d'usage de bases d'exercices en ligne. In J.-B. Lagrange & A. Robert (Eds), *Les technologies numériques pour l'enseignement usages, dispositifs et genèses* (pp. 37-54). Octarès éditions.

Baillat, A., Emprin, F., & Ramel, F. (2018). On words and discourse : From quantitative to qualitative. In G. Devin (Éd.), *Resources and Applied Methods in International Relations* (pp. 151-165). Springer International Publishing. https://doi.org/10.1007/978-3-319-61979-8.

Bart, D. (2011). L'analyse de données textuelles avec le logiciel ALCESTE. *Recherches en Didactiques*, *12*(2), 173-184. https://doi.org/10.3917/rdid.012.0173.

Booms, A. (2022). *Les pratiques enseignantes auprès d'un élève présentant des troubles de l'acquisition des coordinations et équipé de matériel pédagogique adapté*. Thèse de doctorat, Université de Reims Champagne-Ardenne, France. https://theses.hal.science/tel-03887749/document.

Brousseau, G. (1997). *La théorie des situations didactiques*. http://guy-brousseau.com/wp-content/uploads/2011/06/MONTREAL-archives-GB1.pdf.

Chesnais, A., & Coulange, L. (2022). Rôle du langage verbal dans l'apprentissage et l'enseignement des mathématiques. Synthèse et perspectives en didactique des mathématiques. *Revue Française de Pédagogie*, *214*, 85-121. https://doi.org/10.4000/rfp.11357.

Derobertmasure, A., Dubois, L.-A., Delbart, L., & Bocquillon, M. (2022). L'analyse textuelle outillée : L'exemple des traces de réflexivité dans les écrits professionnels. In B. Albero & J. Thievenaz (Éds), *Enquêter dans les métiers de l'humain* (pp. 564-577). Éditions Raison et Passions. https://doi.org/10.3917/rp.alber.2022.02.0564.

Emprin, F. (2018). Les apports d'une analyse statistique des données textuelles pour les recherches en didactique : L'exemple de la méthode Reinert. *Annales de Didactique et de Sciences Cognitives*, *23*, 179-200. https://doi.org/10.4000/adsc.458.

Margolinas, C. (2004). *Points de vue de l'élève et du professeur. Essai de développement de la théorie des situations didactiques*. Note de synthèse pour l'habilitation à diriger des recherches, Université de Provence - Aix-Marseille I, France. https://tel.archives-ouvertes.fr/tel-00429580v2

McDonald, C. V. (2016). STEM Education : A review of the contribution of the disciplines of science, technology, engineering and mathematics. *Science Education International*, *27*(4), 530-569.

Ratinaud, P. (2018). Amélioration de la précision et de la vitesse de l'algorithme de classification de la méthode Reinert dans IRaMuTeQ. In *Proceedings of the 14th international conference on statistical analysis of textual data* (I, pp. 616-625). http://lexicometrica.univ-paris3.fr/jadt/JADT2018/actes-jadt18.pdf.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique : Application à l'analyse lexicale par contexte. *Les Cahiers de l'Analyse des Données*, *2*(8), 187-198.

Reinert, M. (1986). Un logiciel d'analyse lexicale. *Les Cahiers de l'Analyse de Données*, *11*(4), 471-484.

Reinert, M. (1990). Une méthode de classification des énoncés d'un corpus présentée à l'aide d'une application. *Les Cahiers de l'Analyse de Données*, *15*(1), 21-36.

Reinert, M. (1993). Les « mondes lexicaux » et leur « logique » à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et Société*, *66*(1), 5-39. https://doi.org/10.3406/lsoc.1993.2632.

Reinert, M. (2007). Postures énonciatives et mondes lexicaux stabilisés en analyse statistique de discours. *Langage et Société*, *121-122*(3), 189-202. https://doi.org/10.3917/ls.121.0189.

Reinert, M. (2008). Mondes lexicaux stabilisés et analyse statistique de discours. In *Actes de la JADT 2008* (pp. 981-993). JADT.

Robert, A. (2007). *Stabilité des pratiques des enseignants de mathématiques (second degré) : Une hypothèse, des inférences en formation*. *Recherches en Didactique des Mathématiques*, *27*(3), 271-312.

Van der Maren, J.-M. (1996). *Méthodes de recherche pour l'éducation* (2. éd). De Boeck Université, Université de Montréal.