On Charles Hermite's style

François Lê*

Preprint. February 2022

In 1899, historian of science Paul Tannery published a paper entitled "Stylometry: its origins and its present" in *Revue philosophique de la France et de l'étranger*, [Tannery 1899]. This paper was a critical review of a book by Wincenty Lutoławski which had appeared shortly before, and which aimed at establishing the writing chronology of Plato's works by studying (the evolution of) his style, [Lutosławski 1897]. Following a statistical approach, Lutosławski investigated textual data which were supposed to characterize this style, such as rare words, sentence length, or the mutual ratios of the numbers of nouns, verbs, adjectives, and adverbs employed by Plato. Before him, other scholars had already taken into account such textual statistics to determine the chronology of the platonic dialogues, but Lutosławski differentiated himself by proposing basic rules meant to ensure the soundness of this method, which he christened "stylometry."[1] Although Tannery expressed a guarded opinion on the quality of these rules, he still admitted that "stylometry would be invaluably helpful if it was grounded scientifically." Here he referred not so much to chronological matters as to authenticity questions and authorship attribution, for stylometry would allow "bringing to light the particular and multiple causes which create the overall impression left by the style of an author."[2]

More than a century later, thanks to the development of computers and of the statistical analysis of textual data (sometimes called lexicometry or textometry), stylometry has extended its field of action, and

*Univ Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5208, Institut Camille Jordan, 43 blvd. du 11 novembre 1918, F-69622 Villeurbanne Cedex, France.

[1] For a global presentation of Lutosławski and an analysis of his research in linguistics, see [Pawłowski 2008].

[2] "La stylométrie ne peut en effet que prétendre à mettre en évidence les causes particulières et multiples qui produisent l'impression générale que laisse le style d'un auteur." Further: "Que la stylométrie puisse rendre d'inappréciables services, si elle est scientifiquement justifiée et appliquée suivant des lois reconnues valables, cela va de soi." [Tannery 1899, pp. 161–162].

researchers have been reflecting on both its theoretical basis and its technical implementation. Among other kinds of results, semantic and syntactic specificities of literary writers, which were hard to detect to the naked eye, have been revealed; literary genres have been correlated to the over- or under-use of some parts of speech; and quantified approaches of the phenomena of rhythm and rhyme have renewed the analyses of poetic corpuses.[3] Such statistical techniques thus offer a particular way to address the thorny question of style, of which the difficulties and the resistance both to be theorized and turned into an univocal, operative category are notorious—an observation which does not mean that other highly interesting stylistic issues cannot be dealt with successfully.[4]

In the history and the philosophy of mathematics, the notion of style has also been tackled several times during the last decades. Without entering into details, let me just recall that most of these contributions proposed to characterize mathematical styles with the help of criteria linked to the manner of how mathematicians of the past thought of mathematical objects and correspondingly integrated them in their works, how their demonstrations were made with respect to certain values, methods, or disciplinary preferences, or how they included more or less examples in their publications.[5]

These proposals, with all their nuances, have obviously their own merits, and there is no question of diminishing or discussing them here. My intention in this article is to approach the issue of style (an indeed, that of Charles Hermite) by following the alternative path of stylometry.

Hence, both my focus and methodology are different from those of the cited historical and philosophical research. Quite paradoxically, they appear to be more literary and more mathematical, respectively—or, if one prefers, less mathematical and less literary.

I say more literary because my aim is not to dissect some of Hermite's proofs, nor to understand how a given theorem is stated with regard to some mathematical values, nor to account for particular disciplinary articulations in his work.[6] Rather, what will be scrutinized are the

[3]See for instance [Muller 1967; Brunet 1978] or, more recently, [Beaudouin 2002; C. Labbé and D. Labbé 2018]. I am indebted to Catherine Goldstein for having brought Valérie Baudouin's works to my attention.

[4]The literature on this topic is huge. Here I only refer to the introduction and the different contributions of [Himy-Piéri, Castille, and Bougault 2014], to [Herschberg Pierrot 2005], or to the interesting investigations on the figure of speech paradox presented in [Gallard 2019].

[5]See the synthesis [Mancosu 2009/2021] and its references, as well as [Rowe et al. 2010] and [Rabouin 2017].

[6]Moreover, it is not about establishing a Hermitian lexicon with the help of

words used by Hermite, especially those which are *not* directly linked to mathematical objects. Thus the attention will be brought on the words from the natural language which have a more functional role in the mathematical discourse: personal and demonstrative pronouns, conjunctions, non-technical nouns and verbs, and so forth. It is with such elements that I propose to describe one facet of Hermite's style. The other major facet which will be investigated is that of lexical richness, which will be evaluated from three different viewpoints: vocabulary extent, number and nature of the hapaxes, i.e. of words which are used only once by Hermite, and, to a lesser extent, vocabulary sophistication.[7]

I also say more mathematical, because the analysis will be constantly supported by statistical calculations and indicators, which will help quantify the description of Hermite's mathematical prose and, to a certain degree, objectify "the overall impression that leaves the style of [this] author," to borrow Tannery's words again. Of course, this does not mean that the statistical, computer-aided tool is able to provide a purely objective results: the human researcher remains present throughout the whole process, from the initial technical conventions to the very selection of the questions to be tackled, and to the interpretation of the given numbers. In particular, any blind reliance on such numbers will be excluded: to understand and put these numbers into perspective, the corresponding words will always be studied within their textual environment.

Hermite's style will be appraised from his published papers, more specifically from those which are technical in nature and written in French. Thus, starting from the texts gathered in his *Œuvres complètes*, two papers written in English and Italian have been dismissed, as well as texts such as addresses, obituary notices, and prefaces of books by other mathematicians.[8] This operation leads to a corpus of 186 texts, published between 1842 and 1901.[9]

---

statistical tools, as it has been made for Francis Bacon for instance, [Fattori 1980]. Among other recent research devoted to the vocabulary of scientists, see [Giacomotto-Charra and Marrache-Gouraud 2021].

[7] As will be seen, indeed, vocabulary sophistication has been harder to handle. I will explain why, and comment on what I tried to do.

[8] The *Œuvres* also contain a report written by Augustin-Louis Cauchy on a memoir of Hermite, which has obviously been excluded. Therefore the resulting corpus is almost the same as that studied in a prosopographic perspective in [Goldstein 2012].

[9] Here and in the rest of the paper, the word "text" will refer to the different items that are distinguished in the *Œuvres*, where several series of notes published in *Comptes rendus hebdomadaires des séances de l'Académie des sciences* have been fused. Moreover, the content of these texts is not strictly identical with that of the original publications: a number of misprints have been corrected, and Émile Picard,

To determine if the different textual data that can be measured in this corpus are actually characteristic of Hermite, a comparative corpus has to be considered. In the absence of a bigger set of texts which would be available and ready for the textometric treatment, the point of comparison that has been chosen is Camille Jordan: this French mathematician was more or less a contemporary of Hermite, shared a number of research topics with him, and produced a work whose dimensions are roughly similar to Hermite's. Selecting Jordan's papers from his *Œuvres complètes* in the same way as has been done for Hermite yields a corpus of 122 texts published between 1861 and 1920. Thus the temporal width of Jordan's corpus is the exact same as Hermite's, although it begins two decades later. Incidentally, these two decades correspond approximately to the age difference of the two mathematicians, since Hermite is born in 1822 and Jordan in 1838.

In accordance with what has been announced above, the investigation is divided in two main parts, which deal with the notion of lexical richness and with that of grammatical and lexical specificities, respectively—the definitions of these terms will be given in due time. Among other results, we will see that, compared to Jordan, Hermite possesses a higher lexical richness, both in consideration with the vocabulary extent and the hapaxes. Moreover, we will see that the two corpuses differ largely with respect to the use of some grammatical categories. In particular, such unbalanced grammatical distributions will serve to characterize Hermite's mathematical writing as a sort of narrative involving to a great degree the person of Hermite himself through the use of many personal pronouns, and of many specific verbs which describe the mathematical process in a lively way.

In the conclusion, I will come back on the very interest and on the soundness of the proposed approach. In particular, a number of issues which arise naturally will be made explicit: that of the non-synchronicity of Hermite and Jordan, but also that of the possible influence of mathematical domains on the question of style.

1. An overview

For the needs of TXM, the textometry software that has been used,[10] each text of the corpuses gave rise to one `txt` file, formatted with the conventions described in [Lê 2022]. Thus, apart from correcting misprints and standardizing a few words (such as the noun "Tchebichef," originally

the editor, occasionally deleted parts which he indicated to be inaccurate.

[10] Its technical presentation is given in [Heiden 2010].

present in different transliterated forms), the main preliminary operation consisted in deleting the content of every mathematical formula and replacing them by a mere symbol $*$ or #, if the formula was part of a paragraph or was a centered one, respectively. The software then spotted the words of every text, counted them, and associated them with their lemma (i.e. the entry which would correspond to the given word in a dictionary) and the grammatical category to which they belong.

These data form the ground for all the functionality of TXM: exhaustive searches and counts of (sequences of) words, lemmas, or grammatical categories, related to diverse queries (lemmas which belong a given grammatical category, words of which the lemma begins with, or contains, a given chain of characters, etc.); lists of concordances, which situate the results of such queries in their close textual neighborhood, and allow to sort them in a variety of ways; inspection of the whole texts as soon as needed; and other, more advanced tools, of which some will be used and presented later in this paper.

For Hermite, the preliminary treatment inventories 364,412 words counted with repetition, which correspond to 6,334 distinct forms (called the tokens) and 2,740 lemmas. Among these words, 19,542 symbols $*$ and 10,869 symbols # are to be found.

In spite of the lower number of his texts, Jordan's corpus counts more words, since they are 591,732 in number, distributed into 6,983 tokens and 2,852 lemmas. The mathematical substitutes are divided into 55,252 symbols $*$ and 6,714 symbols #.

In both cases, apart from mathematical symbols, the most frequent words are functional words of the French language: articles, conjugated forms of the auxiliary verbs *être* and *avoir*, pronouns, prepositions, usual adverbs, and punctuation marks. The first substantives in the lexicons of Hermite and Jordan are nouns of mathematical objects, which reveal at once some of the thematic predilections of our two authors: substitutions and groups for Jordan, equations, functions, and forms[11] for Hermite (see table 1).

As stated above, this angle of the textual analysis, related to the technical, mathematical words, will not be investigated further. Instead, and before delving into the issue of the lexical richness, let me consider table 2. It shows how the frequency of words and the contribution of the

[11] Just as in English, the French word *forme* can designate either the mathematical object (e.g. a quadratic form) and the aspect of something. Without entering into details, let me just remark that a great number of *forme* do refer to the mathematical objects which bear this name; consistently, the plural *formes*, which is more likely to refer to these objects, is also present is the given list.

Hermite	
Word	Frequency
équation	1,395
forme	1,368
fonctions	1,304
fonction	1,146
expression	942
nombre	905
degré	894
formes	871
racines	870
coefficients	834

Jordan	
Word	Frequency
substitutions	4,816
groupe	3,295
substitution	2,682
forme	2,299
nombre	2,052
ordre	1,951
lettres	1,915
variables	1,616
cas	1,425
système	1,372

Table 1: The ten first common nouns in Hermite's and Jordan's lexicons.

latter to the vocabulary are correlated—to make things clear, the term "frequency" designates the absolute number of occurrences of a word (or a lemma...) in a given corpus, and the term "vocabulary" refers to the set of all the tokens in such a corpus.

Frequency class	Words		Tokens	
$f \geqslant 1500$	182,590	50.1%	29	0.5%
$1000 \leqslant f \leqslant 1499$	20,064	5.5%	16	0.3%
$500 \leqslant f \leqslant 999$	30,795	8.5%	42	0.7%
$100 \leqslant f \leqslant 499$	71,586	19.6%	323	5.1%
$2 \leqslant f \leqslant 99$	57,332	15.7%	3,879	61.2%
$f = 1$	2,045	0.6%	2,045	32.3%

Table 2: Distribution of Hermite's lexicon into frequency classes. The given percentages are relative to the total numbers of words (364,412) and of tokens (6,334), respectively.

The very high frequencies[12] ($f \geqslant 1500$) take up the half of Hermite's corpus although they represent only 0.5% of the vocabulary; conversely, the hapaxes embody 0.6% of the word mass but almost one third of the tokens. Said differently, a handful of words are repeated tremendously, but the vocabulary is concentrated in the very low frequencies: almost all the entries of the vocabulary occur in the corpus with a frequency lower than 100. This phenomenon is quite general, and can also be seen in

[12]The number and the extent of the frequency classes of table 2 are arbitrary. They roughly follow the model given in [Kastberg Sjöblom 2002, § 2.3].

literary texts, with some nuances: for instance, hapaxes tend to represent a tiny part of such texts, but they contribute up to 30% – 45% of the vocabulary.[13]

The very high frequencies correspond to the two symbols that signal the existence of mathematical formulas, and to diverse functional words of the language such as punctuation marks, articles, conjunctions, and prepositions (*de*, *la*, *et*, etc.: "of," "the," "and"). The high frequencies ($1000 \leqslant f \leqslant 1499$), for their part, are mostly composed with pronouns, together with the four substantives *équation*, *forme*, *fonction*, *fonctions*, and the numeral *deux* ("two"). Verbs which are not auxiliary verbs begin to appear in the medium frequencies ($500 \leqslant f \leqslant 999$), with conjugated forms of *pouvoir* et *donner* ("can / to be able to," "to give"). Such verbs occupy more and more room as frequencies get lower; eventually, they represent about half of the hapaxes.[14]

Since I do not wish to make a detailed comparison with Jordan on this point, the analogous numbers which correspond to his corpus will not be provided, and I will just indicate that the ratios are very similar to those of Hermite, either for the distribution of words or for that of the tokens.

That said, two components of the notion of lexical richness are related to some of the numbers which have been given above for our two mathematicians. The first one concerns the number of the tokens in the corpuses, while the second one investigates more closely the category of the hapaxes.

2. Lexical richness

Two main aspects form the notion of lexical richness in its traditional meaning.[15] On one hand, one is interested in the numerical side of the matter exclusively, in the sense that the vocabulary and some of its subsets (typically, the hapaxes) are considered in respect to their size only: this aspect is usually referred to as vocabulary diversity. On the other hand, the semantic content of the vocabulary is at the core of the issue of sophistication, where one tries to evaluate the degree of refinement, or eccentricity, of the terms used by an author. Both of these aspects

[13] See the numbers given in [Kastberg Sjöblom 2002, § 2.3; Lebart, Pincemin, and Poudat 2019, p. 51], as well as the examples of *Les Misérables* and *Germinal* which we present below.

[14] The description being based on words, and not lemmas, this observation is to be linked to the fact that in French, there exist many more inflected forms of verbs than inflected forms of nouns and adjectives, for example.

[15] See [Muller 1977, p. 115].

are most relevant when they are put into a comparative framework: the extent of a vocabulary and the rarity of words are notions which need external references to be gauged. Moreover, even if vocabulary diversity and sophistication are often related to one another in practice, they must be clearly differentiated: in a given text, a writer might use a lot of different, yet completely banal words or, conversely, repeat some advanced terms over and over.

Before beginning the analysis, three preliminary remarks on our situation should be made explicit. The first one concerns the problem posed by mathematical symbols. Because of the text formatting that has been mentioned earlier, it is impossible to take into account the diversity of these symbols as they appear in the original publications of Hermite and Jordan. Therefore, even if it would be interesting to include them in our reflection, the lexical richness will be evaluated only on the basis of the words expressed in the natural language. Accordingly, in this whole section, the numbers of words and tokens will always exclude the symbols $*$ and $\#$.[16]

The second point to bear in mind stems from the very structure of the Hermitian corpus. Indeed, this corpus contains 70 letters or extracts of letters that have been published in journals at the time: they represent 38% of the number of texts and 26% of the number of words. On the contrary, only one such letter is to be found in Jordan's corpus—a letter which happens to be very short. A question, then, is to ascertain if and how the epistolary format may influence the lexical richness. Consequently, the case of the letters will be treated separately when needed.

The last remark concerns the difference between the sizes of Hermite's and Jordan's corpuses (which count 334,001 and 529,766 non-mathematical words, respectively). In fact, this well-known issue goes beyond our case study. It is rooted in the fact that the vocabulary extent of a corpus is not a linear function of the number of words: as new words are added to a text, its vocabulary grows too, but this growth is slower since a part of the adjoined words have already been used before. Hence, a crucial issue is to be able to confront, in a relevant way, the extents of the vocabularies of two corpuses whose sizes are markedly different.

[16]Actually, every calculation has been made twice, to see the possible influence of the mathematical symbols in the results. It turns out that, on the whole, these results are very similar.

2.1 Lexical diversity

In particular, comparing the ratios between the numbers of words and of tokens can be seen as a first indicator of lexical diversity, but it has to be refined in most cases. To put things into perspective, let me (naively) compare Hermite's corpus with two French novels of the second half of the nineteenth century, namely Victor Hugos's *Les Misérables* (1862) and Émile Zola's *Germinal* (1885).[17] A direct examination of the numbers of table 3 shows that even if *Germinal* possesses less words than Hermite's texts, it has more than the double of tokens and more than the triple of hapaxes: this configuration leaves no doubt on the fact that *Germinal* has a higher lexical diversity. The comparison with *Les Misérables* is of the same vein, although this case is a bit different: even if the numbers of words and of tokens are ordered in the same way as in Hermite's corpus, the diverse orders of magnitude of the data do seem to indicate clearly a higher richness for Hugo's book.

	Words	Tokens	Hapaxes
Hermite	334,001	6,332	2,045
Jordan	529,766	6,981	2,014
Germinal	211,379	14,458	6,305
Les Misérables	645,243	31,685	14,394

Table 3: Numbers of words, tokens, and hapaxes which are not the symbols $*$ and #.

More delicate is the comparison between Hermite and Jordan: the latter's texts contain more words and more tokens than Hermite's, but the numbers of tokens are relatively close to one another, and there is no inversion of their order compared to the sizes of the corpuses. This is a typical case where caution has to be taken: to what extent is the superiority of Jordan's vocabulary extent due to the bigger size of the corpus itself?

To answer such a question, several solutions have been proposed by researchers working in lexical statistics.[18] Among these solutions, the techniques of text shortening consist in estimating what would be the vocabulary extent of the longest of two (corpuses of) texts if, considering its frequency structure, its size was the same as the shortest one—for the

[17] The data presented in the following lines are those which I obtained with the help of TXM, by using the versions of these novels which are available on https://fr.wikisource.org.

[18] Apart from those which will be cited and used below, see the references given in [Lebart, Pincemin, and Poudat 2019, p. 50].

reasons which have been evoked above, acting by mere linearity is not seen as adequate. Here I decided to use the method founded on what has been called the coefficient of vocabulary partition, [D. Labbé and Hubert 1997].[19]

By reducing Jordan's corpus to the size of Hermite's, the corresponding calculations[20] give the numbers listed in table 4. They mean that, knowing the actual and complete structure of Jordan's corpus, one estimates that it would count 5,825 tokens if it was made of 334,001 words. This represents a relative difference of about 8% in comparison with Hermite.[21]

	Hermite		Jordan (reduced)
	Words	Tokens	Tokens
Letters	85,540	3,824	3,434
Non-letters	248,461	5,539	5,196
Whole corpus	334,001	6,332	5,825

Table 4: Expected vocabulary extents of Jordan's corpus, if it was reduced to the sizes of Hermite's whole corpus and of its two sub-corpusus made of the letters and the non-letters.

Moreover, the same observation holds when Jordan's corpus is reduced to the sizes of Hermite's sub-corpuses made of the letters and the non-letters, respectively. The relative differences change a little bit, however, since the one associated with the letters equals approximately 10%, while the other one is close to 6%. In particular, it is interesting that even though the epistolary format does seem to favor a higher lexical diversity, Jordan is still characterized with a lower diversity when compared to the non-epistolary part of Hermite's corpus.

To complete these observations, let me eventually remark that Hermite's letters do have a greater lexical diversity than his other publications. This can already be seen in the previous indicators (*via* an intermediate comparison with Jordan). But it is also possible to apply the shortening technique to these sub-corpuses: the one made of the non-letters would

[19]This method refines, in the case of corpuses with a certain vocabulary "specialization," the classical technique of Charles Muller based on a probabilistic, binomial model, [Muller 1977, ch. 20].

[20]Since the software TXM does not include such a shortening process, I proceeded to the calculations on my own.

[21]A mere linear process would have been way more violent, since the vocabulary of Jordan would have been estimated to 4,401 tokens, i.e. 30% less than Hermite's 6,332 tokens.

count 3,522 tokens, that is, 8% less than the letters.[22]

2.2 Hapaxes: numerical comparisons

The presence of a great number of hapaxes in a corpus is often considered as a mark of a high lexical diversity. In the case of the comparison between Hermite and Jordan, a remarkable phenomena is to be observed: even though the corpus of the former is shorter than that of the latter, it contains more hapaxes (see the data already given in table 3). However, to have a clearer picture of the situation, a possibility is, again, to shorten Jordan's corpus. In fact, when it comes to the hapaxes only, the method which has been used above coincides with a simple linear operation. Thus if Jordan's corpus were reduced to the size of Hermite's, it would count 1,270 hapaxes, which represent nearly 62% of Hermite's 2,045 hapaxes: from this viewpoint, Hermite's vocabulary appears as being much more diverse than Jordan's, which echoes the previous results.[23]

That said, an examination of the lists of the hapaxes of our mathematical authors reveals that the grammatical categories to which they belong are distributed differently. On both sides, a significant part of the hapaxes are just numbers (typically, these numbers stand for pages and years, and appear when Hermite and Jordan cite other publications). They are 175 in Hermite and 110 in Jordan. Furthermore, the quantity of hapaxes in Jordan is inflated by about a hundred of ordinal numeral adjectives written in the form *157*^{ème}. All these adjectives come from one paper where Jordan establishes a long enumeration of groups and synthesizes it with the help of phrases such as "*157*^{ème} *groupe*," "*161*^{ème} *à 163*^{ème} *groupes*," etc. [Jordan 1868/1869]. Reciprocally, and contrary to Jordan's corpus, that of Hermite possesses a large number of foreign words (603), of which 218 are hapaxes. Most of these foreign words are constituents of titles and extracts that are cited by Hermite; a handful of them correspond to Latin phrases that Hermite uses here and there. Such non-French terms thus also contribute to extend the hapax number of Hermite.

These hapaxes being neglected,[24] the remaining ones are almost

[22] By shortening Hermite's whole corpus to the size of *Germinal*, its theoretical vocabulary extent would be of 5,181 tokens, a number which is radically smaller than the 14,458 tokens of Zola's novel.

[23] Of course, this result derived from the hapaxes and that on the vocabulary diversity as presented in the previous subsection are not alien to one another, since the hapaxes contribute up to 30% of the vocabulary, in our cases.

[24] The case of the Latin phrases is obviously interesting in respect to the vocabulary sophistication. Their number, however, is too small to have any effect on the global

exclusively content words, i.e. nouns, verbs, (non-numeral) adjectives, and some adverbs. Hermite has 1,640 of them, a smaller number than the 1,808 of Jordan. The relative order of these numbers is thus the opposite of that of the total numbers of hapaxes; however, reducing Jordan's corpus to the size of Hermiet's leads to an estimated number of 1,140 hapaxes, which is markedly less than the 1,640 of Hermite.

Let me finally note that nearly 30% of Hermite's hapaxes come from his letters, whereas these particular texts represent 26% of the total mass of the words of the corpus. The letters, thus, are slightly richer in hapaxes that the rest of the corpus. To have a finer interpretation of these numerical observations, I now consider the meanings of the hapaxes more closely.

2.3 Semantic content of the hapaxes

An important preliminary remark is that a big part of the hapaxes are words which seem to be completely ordinary, and whose very low frequency is surprising at first sight. For example, the words *formée*, *parlant*, *rencontrés*, *essai*, and *Comparaison* ("formed," "talking," "met," "attempt," "Comparison"[25]) are hapaxes for Hermite. One should keep in mind, indeed, that the notion of hapax relates to the very graphical form of words, and not to the associated lemmas: *Comparaison* is not the same as *comparaison*, the verb *parler* ("to talk") actually appears 41 times in Hermite's texts in different conjugated forms, etc.

I will now focus on more particular hapaxes, which seem to characterize to a greater extent Hermite's personal writing.[26]

Here again, it is useful to distinguish the letters from the other publications of Hermite. Indeed, a certain number of hapaxes which come from the letters seem to be in direct connection with a special way of writing, where the marks of personal involvement and anecdotes are multiplied. In this respect, the most emblematic and most extreme text is a 1900 letter to Jules Tannery, which begins as follows—the hapaxes are in bold characters:

> Saint-Jean-de-Luz, **villa Bel-air**, 24 septembre 1900.
>
> Mon cher ami,

counts and comparisons of hapaxes.

[25]The French "formée" and "rencontrés" are past principles. The former is feminine singular, the latter is masculine plural.

[26]The following lines thus can be seen as a first view of Hermite's lexical sophistication, although the impression of such a sophistication is totally subjective.

> Je viens **dégager** ma **parole** et m'**acquitter** bien **tardivement**, il me faut l'**avouer**, de ma **promesse** de vous démontrer les formules concernant les quantités $\varphi\left(\frac{c+d\omega}{a+b\omega}\right)$ données dans mon **ancien** article *Sur l'équation du cinquième degré.*
>
> Le bon **air** de la **mer** m'a aidé à surmonter la **torpeur** qui **faisait obstacle** à mon travail ; j'en profite pour **échapper** aux **remords** de ma **conscience**, et, en pensant que vous avez sous les yeux cet article, j'aborde comme il suit la question.[27] [Hermite 1902, p. 13]

The following pages of the letter are more technical and contain almost no hapax. Such words appear again massively in the conclusion of the letter:

> Et nous **causerons** aussi d'autre chose que d'Analyse, nous **argumenterons**, nous nous **disputerons**. De ma **proximité** de l'**Espagne**, je rapporte des **cigarettes** d'**Espagnoles** ; si vous ne **venez** pas en **fumer** avec votre **collaborateur** d'aujourd'hui, votre professeur d'autrefois, c'est que vous avez le cœur d'un **tigre**. ***Totus tuus*** *et* ***toto*** *corde.*[28] [Hermite 1902, p. 21]

The two themes that are revealed by these hapaxes—that of the delay and the excuses associated with the epistolary answers, and that of the sociable chat on occasional topics—can be observed in other papers. For example, at the other tip of the Hermitian chronology, in one of his letters to Carl Gustav Jacob Jacobi on number theory:

> **Près** de deux années se sont **écoulées**, sans que j'aie encore **répondu** à la lettre pleine de bonté que vous m'avez fait l'honneur de m'écrire. **Aujourd'hui** je viens vous **supplier** de me pardonner ma longue **négligence** et vous exprimer toute la joie que j'ai **ressentie** en me **voyant** une place dans le recueil de vos Œuvres.[29] [Hermite 1850, p. 100]

[27] "Saint-Jean-de Luz, villa Bel-air, September 24 1900. My dear friend, I am coming to free my word and to fulfill, with, I must confess, a great delay, my promise of demonstrating the formulas on the quantities $\varphi\left(\frac{c+d\omega}{a+b\omega}\right)$ that I gave in my old article *On the equation of the fifth degree.* The good air of the sea helped me overcome the torpor which hindered my work; I take advantage of the situation to escape the remorse of my consciousness, and, imagining that you have this article before your eyes, I tackle the question as follows." The pages given in my citations refer to the pages in Hermite's and Jordan's *Œuvres complètes.*

[28] "And we will chat about other things than Analysis, we will argue, we will quarrel with each other. From my proximity with Spain, I bring back cigarettes of Spanish women; if you do not come smoking them with your colleague of today, your professor of the past, then you have the heart of a tiger. *Totus tuus et toto corde.*"

[29] "Almost two years have passed, and I have not yet responded to the letter, filled with kindness, that you made me the honor to write to myself. I am coming today to beg you to forgive my long negligence, and to express all the joy that I have felt by seeing for myself some room in the collection of your works."

Or, as he wrote to Eugenio Beltrami in 1881 about Domenico Chelini:

> Je n'ai point connu, **personnellement**, l'homme excellent et le géomètre si distingué dont vous **voulez honorer** la mémoire, mais j'ai **recueilli** l'**éloge** de son talent et de ses **vertus** de la **bouche** de votre éminent **compatriote** M. Brioschi.[30] [Hermite 1881a, p. 87]

However, the hapaxes of the letters contain many other terms which have nothing to do with such themes, and which deal with mathematical questions more directly. For instance, as he was commenting a result of Leopold Kronecker (about a certain function) in a letter to Joseph Liouville in 1862, Hermite declared:

> M. Kronecker, en la donnant comme l'expression analytique d'un de ses théorèmes, avait bien évidemment **pressenti** la signification qu'elle **recevrait** dans la théorie des fonctions elliptiques, et, à cet égard, je ne puis trop **admirer** la **pénétration** dont il a fait **preuve**.[31] [Hermite 1862, p. 120]

This kind of meliorative comments seems to be more present in the letters, which thus appear to encourage Hermite's personal expression.

Hapaxes which reflect such comments can also be found in the other publications, yet to a lesser extent: without citing the texts in which they are contained, let me note that terms such as *mystère*, *paradoxe*, *prestige*, *lumière*, *guide*, *inattendu*, *magnifiques*, and *stérile* ("mystery," "paradox," "prestige," "light," "guide," "unexpected," "magnificent," and "sterile") are examples of hapaxes which are associated with sentences where Hermite develops his viewpoints on his own works, on some of his colleagues', or on the mathematical objects, theorems, and theories themselves.

Finally, a non-negligible number of hapaxes are terms of a purely technical nature, which reveal some thematic specializations: *Émanants* is the name, proposed by James Joseph Sylvester, of objects of the theory of forms and invariants; the word *couronne* ("annulus") appears (together with two occurrences of the singular *couronne*) in a paper on Laurent series; and the *[points] stationnaires [d'une] quadrique* ("stationary [points of a] quadric") are just mentioned in the post-scriptum of a letter to Lazarus Fuchs on elliptic functions and, to a lesser extent, cubic curves.

[30] "I have not known personally the excellent man and the so distinguished geometer whose memory you want to honor, but I gathered the praise of his talent and his virtues from the mouth of your eminent countryman Mr. Brioschi."

[31] "Mr. Kronecker, by giving it as the analytic expression of one of his theorems, had obviously foreseen the sense that it would receive in the theory of elliptic functions, and, in this respect, I cannot but admire the insight he demonstrated."

On Jordan's side (which will be treated without providing the same amount of detail), a particularity is that the hapaxes comprise many more words which correspond to technical terms, and which are associated with topics to which Jordan devoted one or two papers in the corpus. It is the case of a memoir on the stability of floating bodies [Jordan 1867/1868], where the semantic field of navigation manifests itself through the intermediary of hapaxes such as *navires*, *émersion*, *submergé*, etc. ("ships," "emersion," "submerged"). Similarly, the theme of mountainous geography is developed in a paper called "On the lines of crest and thalweg", [Jordan 1872].

The hapaxes which refer to the expression of Jordan's personal viewpoints are much scarcer than in Hermite's case. In fact, most of them are linked to the lexical field of polemics, and come from the comments that Jordan writes during the 1874 controversy with Kronecker on bilinear forms[32]: the words *contradicteur*, *excusable*, *objective*, *incontestable*, *jugé*, and *complaisance* ("detractor," "excusable," "objective," "unquestionable," "judged," "complacency") are but a few examples of them.

Hermite's corpus, hence, is characterized with a higher lexical diversity, and this diversity comes in part from a greater number of hapaxes which, contrary to Jordan, concern as much the mathematical objects as the expression of the author's viewpoints, anecdotes, and ways of opening his letters.

2.4 On lexical sophistication

It is more difficult to me to draw solid conclusions about the lexical sophistication of Hermite and Jordan. One of the principal reasons is that the affectation of a word is a characteristic trait whose evaluation is linked, *a priori*, to a high degree of subjectivity, which is something I wish to avoid as much as I can. A possible way to bypass this problem is to connect it with the notion of rarity of use in corpuses which could be taken as representatives of the writing norms of a given time period. Unfortunately, as already stated, I do not possess enough mathematical texts of the nineteenth century for such a quantitative treatment yet.

Nevertheless, I would like to explain what I tried to do to tackle the problem, and what obstacle stood in the way.

The starting point was to consider the two components of the symmetrical differences of the lexicons of Jordan and Hermite, that is, the sets of the words which are used by one of them and not the other. Assuming that technical words do not contribute to the sophistication issue,

[32] On this controversy, see [Brechenmacher 2007].

I then removed them from these sets. Then, for each of the remaining terms, I noted how many times it was used in corpuses of reference which I constructed with the help of the online database Frantext, made of literary works mostly:[33] the idea was to assess how often the terms which are proper to Hermite or Jordan are used in the literary production of their time, in order to get rid of my own (anachronistic) impression of sophistication.

The experience turned out to be difficult to interpret. For instance, the verb *vaincre* ("to vanquish / to defeat"), which appears only in Hermite's texts, is common in the literary production of the period 1842–1901, but its occurrences in mathematical texts seem to be much more singular, and give a particular taste to the Hermitian writing, as in: "*Les formes de degrés* pairs *m'ont présenté de plus grandes difficultés, que dès longtemps je ne puis espérer vaincre.*"[34] On the contrary, a verb like *ensuivre* ("to ensue") is relatively rare in the literature of the time but quite usual in Hermite's corpus. The same phenomena can be observed in Jordan's case. To take but one example, the use of *condamner* ("to condemn") in the phrase: "*L'hypothèse dont nous venons de partir se condamne d'elle-même*"[35] sounds precious in a mathematical text even though the verb is really usual in novels, poems, and plays from the period 1861–1920.

As these examples show, the question of the lexical sophistication must wait until larger mathematical corpuses are ready to be investigated and taken as points of comparison, if one wants to treat it in a quantitative way.

Finally, no clear conclusion came out of a direct, non-quantified examination of the symmetric difference of Hermite's and Jordan's vocabularies, in particular because both of them seemed to possess advanced words, which could be substituted with one another. For example, to the adverbs *éminemment*, *hardiment* and *obscurément* ("eminently," "boldly," "obscurely") which appear only on Hermite's side, respond Jordan's *subsidiairement*, *prodigieusement* and *promptement* ("additionally," "prodigiously," "promptly"): it is delicate to tip the scales in favor of one

[33] In February 2022, Frantext counted 5,555 French references, for a total of 264 millions of words. I considered two comparative corpuses adapted to Hermite's and Jordan's periods of publication, in order to erase the possible effects of the time shift between these periods. The reference corpus for Hermite is composed of 696 texts and about 42 millions of words; the one for Jordan comprises 750 texts and 38 millions of words.

[34] "The forms of *even* degree presented greater difficulty, which I cannot hope to defeat since a long time." [Hermite 1856a, p. 351].

[35] "The hypothesis from which we just started condemns itself." [Jordan 1861, p. 151].

or the other side.

Thus I turn away, with terror and horror, from this lamentable plague of lexical sophistication, and come to the grammatical and lexical specificities of our two authors.

3. Specificities

To begin with, it may be helpful to explain on an example the meaning of the technical term "specificities." Let us suppose that we want to assess if Hermite uses markedly more adverbs than Jordan, taking into account the dissimilarity between the sizes of their corpuses. We know that there are 14,158 adverbs in Hermite and 25,032 adverbs in Jordan, but to compare these numbers by a simple linear reduction is not seen as a satisfactory solution.

One way to deal with this issue[36] is to consider the reunion of Hermite's and Jordan's corpuses: it contains many subsets whose cardinality is equal to that of Hermite's corpus, and, obviously, the latter is one of them. Then, supposing that the 39,190 adverbs are equidistributed in the reunion, one evaluates the expected number of adverbs in any such subset in the framework of a hypergeometric distribution. If the number of Hermite's adverbs is larger (resp. smaller) than this expected number, there is an over-representation (resp. under-representation) of adverbs in Hermite's corpus. In the process, a coefficient called the specificity score is calculated. It helps quantify the over- or under-representation, which correspond to a positive or negative score, respectively.

Naturally, the adverbs that have been taken for the example can be replaced by the results of any textual query: words, lemmas, grammatical categories, sequences of words, etc. In any case, it must be emphasized that the over-representation of a word in Hermite does not mean that it is not used by Jordan: it means that it is abnormally more used by Hermite than by Jordan, from a statistical point of view.

In this section, specificity calculations are used at several scales. Firstly, they serve to make a general comparison of the over- and under-uses of grammatical categories in our two corpuses: the results are presented in table 5. For instance, one sees in this table that proper nouns are over-represented in Hermite (and thus under-represented in Jordan) whereas common nouns are banal, or that Hermite favors centered mathematical formulas, while non-centered formulas proliferate in

[36]The model has first been proposed by Pierre Lafon [1980]. Among others, see the caveat explained on p. 137. The software TXM integrates the calculation of specificities.

Jordan's texts.[37]

At this stage, comparisons pertain only to the numbers of proper nouns, common nouns, etc., and not to the words that compose these grammatical categories. To help interpret such results, a possibility is to enter into details by inspecting both the (absolute) frequencies of the constituents of a given category and the specificities of these constituents within the category.

In general, stylometry suggests taking particular care of grammatical categories which correspond to function words, that is, words which are not nouns, verbs, adjectives, and adverbs: words which correspond to those categories of content words, indeed, would convey the actual content of a text and would thus take the researcher away from the stylistic inquiry.[38] In our case, however, to include content words in the analysis allows highlighting some writing features which are not of purely functional nature, and do not either relate to the actual mathematical content. Moreover, content words will sometimes be necessary to interpret correctly some over- or under-representations of function words; conversely, such imbalances of function words can reflect differences of mathematical content.

3.1 Proper nouns and marks of citations

Let me first very briefly comment on the case of proper nouns, which are clearly over-represented in Hermite. As has been explained elsewhere,[39] proper nouns of persons have different status in Hermite's *Œuvres*: they can be noun complements in the designation of mathematical objects, they can designate journals through the name of their editor, and they can refer to the people whom Hermite is writing to, to translators, or to authors of works upon which Hermite expands.

Citations, at least in the way in which they are formulated by Hermite and Jordan, seem to play a role in the imbalance of proper nouns, as

[37]The imbalance between the two types of mathematical symbols is in part due to the fact that Hermite's corpus include many papers of analysis, where formulas are larger and thus need to be centered. Without going further into detail here, I refer to [Lê 2022], where this point is discussed in another case.

[38]As is well known, however, content words can have a purely functional role in a text. This is obviously the case of some adverbs (such as "then," in English) or some nouns which are parts of fixed syntagmas, such as "as a result."

[39]See the second section of [Goldstein 2012], of which I take the results. The conventions used in the present paper leads to 1,842 proper nouns, which are represented by 232 tokens: this is more than in the given reference, which concentrates on the nouns of persons. Our 1,842 occurrences include places, institutions, and (abbreviations of) first names.

Category	Frequency	Frequency H.	Spec. score
#	17,583	10,869	1,000.0
Proper nouns	2,427	1,842	313.9
Pers. pronouns	44,630	20,153	211.6
Weak punct.	80,462	34,540	186.9
Verbs, present	44,722	19,587	139.1
Foreign words	764	603	116.8
Verbs, infinitive	15,774	6,815	39.2
Verbs, pres. part.	12,941	5,576	30.9
Prepositions	108,375	43,034	30.0
Verbs, past part.	17,560	7,297	20.7
Articles	78,598	31,159	19.7
Abbreviations	2,543	1,183	17.5
Symbols	414	234	13.6
Conjunctions	47,555	18,699	7.8
Quot. marks	54	36	4.7
Dem. pronouns	18,214	7,097	2.1
Verbs, sple past	20	10	0.7
Rel. pronouns	18,553	7,092	0.4
Poss. pronouns	18	7	0.3
Common nouns	144,114	54,849	−0.5
Prep. + det.	24,673	9,356	−0.6
Pronouns (other)	39	6	−2.7
Verbs, imperative	3,260	1,116	−5.6
Adjectives	61,534	22,917	−5.7
Verbs, impf.	1,149	331	−10.7
Verbs subj. impf.	178	19	−15.8
Adverbs	39,190	14,158	−16.2
Poss. adj.	4,619	1,220	−63.1
Verbs subj. pres.	5,932	1,488	−100.6
Verbs cond.	2,388	395	−117.4
Strong punct.	38,165	11,882	−184.9
Verbs sple fut.	18,474	5,032	−216.6
Numerals	16,592	4,318	−241.4
Indef. pronouns	9,836	1,950	−1,000.0
∗	74,794	19,542	−1,000.0

Table 5: Distribution and specificity scores of grammatical categories. The second column lists the frequencies in the reunion of Hermite's and Jordan's corpuses. The third one is related to Hermite only.

indicate other grammatical specificities too. Thus the overabundance of foreign words in Hermite, which we already mentioned, is mostly supplied by titles of cited works in German, Latin, Italian, or English, and by quotes of words and sentences written in these languages—this aspect is also revealed by the (lighter) over-use of quotation marks in Hermite. Moreover, the positive specificity of the category of abbreviations echoes such observations, and is due to a massive use of *M.*, *p.*, and *t.*, which stand for *Monsieur*, *page*, and *tome*.

In any case, the multitude of proper nouns has the effect of marbling Hermite's texts with the presence of individuals and collectives of all sorts, and thus brings to these texts a certain human color, which appears to be less bright in Jordan's corpus.

3.2 Common nouns, adjectives, and adverbs

Common nouns, adjectives, and adverbs are categories whose specificity scores are not high, compared to the others in table 5: common nouns, as stated above, are actually completely banal, while adverbs and, in a more minor way, adjectives are a bit under-represented in Hermite. However, the examination of the words composing these three categories, and especially the words that are specific to one author or the other, reveals interesting phenomena.

A great number of these specific terms are of technical, mathematical nature: it is the case of *polynôme*, *intégrale*, *elliptique*, and *doublement* ("polynomial," "integral," "elliptic," "doubly") for Hermite, and of *groupe*, *lettres*, *échangeable*, or *transitivement* ("group," "letters," "exchangeable," "transitively") for Jordan.[40] Since these words reflect mathematical topics themselves, I will disregard them in order to focus on features that are closer to the issue of style.

Some nouns, adjectives, and adverbs, although being specific to Hermite or Jordan, activate in fact the same meanings: for instance, Hermite is fond of the adverbs *immédiatement* and *facilement* ("immediately," "easily"), which certainly energize his writing, while Jordan prefers to use *évidemment* ("obviously") over and over: these adverbs do not seem to have really different senses,[41] and thus appear as mere personal preferences of our two authors.

Other specific terms create lexical fields that are only present on one side. Concerning Hermite, a whole set of nouns and adjectives reflects

[40] For the sake of brevity, I will not systematically present tables with frequencies and specificity scores when examining particular categories.

[41] Such assertions have systematically been supported by a direct inspection of the occurrences of the said words in their textual context.

the frequent expression of his personal views on the sequence of mathematical events, on objects, theorems, and works of the past: *méthode*, *recherche*, *facile*, *important*, *beau*, *essentiel* ("method," "research," "easy," "important," "beautiful," "essential") are but a few such specific terms through which the person of Hermite is made visible in the text, and of which there is no equivalent in Jordan. The corresponding semantic field is thus the same as that which has been already detected with the hapaxes.

The terms that are specific to Jordan and that are not directly linked to mathematical objects relate to the *reductio ad absurdum*, with the nouns and adjectives *hypothèse*, *absurde*, *inadmissible*, *contraire* ("hypothesis," "absurd," "inadmissible," "contrary").[42] As for the adverbs, the over-representation of the twin negative markers *ne* and *pas* seems to correspond to the same characteristic.[43] The proof by contradiction thus appears to be carefully avoided by Hermite. Examining the absolute frequencies shows it even more clearly, since *inadmissible* is never employed by Hermite and *absurde* is used four times only—*contraire* is a bit more frequent, with 81 occurrences of words having this lemma, but most these occurrences concern quantities of opposite signs or to the opposite types of monotony of functions.

3.3 Verbs, personal pronouns, conjunctions

The quasi absence of proof by contradiction in Hermite has also an impact on the specificities of verb tenses and moods. Indeed, as table 5 shows, the conditional, the subjunctive, and, to a lesser degree, the imperfect indicative are under-represented in Hermite's texts. But conditional and subjunctive are two ways to formulate hypotheses and their possible consequences,[44] which is in accordance with Jordan's predilection (compared with Hermite) of the proof by contradiction.

To carry on with the discussion on verbs, let me enumerate the first ones that are specific to Hermite, in their word form, and by decreasing

[42] If *contraire* appears both as a specific adjective and a specific noun, *absurde* is present only in its adjectival form.

[43] In terms of absolute frequencies, these two words inundate Jordan's adverbs, which explains why the category itself is over-represented in the latter. Moreover, the word *si* ("if"), which is the introducer of hypotheses *par excellence*, is among the conjunctions that are largely over-represented in Jordan's texts.

[44] Moreover, in French, the imperfect indicative usually accompany the conditional, when it comes to sentences that are introduced by *si*: "*Si le groupe était abélien, il serait résoluble.*" Jordan's over-use of *Supposons* ("Let us suppose"), which will be seen in a few lines, is to be linked to that of the subjunctive.

order of specificity: *ai*, *savoir*,[45] *conduit, tire, faisant, donne, supposant, vais, conclut, trouve, obtient, obtenir, écrire, observe, a, été, remarque, employant, parvenir, trouvera...* Conversely, for Jordan : *sera, contient, formé, contiendra, pourra, contenu, aura, Soient, contenant, déplace, Supposons, forme, être, existe, serait, déplacent, seront, transforme, succéder...*

Several lessons can be learned from the comparison of these lists. First, Jordan's specific verbs include more technical terms (in particular with the diverse forms of *contenir* and *déplacer*: "to contain," "to move"), while those on Hermite's side rather evoke the description of processes: *conduire*, *tirer*, *observer*, *écrire* ("to lead," "to draw," "to observe," "to write") as well as *[je] vais*, ("[I] am going to"). Hermite's mathematical narration, therefore, is marked out by such verbs which contribute to vitalize the action and to recall the involvement of a human person in charge of this action.[46]

Furthermore, the specificities of the personal pronouns used by Hermite and Jordan agree with this conclusion. Indeed, Hermite over-uses those which are linked to the first person singular, the (semi-)impersonal *on* ("one"), and the second person plural, which is associated with the French *vouvoiement*. Interestingly, if *vous* is a definite trace of the epistolary genre, it is not the case of *je*: the letters being excluded from the corpus, the first person singular is still overabundant in Hermite, whereas the second person plural disappears almost completely from the counts. On the contrary, Jordan makes considerable use of *il*, *elle* ("he / it," "she") and their diverse plural and reflexive declensions. About the *il*, it is helpful to delineate its different uses in Jordan's texts: almost non of them stand for a person, about a third represent mathematical objects, and the rest is assigned with an impersonal value in phrases such as *il y a*, *il faut*, and *il existe* ("there is / there are," "one has to,"“there exists / there exist").

These are the specificities of personal pronouns within their own

[45]This infinitive is exceptional in this list, for it is the only one to be employed almost exclusively within the fixed phrase *à savoir*, used as a synonym of *c'est-à-dire* ("that is / that is to say").

[46]On this point, see [Goldstein 2007, p. 398]: "More difficult to pinpoint, but quite characteristic, the flavour of Hermite's mathematical prose itself reminds the reader strongly of these French authors [Lagrange, Legendre, Cauchy, Fourier]. The style is discursive and oriented towards the description of processes." Moreover, the over-representation of *observer* may be linked to Hermite's predilection of observing formulas, although phrases such as *j'observe que* do not necessarily introduce reflections on what is observed, but rather serve as a way to state intermediary results in the course of a proof, for instance. On Hermite and observation, see [Goldstein 2011].

category, but table 5 shows that, when it comes to global numbers, Hermite over-employs these pronouns. This can be explained by the fact that Hermite's specific verbs, which have been listed above, cannot have mathematical objects as subjects, and are almost systematically associated with the pronouns *je* and *on*. No similar phenomenon seems to exist on Jordan's side: as mathematical objects are often the subjects of the verbs, some of them are represented by personal pronouns, but others are written as a noun or as a mathematical symbol, as in: *Ainsi, G contient...* ("Thus, G contains..."), and this makes the total number of personal pronouns decrease.

Finally, the lists of Hermite's and Jordan's specific verbs also reflect the global imbalance of the verb tenses that can be seen in table 5. The present indicative, as well as the infinitives and the past and present participles are overabundant in Hermite, while the simple future is Jordan's feature. At this point, it is perhaps useful to recall that in French, the simple future is one way, among others, to express the future; another one is to combine a conjugated form of *aller* with an infinitive (e.g. *je vais observer* and *j'observerai*: "I am going to observe" and "I will observe"). Hermite and Jordan both use these two ways of expressing the future, but the specificities of the tenses show that they use them in different proportions.

The over-representation of the simple future in Jordan is supplied in part by verbs that express mathematical facts: such simple futures have a gnomic value, as in "*Ce système ne contiendra donc en général qu'une fraction des substitutions du système primitif*" [Jordan 1861, p. 132], or "*Il est clair qu'une partie quelconque d'une ligne géodésique sera elle-même géodésique*"[47] [Jordan 1866]. On the other hand, the futures which are expressed with the verb *aller* and an infinitive are over-used in Hermite's texts, and are often associated with the first person (singular and plural). The verbs are in great part those of the description of processes; other examples than those which we already saw include: "*Cette remarque faite, je vais étudier de plus près les quotients...*" [Hermite 1856b, p. 381] and "*La notion de coupure se présente de la manière la plus simple dans un cas particulier que je vais maintenant considérer.*" [Hermite 1881b, p. 63].[48]

[47] "Thus, this system will generally contain only a fraction of the substitutions of the primitive system"; "It is clear that any part of a geodesic line will be geodesic itself".

[48] "This remark being made, I am going to study more closely the quotients..."; "The notion of cut presents itself most simply in a particular case that I am now going to consider."

Such an expression of the future colors Hermite's texts with a certain vitality, with an immediacy of the described mathematical action, which is also fueled by the use of the present indicative, and of the different participles.[49] In this respect, it is particularly telling that the present participle *supposant* is characteristic of Hermite, while the imperative *Supposons* is preferred by Jordan: both have the same meaning, of course, but the former is a trace of a prose which is more energetic than that conveyed by the latter. As for the present indicative, the same effects were already noticeable in the verbs that describe processes, and which we mentioned above.

The vigor of Hermite's writing can also be detected, although to a lesser extent, in the conjunctions which are specific to him. Among those that he tends to over-use, one can list *comme*, *et*, *lorsque*, *or*, *afin*, and *quand* ("as," "and," "when," "but / now," "so (that)," "when"); on Jordan's side, *donc*, *car*, *ou*, and *ni* ("so," "for," "or," "nor") prevail. In particular, Hermite favors subordinating conjunctions; those such as *quand* and *lorsque*, which are used relatively rarely by Jordan, are often parts of phrases such as *quand on ajoute*, *quand on remplace* or *lorsqu'on suppose* ("when one adds," "when one replaces," "when one supposes"). In general, such conjunctions could be replaced by *si* ("if"), but *quand* et *lorsque* have a temporal connotation which, again, evoke the dynamics of Hermite's prose.

3.4 Sentences, demonstrative categories, favorite phrases

But the overabundance, in Hermite's corpus, of the category of conjunctions itself is linked to another disequilibrium displayed in table 5: that between the weak punctuation marks, which are over-represented in Hermite, and the strong punctuation marks, which proliferate in Jordan's corpus. These clues point to a difference between the average length of the sentences written by Hermite and Jordan. Incidentally, this is easily confirmed and refined by considering the absolute numbers in question: about 10,247 "actual" marks of strong punctuation (among which 10,227 periods, 18 question marks and 2 exclamation marks) can be counted on Hermite's side, which represent the same number of sentences.[50] In

[49]These tenses and moods are still over-represented in Hermite when the different forms of *aller*, *être*, and *avoir*, which are associated with the future and the past participles, are not taken into account for the calculation of the specificities.

[50]This number is the result of the subtraction, from the total number of strong punctuation marks, of the numbers of abbreviating dots and of dots that are appended to numbers which numerate paragraphs and sections. This is an approximate manner of counting the number of sentences, which does not include a reflection on what is

relating this number to that of the words in the corpus, one finds that the Hermitian sentence counts about 36 words on average. The analogous estimations for Jordan yield a number of 23,473 sentences (almost the double of Hermite's), each of them having 25 words on average. Hence the Hermitian sentence is wider, and this width is supported with the over-representation of weak punctuation marks (mostly commas) and of conjunctions.

The category of demonstrative pronouns, for its part, is quite banal in terms of specificity. However, inside this category, the words *c'*, *cette*, *C'*, *cet*, and *Cela*[51] are over-used in Hermite. Inspecting the occurrences of these words within their context brings to light several phrases which Hermite seems to favor: *Cela étant*, *pour cela*, *c'est-à-dire*, and *à cet effet* ("That being said," "for this," "that is (to say)," "to this end") are examples of phrases which are commonly used by Hermite, and almost never by Jordan.[52]

Furthermore, the phrases *pour cela* and *à cet effet* are often preceded or followed by the characteristic verbs that we listed previously, which mark the liveliness of the process depiction: "*J'observe, à cet effet*," "*À cet effet, nous remarquerons que...,*" "*Je vais établir pour cela que...*" etc.[53] The *C'* is also typical of Hermite, who writes it four times more frequently than Jordan; it expresses the introduction and the presentation of the information in a very active way: "*C'est à ce même résultat que je dois parvenir en me plaçant dans la seconde hypothèse*" and "*C'est ce qui résulte immédiatement des expressions...*"[54] are examples of beginnings of sentences which energize the mathematical speech.

Speaking of beginnings of sentences, it is also possible to investigate those, made of two words, that are specific to Hermite and Jordan (see

a sentence. In particular, it is blind to the extreme cases of word-sentences such as "*Théroème.*" These cases are characteristic of Jordan but, since they are marginal from a numeric point of view, they do not change the proposed interpretation on sentence lengths.

[51]All these words can be translated by "this," "that," or "it." The words *c'* and *C'* are the elided forms of *ce* and *Ce*; on the contrary, the final *t* in *cet* appears when the following word begins with a vowel or a silent *h*.

[52]Another phrase whose Hermite has the quasi-exclusivity—and which does not involve a demonstrative pronoun—is *par conséquent* ("consequently"). It is used 318 times by Hermite and only 5 times by Jordan. More generally, it is surprising to see that some words or phrases are completely absent from one author or the other, although they seem to be absolutely commonplace. For instance, Hermite never employs the adverb *pourtant* ("yet")!

[53]"I observe, to this end,..."; "To this end, we remark that..."; "For this, I am going to establish that.."

[54]"It is this same result that I must reach by placing me in the second hypothesis"; "It is what immediately results from the expressions..."

Sentence beginning	Freq.	Freq H.	Spec. score
Cela étant	198	195	93.5
C'est	289	234	68.0
Or,	319	283	61.0
De là	88	80	31.0
Je me	56	52	21.4
J'observe	39	39	19.7
Effectivement,	45	43	19.1
Je remarque	40	39	18.3
J'ai	65	53	16.1
Ainsi,	82	62	15.8
On trouve	45	40	14.9
Maintenant,	33	32	14.8
Voici maintenant	29	29	14.7
⋮	⋮	⋮	⋮
D'autre	183	3	−24.9
Soit ∗	580	68	−28.1
Soient ∗	451	42	−29.4
En effet	621	74	−29.4
D'ailleurs	414	29	−33.0
Les substitutions	290	5	−38.9
On aura	566	33	−50.4
Le groupe	334	2	−50.6
Donc ∗	421	4	−61.2
Si ∗	824	11	−115.2

Table 6: Some beginnings of sentences made of two words, ordered by specificity index.

table 6). Somehow, they sum up a number of the results that have been described until now, as they clearly show some of the most striking differences between our two mathematicians. Indeed, while the beginnings of Jordan's sentences involve (or make us guess the involvement of) mathematical objects as subjects, the first person singular is apparent already in the attacks of Hermite's sentences, and is associated with some of the characteristic verbs and syntagmas that we already brought to light. The other ones that appear in this list, such as *De là*, *Effectivement*, *Maintenant*, and *Voici maintenant* ("From this," "Indeed," "Now," "Here [is] now"), are yet other testimonies of the liveliness of the Hermitian prose.

3.5 The case of indefinite pronouns

Among the few grammatical categories which are deeply unbalanced in table 5, the case of indefinite pronouns has not been examined yet.[55] Actually, this case seems to be quite particular with regards to the issue of style: even if indefinite pronouns are function words, their over-representation in Jordan appears to be explained by the very content of the mathematics he develops.

Indeed, the inspection of the absolute numbers of these pronouns reveals that the category is over-used by Jordan because of the massive presence of *toutes*, *une*, *tous*, *un*, *chacun*, etc. ("every," "one," "all," "each"), that is, because of pronouns which relate to wholes by referring to their constituents (whether these constituents are individuated or not). Now, the nouns which are most frequently associated with these pronouns are letters, systems, substitutions, and groups, and we saw that some of Jordan's specific verbs are the diverse conjugated forms of *contenir*. All these clues seem to indicate that the over-representation of indefinite pronouns is rooted in that fact that, in comparison with Hermite, Jordan has a deeper interest in questions which have to do with different kinds of sets, their subsets and their constituents.

4. Concluding remarks

Mathematical style has often been described as the manner of expressing or presenting mathematical truths, or mathematical facts: a theorem, a proof, a more or less coherent set of results, could thus be expressed algebraically or geometrically, rigorously or intuitively, in a set-theoretic way, by starting from axioms, by following a given method.

Such a position differs from the approach that has been adopted here, not so much on the fact that studying the style of a mathematician comes down to studying how the latter expresses himself or herself, but rather on the issue of what is the "expression" and, simultaneously, what is the object of this expression. Following the literary path, I tried to turn away from what is *a priori* linked to mathematical objects, results, and other disciplinary features, and to focus on the diverse facets of lexical richness, as well as on the specificities of certain categories of words which do not belong to the technical, mathematical lexicon—however, the examples of the indefinite pronouns and of the proof by contradiction showed that

[55] Another such category is that of the determinants, whose overabundance in Hermite seems to be more difficult to interpret. In particular, it is not clear to me whether its over-representation is to be linked to the deficit of indefinite pronouns.

there is no question of associating, without due process, content words with mathematical content, and function words with the expression of this content.

Several characteristics of Hermite's writing have been brought to light and quantified, especially through the examination of many words which embody the mathematical narration. Thus Hermite favors the use of the first person singular, which, to a large extent, is associated with many verbs in the present indicative that relate, step by step, the mathematical process as Hermite thinks of it: observations, remarks, deductions, and conclusions are closely tied to the person of Hermite, whose presence remains visible in many other places of the mathematical writing. The individual person of Hermite, further, is not the only human trace in his corpus, since many persons are, in one way or the other, summoned through the different proper nouns. Hermite's writing is also characterized by a remarkable lexical diversity. This diversity is particularly due to the many hapaxes used by Hermite, and which appear to be in part linked to the expression of the viewpoints of Hermite on mathematics, in part to the personal anecdotes and the sociable talk that are mostly visible in letters.

Many of these conclusions, as has been emphasized several times, are relative in nature and stem from the comparison with Jordan's corpus. Some of them can also be accepted as absolute results: the involvement of Hermite's person in his texts, the description of mathematical processes, or the particularities of the epistolary form. However, my point is that when it comes to the issues of style, such features are best interpreted when confronted with an external reference, for an important question is to ascertain to what extent an author stands out from others.

Inevitably, this raises many questions related to the decision of having compared Hermite to Jordan.

Jordan, indeed, is a mathematician whose period of activity began about two decades after Hermite's. Thus one naturally wonders whether features such as the higher personalization of Hermite's prose actually reflect a difference between two authors, or between two representatives of two generations. The latter hypothesis would accredit, or would be accredited by, the fact that as time would pass, mathematicians would favor more impersonal formulations. But several studies, involving a sufficiently great number of mathematical texts of different authors, would be necessary to validate such an hypothesis. It would also be interesting to confront Hermite with mathematicians who are be closer to him in time, and then to determine whether Hermite's stylistic features subsist or not.

Another difficulty would be to take into account the possible impact of the existence of different research themes in the works of the considered authors. Indeed, as has been showed in [Lê 2022], grammatical specificities of two corpuses associated with two disciplines can be interpreted in the light of different, collective writing practices: for instance, texts from the theory of algebraic surfaces are markedly richer in common nouns than texts from invariant theory, because of the different ways of writing geometry and invariant theory at the time.

In this perspective, it must be emphasized that the results that have been presented in this paper are completely blind to possible specificities which would be internal to Hermite's corpus, especially in regards with (sub-)disciplines. Since I have no room to develop this point, I will confine myself with a teaser on this issue. Hermite's papers can be grouped into several subsets, according to their classification in the *Catalogue of scientific papers*. Among these subsets, the largest ones (in terms of word number) deal with number theory, linear substitutions, equation theory, algebraic functions, other special functions, complex functions, and the foundation of analysis.[56] Then, for instance, when confronted to the other subsets, number theory is characterized by a certain over-representation of indefinite pronouns and verbs in conditional and simple future forms, and an under-representation in verbs to the present indicative. In other words, within Hermite's corpus, number theory has some of the specificities that Jordan has in comparison with Hermite.[57] The question of interpreting such results correctly and finely remains open for the moment.

The objective that had been set in this paper was to tackle the notion of style with the help of stylometry, in particular to quantify, even if imperfectly, the impression that the reader feels when reading Hermite's mathematical works. Are such reading impressions even relevant for the history of mathematics? My personal conviction is yes. I do believe that the favorite expressions of a mathematician, his or her writing peculiarities, and, more generally, the way he or she elaborates the mathematical narration, are part and parcel of his or her works, even if they are not directly connected with the mathematical core. Moreover, and for the very reason that such elements concern impressions of reading, it is necessary to construct an adequate methodological framework which allows accounting for them, by making as explicit as possible the decisions

[56]These keywords are parts of the corresponding sections of the *Catalogue*.

[57]The scores of the grammatical specificities observed in the case of the subsets of Hermite's corpus, however, are way lower than those relative to the Hermite-Jordan comparison.

that must be made, the different steps of the investigation, and its limitations. The approach that I followed here is one proposal among others which, by its statistical nature and its systematic inclination, appears to me as a good candidate to make progress in this direction.

References

Beaudouin Valérie (2002), *Mètre et rythmes du vers classique*, Paris: Honoré Champion (↑ 1).

Brechenmacher Frédéric (2007), "La controverse de 1874 entre Camille Jordan et Leopold Kronecker", *Revue d'histoire des mathématiques* **13** (2), 187–257 (↑ 15).

Brunet Étienne (1978), *Le vocabulaire de Jean Giraudoux, structure et évolution*, Genève: Slatkine (↑ 1).

Fattori Marta (1980), *Lessico del* Novum organum *di Francesco Bacone*, Rome: Edizioni dell'Ateneo e Bizzarri Roma. Two volumes (↑ 2).

Gallard Pierre-Yves (2019), *Paradoxes et style paradoxal : l'âge des moralistes*, Paris: Classiques Garnier (↑ 2).

Giacomotto-Charra Violaine and Marrache-Gouraud Myriam (eds.) (2021), *La Science prise aux mots : Enquête sur le lexique scientifique de la Renaissance*, Paris: Classiques Garnier (↑ 2).

Goldstein Catherine (2007), "The Hermitian Form of Reading the *Disquisitiones*", in Catherine Goldstein, Norbert Schappacher, and Joachim Schwermer (eds.), *The Shaping of Arithmetic after C. F. Gauss's Disquisitiones Arithmeticae*, Berlin: Springer, 377–410 (↑ 22).

— (2011), "Un arithméticien contre l'arithémtisation : les principes de Charles Hermite", in Dominique Flament and Philippe Nabonnand (eds.), *Justifier en mathématiques*, Paris: Maison des Sciences de l'Homme, 129–165 (↑ 22).

— (2012), "Les autres de l'un : deux enquêtes prosopographiques sur Charles Hermite", in Philippe Nabonnand and Laurent Rollet (eds.), *Les uns et les autres... Biographies et prosopographies en histoire des sciences*, Nancy: Presses Universitaires de Nancy, 509–540 (↑ 3, 18).

Heiden Serge (2010), "The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme", in *24th Pacific Asia Conference on Language, Information and Computation*, Sendai, Japan, 389–398 (↑ 4).

Hermite Charles (1850), "Extraits de lettres de M. Ch. Hermite à M. Jacobi sur différents objets de la théorie des nombres", *Journal für die reine und angewandte Mathematik* **40**, 261–278, 279–315. Œuvres I, 100–163 (↑ 13).

Hermite Charles (1856a), “Sur la théorie des fonctions homogènes à deux indéterminées. Premier mémoire”, *Journal für die reine und angewandte Mathematik* **52**, 1–17. Œuvres I, 350–371 (↑ 16).
— (1856b), “Sur la théorie des fonctions homogènes à deux indéterminées. Second mémoire”, *Journal für die reine und angewandte Mathematik* **52**, 18–38. Œuvres I, 372–396 (↑ 23).
— (1862), “Sur la théorie des fonctions elliptiques et ses applications à l’Arithmétique. Lettre adressée à M. Liouville.”, *Journal de Mathématiques pures et appliquées*, 2nd ser. **7**, 25–48. Œuvres II, 109–124 (↑ 14).
— (1881a), “Sur les fonctions $\Theta(x)$ et $H(x)$ de Jacobi (estratto di lettera al Prof. E. Beltrami)”, in Luigi Cremona and Eugenio Beltrami (eds.), *In memoriam Dominici Chelini. Collectanea mathematica*, Milan, Naples, Pise: Hoepli, 1–5, Œuvres IV, 87–91 (↑ 13).
— (1881b), “Sur quelques points de la théorie des fonctions (extrait d’une lettre à M. Mittag-Leffler)”, *Journal für die reine und angewandte Mathematik* **91**, 54–78. Œuvres IV, 48–75 (↑ 23).
— (1902), “Lettre de M. Charles Hermite à M. Jules Tannery sur les fonctions modulaires”, in, *Éléments de la théorie des fonctions elliptiques*, vol. 4, Paris: Gauthier-Villars, 294–302, Livre de Jules Tannery et Jules Molk. Lettre de 1900. Œuvres II, 13–21 (↑ 12, 13).
Herschberg Pierrot Anne (2005), *Le style en mouvement. Littérature et art*, Belin (↑ 2).
Himy-Piéri Laure, Castille Jean-François, and Bougault Laurence (eds.) (2014), *Le style, découpeur de réel*, Rennes: Presses universitaires de Rennes (↑ 2).
Jordan Camille (1861), “Mémoire sur le nombre de valeurs des fonctions”, *Journal de l’École polytechnique* **22** (cahier 38), 113–194. Œuvres I, 1–82 (↑ 16, 23).
— (1866), “Recherches sur les polyèdres”, *Journal für die reine und angewandte Mathematik* **66**, 22–85. Œuvres IV (↑ 23).
— (1872), “Sur les lignes de faîte et de thalweg”, *Comptes rendus hebdomadaires des séances de l’Académie des sciences* **74, 75**, 1457–1459, 625–627, 1023–1025. Œuvres IV (↑ 14).
— (1867/1868), “Mémoire sur la stabilité des corps flottants”, *Annali di Matematica Pura ed Applicata*, 2nd ser. **1**, 170–221. Œuvres IV (↑ 14).
— (1868/1869), “Mémoire sur les groupes de mouvements”, *Annali di Matematica Pura ed Applicata*, 2nd ser. **2**, 167–215, 322–345. Œuvres IV (↑ 11).

Kastberg Sjöblom Margareta (2002), “L’écriture de J. M. G. Le Clézio, une approche lexicométrique”, PhD thesis, Université de Nice, http://www.revue-texto.net/Corpus/Publications/Kastberg/Kastberg_LeClezio.html (↑ 5, 6).

Labbé Cyril and Labbé Dominique (2018), “Les phrases de Marcel Proust”, in, *Proceedings of the 14th International Conference on Statistical Analysis of Textual Data*, UniversItalia, 400–410 (↑ 1).

Labbé Dominique and Hubert Pierre (1997), “Vocabulary richness”, *Lexicometrica* **0**, http://lexicometrica.univ-paris3.fr/article/numero0/VocabRichness.pdf (↑ 9).

Lafon Pierre (1980), “Sur la variabilité de la fréquence des formes dans un corpus”, *Mots* **1**, 127–165 (↑ 17).

Lê François (2022), “La théorie des surfaces algébriques dans les *Mathematische Annalen* à l’épreuve de la textométrie (1869-1898)”, *Revue d’histoire des mathématiques* **29**, COMPLÉTER (↑ 4, 17, 29).

Lebart Ludovic, Pincemin Bénédicte, and Poudat Céline (2019), *Analyse des données textuelles*, Presses de l’Université du Québec (↑ 6, 9).

Lutosławski Wincenty (1897), *The Origin and Growth of Plato’s Logic: With an Account of Plato’s Style and of the Chronology of his Writings*, London, New York, Bombay: Longmans, Green, and Co. (↑ 1).

Mancosu Paolo (2009/2021), “Mathematical Style”, in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Winter 2021 Edition, https://plato.stanford.edu/entries/mathematical-style/ (visited on 01/25/2022) (↑ 2).

Muller Charles (1967), *Étude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille*, Paris: Larousse (↑ 1).

— (1977), *Principes et méthodes de statistique lexicale*, Paris: Hachette (↑ 7, 9).

Pawłowski Adam (2008), “Les aspects linguistiques dans l’œuvre scientifique de Wincenty Lutosławski”, *Organon* **37** (40), 149–176 (↑ 1).

Rabouin David (2017), “Styles in Mathematical Practice”, in Karine Chemla and Evelyn Fox Keller (eds.), *Cultures without Culturalism: The Making of Scientific Knowledge*, Durham, London: Duke University Press, 262–306 (↑ 2).

Rowe David E., Volkert Klaus, Nabonnand Philippe, and Remmert Volker (eds.), *Disciplines and Styles in Pure Mathematics, 1800-2000*, vol. 7, 1, Oberwolfach Reports (↑ 2).

Tannery Paul (1899), “La stylométrie : ses origines et son présent”, *Revue philosophique de la France et de l’étranger* **47**, 159–169 (↑ 1).