

Teaching the specialized language of Mathematics with a data-driven approach: what data do we use?

Cecilia Fissore¹[0000-0001-8398-265X], Francesco Floris¹[0000-0003-0856-2422],
Marina Marchisio Conte¹[0000-0003-1007-5404], Matteo Sacchet¹[0000-0002-5630-0796]

¹ University of Turin, Turin, Italy
{cecilia.fissore, francesco.floris, marina.marchisio,
matteo.sacchet}@unito.it

Abstract. Numerous studies in Mathematics Education have shown that among the causes of disciplinary learning difficulties are the acquisition and understanding of its specialized language. Data-Driven Learning (DDL) is a didactic approach that treats language as data, and which sees students as researchers doing guided discovery activities. The exploration of corpora can effectively support reflection on the specialized languages of Mathematics. What data should be used? Students use the most recent Large Language Models and Google daily, which can be used for linguistic investigations. However, it must be remembered that there is no control over the data on which searches are carried out, the results of the searches, and the type of language they use. In this paper, we present a recent DDL research activity with 80 secondary school students on the specialized language of Mathematics. The students carried out linguistic investigations on a specially designed corpus and carried out corpus-based activities with automatic formative assessment within a Digital Learning Environment. The results show that the students appreciated the proposed activities. They develop linguistic and mathematical skills and become more aware of the importance of the language they use. Students developed digital skills in browsing, searching, and filtering data, but also in evaluating data, information, and digital content. In the future, given the notable diffusion of new AI tools, it is important to train teachers and students on their strengths and weaknesses, and how they influence teaching and learning.

Keywords: Data-Driven Learning, Large Language Model, Language for Specific Purposes, Mathematics Education.

1. Introduction

Every subject from Geography to Mathematics has a specialized language, which is the typical way of expressing itself in that area. Research in mathematics education has studied how learning difficulties in Mathematics depend on linguistic difficulties, that is, on the acquisition and understanding of its specialized language [1]. Understanding the language of Mathematics is considered by didactic research as one of the major obstacles for learning the discipline, at all school levels [2]. Furthermore,

Mathematics has always been one of the school disciplines in which students of all grades encounter the greatest difficulties. For this reason, many researchers in mathematics education study the language that students use in class (with their peers or with the teacher) for questions or collaborative or problem-solving activities; that teachers use during explanations; which is present in textbooks; and so on.

To help students develop language skills, it is possible to use the Data-Driven Learning (DDL) methodology. DDL treats language as data and sees students as researchers doing guided discovery activities [3]. Students can discover grammatical patterns, word meanings, or other aspects of language by searching linguistic data and investigating large amounts of authentic language. This methodology was born for language learning but has recently also been used for learning specialized languages [4]. DDL is closely linked to the concept of corpus, a collection of texts, on which linguistic investigations are carried out.

In recent times, the use of Large Language Models (LLM) by students has spread considerably, also for educational purposes. LLMs are deep neural network models (Deep Learning) capable of acquiring vast knowledge of a language, obtaining it from enormous quantities of texts, mainly taken from the Web [5]. The impact of LLM on education can be enormous [6]. LLMs could change the educational learning goals, learning activities, and assessment and evaluation practices. The advantages of LLM in education include personalized learning, timely support, continuous assessment, resource delivery, collaboration, and so on [7]. One of the most used LLM by students for learning is currently ChatGPT (<https://chat.openai.com/>). ChatGPT is just one of many ways to use LLM. Research is making rapid progress in this area and new models are constantly being developed [5]. The release of ChatGPT as a chatbot based on GPT models by OpenAI marks a pivotal milestone in the development of chatbots and conversational AI [7]. By conversing with ChatGPT, students also try to study and learn Mathematics. There is a lot of research studying the math skills of ChatGPT (see for example [8], [9]). Not as much research studies ChatGPT's linguistic skills in mathematics. LLMs represent a significant step towards the integration of natural language understanding and generation in the field of information technology and artificial intelligence, with applications that extend to multiple sectors and scenarios (mathematical, chemical, legal, etc.). However, ChatGPT does not always manipulate the specialized language of these disciplines. LLM, as well as Google, can be used for linguistic investigations [10]. However, it must be remembered that there is no control over the corpus data on which they carry out searches, over the results obtained from the searches, and over the language they use.

This paper illustrates a project connecting DDL with a Digital Learning Environment (DLE) integrated with an Automatic Assessment System (AAS) to develop competencies in the language of Mathematics for Italian secondary school students [11]. The DLE is a stimulating environment to boost motivation and foster constructivist acquisition though respecting students' learning pace [12]. With the AAS it is possible to create adaptive questions with personalized, immediate, and interactive feedback giving information not only about how the DDL task was performed, but also about the process to be mastered, thus enabling self-regulation and self-monitoring of actions [13] [14]. The DDL approach gives students the language ac-

cess keys to the content and, in the case of Mathematics, proves its effectiveness in helping them to understand and manage a language that can pose actual obstacles to problem solving activities and exercises.

In section 2 the state of the art is presented; section 3 illustrates the research activity and methodology; section 4 shows the main results emerging from the research activity and a discussion on the implications and the conclusions.

2. State of the art

2.1 The specialized language of Mathematics and ChatGPT

Numerous studies in mathematics education have shown that among the causes of difficulties are the acquisition, understanding, and management of its language [1]. Over time, Mathematics has developed a specialized language that has become increasingly universal, precise, concise, and effective. This language has its own semiological code, which is realized in texts in which technical terms, figures and graphs, and symbolic expressions (equations, formulas, algebraic expressions, etc.) coexist; the latter are sometimes inserted in sentences which, for the rest, use the common, everyday language [15]. Mathematicians and researchers in mathematics education have different attitudes and points of view towards language, which derive from two apparently divergent aspects: the specificity of Mathematics and its language and the role of the context in mathematical communication [16]. Teaching is communication and one of its purposes is to encourage student learning, using a language that does not hinder understanding and using metaphors to make concepts understood. At the same time, one of the main objectives is to acquire that specialized language. Often the linguistic aspects of teaching a language are considered only in a CLIL context, but teachers tend to forget that language education is a task that belongs to the training curriculum as a whole [4]. The characteristics of the mathematical language often contrast with the linguistic habits of the students which are difficult to understand [2]. Secondary school students are often forced to deal with the languages of the disciplines without having adequate linguistic scaffolding that allows them to make the necessary differentiations and categorizations of the language for specific purposes with respect to more general language varieties. Students' linguistic difficulties may concern the comprehension or production of verbal texts, symbolic expressions, and figures. In all cases, it is not easy to recognize students' linguistic difficulties, also because they often compete with other difficulties related to gaps in content, careless mistakes, etc. Language difficulties can cause behaviours that can be identified, such as misinterpretations or the production of incoherent texts, but they can also cause no behaviours, such as not responding, or behaviours that are difficult to interpret, such as random answers [16].

In recent times, the use of LLMs by students for learning mathematics has spread considerably [7]. Consciously or not, this also affects students' learning of its specialized language. However, LLM, such as ChatGPT, do not always use specialized language. If we ask ChatGPT “Do you know the specialized language of Mathematics?”, a response is “Yes, I'm familiar with the specialized language of mathematics. Math-

ematics has its own technical vocabulary that includes specialized terms, symbols, and notations” [17]. For example, if we ask ChatGPT “can you define in 100 characters what “intersection points” are in Mathematics?”, the answer is “points where lines or objects meet, sharing coordinates” [17]. The use of terms taken from a common language such as “object” or “meet” can confuse students because in a common language these terms have different meanings. Furthermore, the use of the specialist verb “intersect” is missing, which must be acquired by students. Otherwise, students will have difficulty interpreting the meaning of the concept of intersection and tasks such as “at what point does the function intersect the ordinate axis?”.

2.2 Data-driven learning

To learn specialized languages, it is possible to use the DDL methodology [11], [18]. The main idea behind DDL is that learners can discover grammatical patterns, word meanings, or other aspects of language through searching linguistic data. Corpora are proving increasingly influential in language teaching as sources of language descriptions [4]. A corpus is a collection of texts or parts of them in a finite number in an electronic format processed in a uniform way that makes it manageable and searchable using a computer [19]. Students have to do with a “massive but controlled exposure to authentic input” so fundamental for language learning [20] and such controlled and contextualized contact fosters more language awareness, noticing, and autonomy. Tim Johns [3] argues that at the heart of the approach is the use of the machine as a rather special type of informant. Once the informant answers the question, students must make an effort to “make sense of that response and to integrate it with what is already known” [3]. Corpora provide data, but do not interpret them: it is up to learners’ work and responsibility to evaluate the information found. The easiest way to explore corpus data is directly via concordancers: end-users may display a list of words with their immediate context. A concordance based on KWIC (keyword in context) can reveal a massive amount of information about the language: idioms, collocations, fixed phrases, and frequency data. Since then, literature on the uses and benefits of corpora for language learning has rapidly grown, although there is still little field practice, in Italy at least [4].

LLMs, such as ChatGPT, can be used for linguistic investigations. However, when carrying out linguistic research in a specialized field it is particularly important to design prompts to produce desired outputs. Students, for example, might assume that if they talk about “functions” ChatGPT understands that they are talking about math. However, if we ask ChatGPT “What are the verbs that have “function” as the subject?”, we get an ambiguous answer like: “Verbs that have “function” as the subject are typically used when describing how something operates or works; like functions, operates, works, performs, runs, executes, etc.” [17]. If we ask “What are the verbs that have “function” as the subject in Mathematics?”, the answer is radically different (see Fig. 1). In this case ChatGPT interpreted the task correctly and the results refer exclusively to the mathematics field. As shown in fig. 1, ChatGPT reports some verbs in the form of a list, followed by an example sentence in which the verb is present. The presence of the example can help students understand in which context it is ap-

appropriate to use the verb, and they can reason about its meaning. However, as we have no control over the data, we do not know what determines the choice of verbs and the order in which they are presented. Furthermore, the concept of function is very transversal in Mathematics and is studied from lower secondary school to university. If we add more details to the prompt, the results become more accurate.



In mathematics, "function" is often used as the subject of various verbs to describe its behavior and properties. Here are some common verbs used in mathematical contexts where "function" is the subject:

1. **Evaluates:** "The function evaluates to a specific value at $x = 2$."
2. **Maps:** "The function maps elements from set A to set B."
3. **Transforms:** "The function transforms the input data into a different representation."
4. **Solves:** "The function solves for the roots of the equation."
5. **Graphs:** "The function graphs as a parabola."
6. **Derives:** "The function derives its values from the underlying data."
7. **Integrates:** "The function integrates over the interval $[a, b]$."
8. **Differentiates:** "The function differentiates to find its rate of change."
9. **Computes:** "The function computes the result based on its input."
10. **Approximates:** "The function approximates the solution using a numerical method."

These verbs are commonly used in mathematics to describe how functions operate, the relationships they represent, and the calculations or transformations they perform.

Fig. 1. Example of mathematical linguistic investigation with ChatGPT.

Google itself can be used for linguistic investigations, but positions on the nature of the web as a corpus are multiple and conflicting [4]. The exploration of web resources as a 'mega corpus' responds to the lack of corpora available to study increasingly complex specific linguistic problems. At the same time, materials age quickly compared to the continuous evolution of language, also in relation to new technologies and new means of online communication. However, the World Wide Web is constantly updated, has no finite dimension and was not designed with linguistic intent. For this reason, it cannot be considered a corpus entirely [19].

3. Research activities and methodology

The research activity involved 4 classes of two Italian secondary schools for a total of 80 students in grade 11 and their teachers. The two schools are scientific high schools. The activities with the students involved two classes, while the other two classes made up the control group. The didactic experimentation, from November to December 2021, consisted of four two-hour meetings in the classroom. The teachers partici-

pated in all the meetings with the students and carried out the activities themselves to study the teaching methodologies proposed. Two specific corpora were created for the activities, selecting texts relating to the concept of function from textbooks for secondary school. The first corpus was in Italian, and the second corpus was in English.

Corpus-based activities with automatic formative assessment are questions implemented within the DLE using the AAS [14]. All questions are characterized by immediate and interactive feedback. From a formative point of view, in each question, the students have several attempts to answer. The questions are designed to guide students to consult the corpus and to make them reflect on the mathematical concepts. To carry out the activities, students use the AntConc concordance tool (<https://www.laurenceanthony.net/software/antconc/>) and the DLE (available at the link <https://linguaggispecialistici.i-learn.unito.it>) side by side. Fig. 2 shows an example of a corpus-based activity with automatic formative assessment. The question is divided into two parts: in the first part students are asked to search for the word “funzion*” (which means function) in the corpus by inserting 1R, 2R, and 3R as concordance levels, and to select the verbs that have function as subject; in the second part of the question the students had to complete the proposed statements by inserting the verbs identified in the previous section.

Cerca la parola “funzion*” inserendo come livelli di concordanza 1R,2R e 3R. Quali verbi hanno funzione come soggetto? Una funzione..

- dichiara interseca
 determina associa
 assume incontra
 combina esprime

Sezione Tentativo 1 di 3

Verifica

Completa le seguenti affermazioni utilizzando i verbi individuati nella sezione precedente:

- La funzione gli assi nel punto (0; 0)
- Una funzione agli elementi del dominio un solo valore nel codominio
- La funzione il valore 3 in $x=2$
- Una funzione la relazione tra due grandezze

Sezione Tentativo 1 di 3

Verifica

Fig. 2. Example of corpus-based activity.

Students see the second part after answering the first. At the end of each part of the question, there is a "verifica" button (which means "verify"). Students could click this button after answering the question to obtain immediate feedback on the correctness of the answer. If they made a mistake, they could try to answer the question again. In case of an incorrect answer after finishing the attempts, the students see the correct

answer that is needed to carry out the next part. To answer the first part of the question, students had to use the AntConc software and enter the word "funzion*" in the search bar. Inserting the asterisk symbol is used to search for both the word function in the singular ("funzione") and the plural ("funzioni"). Fig. 3 shows an example of what appears on the screen to students after the search. The levels of concordance specified in the text of the question are placed under the search bar and must be set by the students. Through the levels of concordance one chooses which words to highlight in the search. In this case, by selecting the levels "1R", "2R" and "3R", the search shows all the phrases of the corpus in which the word function appears and the following words are highlighted: the words one position to the right with respect to the searched word (in red), the words two positions to the right with respect to the searched word (in green), and the words three positions to the right with respect to the searched word (in purple). In step-by-step processes with multiple response attempts, students earn partial credits for the correctness of their answers. These points act as a motivational lever and, by expressing intermediate levels between "incorrect" and "correct", also offer teachers and students more precise information about the students' competence in a particular domain.

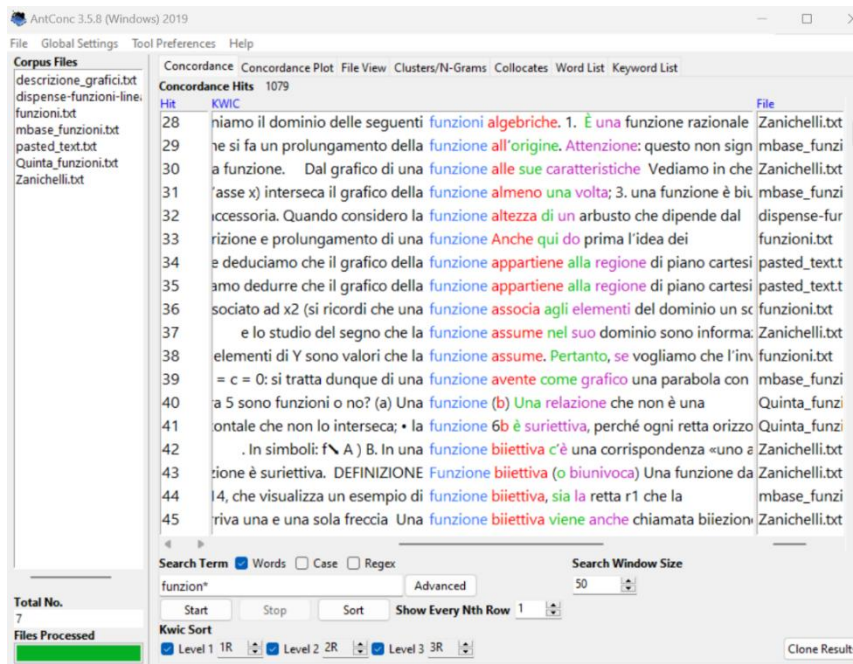


Fig. 3. Example of linguistic investigation with AntConc.

The research methodology of the didactic experimentation included:

- Initial test (before starting the research activity) and final test (at the end of the research activity) to evaluate students' mathematical and linguistic skills;
- Initial questionnaire (before starting the research activity) and final questionnaire (at the end of the research activity) to evaluate students' opinions about mathematics and mathematics lessons.

All 80 students took an initial test and an initial questionnaire before starting the research activity. The test consisted of ten questions and included linguistic questions and mathematical exercises. The questionnaire was composed of 40 Likert-scale questions (on a scale from "1 = very disagree" to "4 = very agree") covering opinions on Mathematics, mathematics lessons, the study of Mathematics for school, school and extracurricular activities in Mathematics, students' attitudes towards Mathematics. At the end of the experimentation, all students took a final test structured like the initial test and a final questionnaire. The questions in the final questionnaire were the same as in the initial questionnaire. Questions were added to the students who carried out the activities in the classroom in order to analyze the satisfaction of the activities and methodologies proposed.

4. Results and conclusions

Students appreciated the proposed methodologies and the mathematical activities that were very different from the traditional ones. The use of DDL approach had a very positive and motivating impact on the students: according to them, the activities were practical and not theoretical, interactive and non-transmissive, and based on real data. According to the students, the activities were interesting (3.23) and understandable (3.43). The activities stimulated their interest in Mathematics (2.60) and raised students' awareness of language (3.14). They appreciated the group discussions and the group activity (3.34). According to students, it was easy to use AntConc (3.09) and it was easy to understand how to use AntConc to answer questions (3.14). They also believe that learning to use text analysis software is useful for the future (3.14). According to students, the exercises helped them to become more aware of the language to use (3.14) and the exercises on language also made them reflect on meanings and contents (2.77).

Students appreciated the formative aspect of the activities: the possibility of viewing the correct answer immediately after answering a question (3.63) and the possibility of retrying the exercise in case of a wrong answer (3.40). They agreed that being able to have the assessment immediately after solving an exercise helped them understand how the question should be answered (3.51). The DDL approach gives students the language access keys to the content and, in the case of Mathematics, proves its effectiveness in helping them understand and manage its language that can pose obstacles to problem solving activities and exercises. Students who were exposed to DDL engaged in an involving activity that improved their language skills in their actual working practice. The study of language does not only concern the choice of words to use in a more specialized language, but that the terms convey meanings, and

by exploring the language one explores the meanings. The new and interdisciplinary nature of the activity greatly impressed the students and also their teachers, who for the first time discovered these methodologies and carried out linguistics activities. These activities allow students to better understand mathematical topics and to be more aware of the importance of using correct language to understand and be understood correctly.

Using the tools and techniques of corpus linguistics for pedagogical purposes, students acquire language and digital skills. The use of the digital learning environment, as well as the use of an automatic assessment system, enables the generation and collection of data on student learning: about learning processes and not just about results. These data are useful to drive and adjust the learning path, make choices and decisions, and support learning in several ways. Teachers appreciated the proposed contents and methodologies very much. In the future, it would be important to train teachers on the use of these methodologies and technologies. In particular, for Mathematics teachers it could be very interesting and formative to learn the analysis of specialized texts through the consultation of a corpus and the design of data-driven learning activities with formative assessment to make students reflect on the specialized language of Mathematics. It is important to promote reflection among all teachers of all levels on the importance of acquiring specialized languages and how linguistic difficulties can influence disciplinary difficulties.

The use of technologies and a student-centered DDL approach allows students to be involved in linguistic tasks, especially in a STEM subject to which they are not accustomed. The exploration of corpora can effectively support reflection on the specialized languages of Mathematics. There are many technologies used daily by students that support linguistic investigations, but it is important to question the types of data on which they are carried out. Students use the most recent Large Language Models and Google daily and these tools can be used for linguistic investigations. However, there is no control over the data on which searches are carried out, the results of the searches, and the type of language they use. Given their considerable diffusion, it is important to train teachers and students on new AI tools, on their strengths and weaknesses, and on how they influence teaching and learning. If appropriately trained, students can use these tools critically and profitably even for less standard tasks such as linguistic research in mathematics. The limitation of not having control over the data and results makes it more difficult to use these tools for the design of educational linguistic activities, in which it would be more appropriate to use specially designed corpora.

These tools are changing mathematics education just as they will change the way of doing research in mathematics education. It is certainly important to also reflect on how these tools can influence students' language, both specialist and non-specialist, and on their linguistic and not just disciplinary skills.

References

1. Ferrari, P. L.: Costruzione di competenze linguistiche appropriate per la matematica a partire dalla media inferiore. *L'insegnamento della matematica e delle scienze integrate*, 26A, 4, 469-496 (2003).
2. Sbaragli, S., Demartini, S., Franchini, E.: Le difficoltà di comprensione e di gestione dei termini specialistici della geometria all'ingresso della scuola secondaria di primo grado. *La matematica e la sua didattica*, 29(1), 7-37 (2021).
3. T. Johns: Should you be persuaded: two examples of data-driven learning. *Classroom Concordancing: ELR Journal*, 1-12 (1991).
4. Corino, E., Onesti, C.: Data-driven learning: a scaffolding methodology for CLIL and LSP teaching and learning. In *Frontiers in education*, 4, p. 7) (2019).
5. Attardi, G.: Il Bello, il Brutto e il Cattivo dei LLM. *Mondo Digitale*, 1-16 (2023).
6. Zhai, X.: ChatGPT user experience: Implications for education. SSRN 4312418, 1-10 (2022).
7. Abedi, M., Alshybani, I., Shahadat, M. R. B., Murillo, M.: Beyond Traditional Teaching: The Potential of Large Language Models and Chatbots in Graduate Engineering Education. *Qeios* (2023).
8. Frieder, S., Pinchetti, L., Griffiths, R. R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., Berner, J.: Mathematical capabilities of chatgpt. arXiv:2301.13867 (2023).
9. Shakarian, P., Koyyalamudi, A., Ngu, N., Mareedu, L.: An Independent Evaluation of ChatGPT on Mathematical Word Problems (MWP). arXiv:2302.13814 (2023).
10. Corino, E.: Didattica delle lingue corpus-based. *Educazione Linguistica Language Education*, 8, 231-248, (2014).
11. Corino, E., Fissore, C., Marchisio, M.: Data Driven Learning activities within a Digital Learning Environment to study the specialized language of Mathematics. In: O'Conner, L. *IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, 167-176 (2022).
12. Barana A., Marchisio, M.: Analyzing Interactions in Automatic Formative Assessment Activities for Mathematics in Digital Learning Environments. In Csapó, B., Uhomobhi, J. *13th International Conference on Computer Supported Education (CSEDU 2021)*, vol. 1, 497-504 (2021).
13. Barana, A., Marchisio, M., Sacchet, M.: Effectiveness of automatic formative assessment for learning mathematics in higher education. In Domenech J., Merello, P., De la Poza, E. *7th International Conference on Higher Education Advances (HEAd'21)*, 1-8 (2021).
14. Corino, E., Fissore, C., Marchisio, M.: Adaptive Exercises and Formative Assessment for English Remedial Action. *Orchestration of Learning Environments in the Digital World*, 3-19 (2022).
15. Sbaragli, S., Demartini, S.: *Lingua e strutture dei testi scolastici di matematica*. Italmatica. Dedalo Editor (2021).
16. P. L. Ferrari, *Educazione matematica, lingua, linguaggi. Costruire, condividere e comunicare matematica in classe*, (2021).
17. OpenAI, ChatGPT, <https://chat.openai.com>, last access 2023/09/15.
18. Fissore, C.: Data driven learning activities in English on the specialized language of Mathematics in secondary school. In: Chova L. G., López Martínez A., Lees J. *14th International Conference on Education and new Learning Technologies (EDULEARN)*, 9326-9333 (2022).
19. Barbera, E. F.: *Linguistica dei corpora e linguistica dei corpora italiana. Un'introduzione*, pp. 1-113 (2013).

20. Cobb, T., & Boulton, A.: Classroom applications of corpus analysis. *Cambridge Handbook of English Corpus Linguistics*, 478–497 (2015).
21. Godwin-Jones, R.: Data-informed language learning. *Language Learning & Technology*, 21(3), 9-27 (2017).