A Concurrent NLP-fMRI Approach to the Brain's Mathematical Network

Internship Thesis

submitted in partial fulfillment of the requirements for the degree of

Master in Cognitive Science

***

M2 CogMaster

at *École Normale Supérieure*

Written by

Samuel Debray

École Normale Supérieure Paris-Saclay

Under the supervision of

Prof. Stanislas Dehaene

Collège de France

INSERM–CEA

Cognitive Neuroimaging Lab

Year 2021 – 2022

Abstract

The progress of AI in the last few years has allowed great progress to be made both in computer science and neuroimaging. Capitalising on techniques used to compare AI and the brain on natural language processing tasks, this thesis presents our attempts to use AI to understand the cognition of advanced mathematics.

The project had three strands. Firstly, we gathered a mathematical corpus from Wikipedia to generate a mathematical vocabulary and its semantic embeddings using the GloVe algorithm. Secondly, we used GPT-fr, a pretrained model of the Transformer, to analyse advanced mathematical and general knowledge statements. We showed that GPT-fr makes a distinction between meaningful and meaningless statements, and that its behaviour correlates with that of non-mathematicians when presented with meaningless mathematical statements. Thirdly, we found that the first principal component of the GloVe embeddings of a global corpus made up of mathematical and non-mathematical pages enables to retrieve the fMRI activations in the brain's mathematical network.

However we conclude that, although it is endowed with great mathematical abilities, AI's handling of mathematics is very different from ours, and that there is still room for improvement regarding AI's contributions to the understanding of mathematical cognition.

Keywords. fMRI, NLP, GloVe, Transformer, GPT-fr, mathematical cognition

Approximate word count. 13,941 words

Acknowledgments

First of all, I would like to thank my supervisor, Stanislas Dehaene, not only for having hosted me in his lab and avdised me throughout my internship, but also for having fed my curiosity and appetite for cognitive science. I would particularly like to thank Christophe Pallier, for having been my office partner and taught me the vary basics of statistics I somehow managed to know nothing about. Thanks also to Alexandre Pasquiou and Mathias Sablé-Meyer for having shared their time and wisdom with me.

As this project relies on those of Antonio Moreno and Marie Amalric, I would like to thank them for having guided me through their data: Marie and I have spent quite a few hours on zoom looking for her behavioural data.

As an intership could not go well without human contact, I would like to thank all the people with whom I have shared time, discussions, lunches, croissants or coffees at NeuroSpin: Maxime Cauté, Alexis Thual, Théo Morfoisse, Antonio Moreno, Marie Pallu, Isabelle Denghien, Vanna Santoro and Manon Pietrantoni.

Finally, I would like to thank Mihaela Sighireanu, head of the Computer Science department of École Normale Supérieure Paris-Saclay, for having supported my reorientation project in cognitive science, and Salvador Mascarenhas for his guidance and advice on my projects in linguistics and psychology of reasoning.

Contents

Declaration of Originality

Mathematical performance of AI and AI's ability to predict natural language processing in the brain are two very prolific strands of the literature in computer science and cognitive science. However, to the best of our knowledge, no work has been done to assess AI's ability to predict the cognition of advanced mathematics. This internship took a first step in this direction.

We used common tools and techniques from the literature (clustering, use of AI models to predict brain activity, etc.) but put them to use for an new purpose, in which the originality of our work lies.

Stanislas Dehaene aims at building upon the bases laid during this internship in his next research project on mathematical cognition.

Declaration of Contribution

Definition of the resarch questions. Result of dialogues between Stanislas Dehaene and myself.

Literature review. Myself, with many suggestions from Stanislas Dehaene.

Choice of methodology. Several tools were proposed by Stanislas Dehaene, Christophe Pallier and Alexandre Pasquiou. The final decisions were made by myself.

Programming. The programming part was made by myself, with occasional contributions of Alexandre Pasquiou and Christophe Pallier. The code for spectral clustering and two-dimensional plot of clusters, was translated into `Python` from Fransisco Pereira's `Matlab` code.

Design of fMRI experiments. Marie Amalric, Antonio Moreno and Stanislas Dehaene.

Subjects recruitment and testing. Marie Amalric, Antonio Moreno and Stanislas Dehaene.

Data analysis. Myself, with the help of Alexandre Pasquiou and Christophe Pallier.

Interpretation of the results. Result of dialogues between Stanislas Dehaene and myself, sometimes with Christophe Pallier too.

Redaction of the thesis. Myself.

Proof-reading of the thesis. Christophe Pallier.

Pre-registration

0.1 Administrative information

The internship will be conducted at NeuroSpin's Unicog team (INSERM – CEA Paris-Saclay) under the supervision of Stanislas Dehaene. Other referees include Alexandre Pasquiou, Mathias Sablé-Meyer (PhD students), Christophe Pallier and Antonio Moreno.

0.2 Introduction

Background. Dehaene and Amalric [AD16; AD18; AD19] provided conclusive evidence for the existence of a mathematical network in the brain, which activates when subjects are asked mathematical questions, no matter the difficulty of said questions. In parallel, natural language processing tools like GloVe [PSM14] and the Transformer architecture [Vas+17] like GPT or BERT proved to be great tools to model words' semantics. For instance, Pereira et al. [Per+18], used 300-dimensional GloVe vectors to provide a semantically-clustered map of the 1,000 most common English words. Our project is to combine fMRI and NLP methods to better understand how mathematics are processed in the brain.

Research questions. We shall try to answer three questions.

(i) How do models of the Transformer like GPT-fr and CamemBERT *understand* mathematics, as compared to human subjects?
(ii) Using the GloVe embbeding of a mathematical vocabulary, can we cluster mathematics in an interpretable way, as did Pereira et al. with language?
(iii) Do the GloVe embeddings of mathematical stimuli enable us to build good predictors of fMRI data in various regions of the brain?

Two types of models for NLP. We will be using both the Transformer and GloVe to perform NLP analysis. We will be using the Transformer to get a score, called perplexity, for an input of any size: this will enable us to analyse behavioural data. Furthermore, we will be using GloVe to perform a word-by-word analyses, which will be more convenient when it comes to analysing fMRI data, for which we have onsets for words. If time allows, we could also try to fit fMRI data with the Transformer's hidden states as well, but this could require a lot of work as far as tokenising is concerned, as well as assuming that the Transformer makes a good prediction of subjects' behaviour.

0.3 Methods

0.3.1 Tools and softwares

All the code will be written in `Python` 3.9. The analysis of fMRI data will be conducted using the `nilearn` library and the Transformer-NLP will be done using the `transformer` library and HuggingFace[1].

Statistical analyses (ANOVAs, regressions, clustering) will also be done in `Python`, using both the `scikit-learn` and the `statsmodels` libraries.

We will be using experimental data (stimuli and brain fMRI pictures) of Marie Amalric (three experiments) and Antonio Moreno acquired in Unicog.

0.3.2 Data analysis

Transformer's model selection. Using HuggingFace, we selected two models of the Transformer trained on a large dataset in French: CamemBERT (fill-mask) and GPT-fr (text generation). We will compare these two models based on the following criteria:

- *tokeniser*: we would like the tokeniser not to cut words, that is we want the ratio $\frac{\text{number of tokens}}{\text{number of words}}$ for each sentence to be as close to 1 as possible, as we do not believe the brain computes surprisal on sub-words;
- *perplexity*: if both models perform equally well on the tokenising task, we will select the model whose perplexity on the stimuli is the lower.

Model's behaviour analysis. Once we will have selected a Transformer's model and obtained its perplexity on each stimulus, we will try to understand how well it allows to discriminate between true and false sentences and meaningful and meaningless sentences. To do so, we will perform two-way ANOVAs whose dependent variable will be the network's perplexity and whose independent variables will be the category of the stimuli (e.g. algebra, geometry, etc.) and the parameter of interest (truth value or meaning).

Comparison between the Transformer and human subjects. So as to compare the Transformer with human subjects, we will collect behavioural data from Amalric and Moreno's experiments and plot, for each stimulus, the Transformer's perplexity against the percentage of human subjects who judged the stimulus was not true. We will then fit a linear regression to see if the Transformer's and subjects' behaviour are correlated.

Mathematical vocabulary. To obtain a mathematical vocabulary, we will scrap all Wikipedia French pages, decide which one are maths-related based on their content, and extract their content. The two (mathematical and non-mathematical) corpora will be preprocessed by removing all punctuation, capital letters and by replace each word with its lemma using `Python`'s `spacy` library. We will then use GloVe to obtain a 50,000-word vocabulary for both mathematical and non-mathematical corpora, which we will then use to create a mathematical vocabulary. To do so, we will compare, for each word w, the frequency $f_{m,w}$ of w in mathematical pages and its frequency $f_{nm,w}$ in non-mathematical pages. If $\log_{10}(f_{m,w}) - \log_{10}(f_{nm,w}) \geq 1$ we will consider w more frequent in mathematical pages and add it to a temporary vocabulary. The final mathematical vocabulary will be obtained from the temporary one by sorting words by number of occurences in maths pages and taking the 1,000 first words of the sorted vocabulary. As a sanity check, we will

[1] `https://huggingface.co/`

manually verify that all the obvious words (numbers, arithmetic operations, geometric shapes, etc.) are present in the vocabulary.

The final non-mathematical vocabulary will consist of the 1,000 most frequent words of GloVe's non-mathematical vocabulary.

Clustered map of mathematics. Once we will have obtained the final vocabularies, we will relaunch a GloVe analysis on the global corpus (mathematical plus non-mathematical) to get all vectors in the same vector space. We will then use spectral clustering [Lux07] on the 300-dimensional GloVe vectors of the 1,000 words of the mathematical vocabulary to obtain approximately 200 clusters, which we hope will be easy to interpret[2]. We will finally project the clusters in two dimensions using tSNE, plot centroids and add Voronoi boundaries to obtain a semantic map of mathematics.

We will then do the same with the final non-mathematical vocabulary to try to replicate Pereira et al.'s semantic map [Per+18] in French and on the 2,000-word global vocabulary (mathematical plus non-mathematical) to see how mathematics cluster within the general language.

Principal Component Analyses of vocabularies. We will perform PCA on GloVe vectors for each of the three vocabularies (mathematical, non-mathematical and global). For the mathematical vocabulary, we will try to see if the first components account for the differences between branches of mathematics (e.g. algebra, geometry, analysis). For the global corpus, we will try to see if mathematics are one an interpretable principal component. For the non-mathematical corpus, we will seek to replicate [Hut+16] in French.

fMRI data analysis. The question we seek to answer is to see if NLP helps us predict fMRI data. We will fit three different GLM and compare them two by two (3 comparisons in total) using brain maps showing the voxel-wise difference of two models' R coefficients (the second-level analysis will consist in a one-sample t-test accross subjects). The first model (the naive one) will only contain one categorical predictor by category of sentence (e.g. analysis meaningless, topology false, etc.). The second model (the medium naive one) will contain three predictors consisting of the mean, for each sentence, of the projection of the GloVe vectors of its words onto the three main principal components for the global model. The third model (the least naive one) will contain the three predictors of the second model in addition to three other predictors consisting of the mean of the projections onto the principal components for the mathematical model.

0.3.3 Statistical significance tests conventions

For fMRI data. All brain activation results will be reported with a cluster-wise threshold of $p < 0.05$ corrected for multiple comparisons across the whole brain, using an uncorrected voxel-wise threshold of $p < 0.001$.

ANOVAs and correlations. For ANOVAs and correlation tests, we will be rejecting the null hypothesis when $p < 0.05$.

[2] [Per+18] gives us good reasons to hope since we are replicating their work with mathematics instead of common language.

CHAPTER 1

Introduction

1.1 Motivations

In the last few years, the field of Artificial Intelligence (AI) has been the subject of great upheaval. The development of state-of-the-art neural networks like the Transformer has made it possible to build bridges between algorithms and the brain. Computer scientists have also realised that AI models are endowed with great reasoning abilities and are now able to beat human players at Go and to automatically prove mathematical theorems.

A fair amount of research has been conducted on the links between AI and the brain's treatment of language and on AI's mathematical abilities. Suprisingly, though, virtually nothing has been done on AI and the cognition of advanced mathematics. This project is a first step in this direction, capitalising on the recent discoveries on mathematical cognition made by Amalric and Dehaene. By looking in this direction, one aims at understanding better the way advanced mathematics are processed in the brain. Ultimately, this could lead to the emergence of artifical scientists, capable of performing scientific research on their own or, at the very least, to help scientists in the most tedious parts of their job; all by imitating the human brain. In short, the mutual benefits of AI and neuroscience are potentially immense.

1.2 Literature review

AI and its relations with mathematics or neuroscience has already been widely studied. This section reviews three strands of the literature: mathematical cognition – without AI – on the one hand, and the use AI for mathematics and neuroscience on the other hand.

1.2.1 The brain's mathematical network

This section presents the findings of Amalric et al. [AD16; AD18; AD19] on the cognition of advanced mathematics. As these works are the building blocks of the present project, Amalric et al.'s results are presented in some detail, but the main information is given in the first paragraph.

While many previous works focus on the processing of numbers [Par+12; Das+13; Shu+13; Ege16; Dai+16] and symbolic mathematics [Mar+12; MPO12], the work of Amalric et al. is, to our knowledge, the only one to tackle a semantic network of advanced mathematics in the brain.

Main findings. Amalric et al. [AD16; AD18] have idenfied areas in the brain which activate during the processing of mathematics using fMRI imaging. They showed that these regions always activate in mathematically-educated subjects, no matter the estimated difficulty or the kind of mathematics involved (e.g. algebra, topology, etc.). However, these regions do not activate when subjects are presented with general-knowledge statements or in mathematically-uneducated subjects who cannot understand the mathematical statements they are presented with. For the latter subjects, the regions involved in the processing of advanced mathematical statements overlap with those involved with the processing of meaningless non-mathematical statements in both groups of subjects. Furthermore, the areas identified by Amalric et al. activate both during the listening period of the mathematical statements and their processing during a reflection period. The areas identified by Amalric et al. are shown on figure 1.1. To further investigate these regions, Amalric et al.

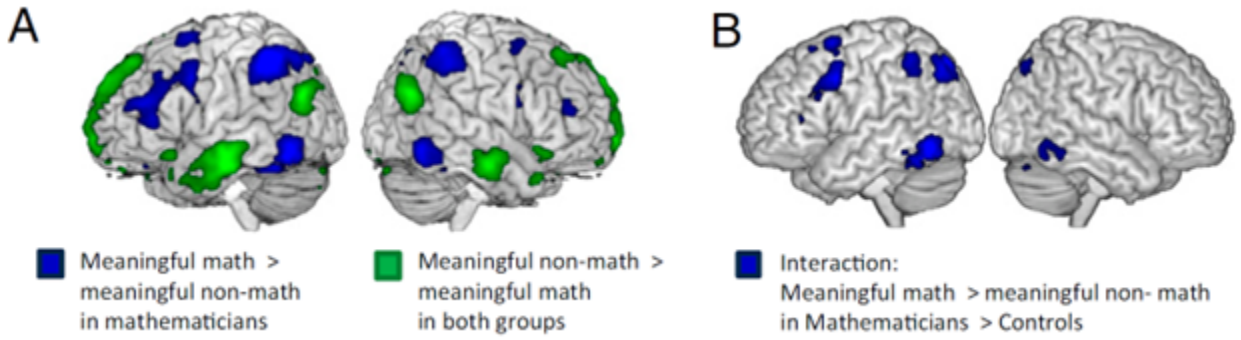


Fig. 1.1. Distinct brain areas for mathematical expertise and for general semantic knowledge. **(A)** Whole-brain view of areas activated during reflection on mathematical statements (blue) versus general knowledge (green). **(B)** Mathematical expertise effect: Interaction indicating a greater difference between meaningful math and nonmath statements in mathematicians than in controls. Maps are thresholded at voxel $p < 0.001$ and cluster $p < 0.05$ corrected for multiple comparisons across the brain volume. Extract from [AD16].

[AD19; AD18] designed two additional fMRI experiments. In the first one, they asked mathematician subjects to judge the truth value of mathematical statements which were much simpler than those used for their first paper [AD16]. The aim was to verify that very simple mathematical statements (some only using rote memory like $(a+b)^2 = a^2 + b^2 + 2ab$) activate the mathematical regions. This experiment replicated the results of that published in [AD16], proving the robustness of Amalric et al.'s findings. As the mathematical network is disjoint from language areas, Amalric et al. designed a second experiment to try to see if combinatorial operations such as quantifiers and negation make the processing of non-mathematical statements shift from language areas to mathematical areas. Interestingly, they found that adding combinatorial operations does not elicit the mathematical network.

Regions of Interest (ROIs). The ROIs identified in [AD16] include the bilateral intraparietal sulci, bilateral inferior temporal regions, bilateral dorsolateral, superior, and mesial prefrontal cortex, and cerebellum. Altough they found that adding combinatorial operations does not elicit the mathematical network, Amalric et al. [AD19] found that negation correlates with activation in the left inferior frontal gyrus (usually associated with syntactical complexity) and quantifiers correlate with deactivation in the right angular gyrus with no overlap with the parietal activation associated with mathematics.

Difference between listening and reflection periods. During listening, Amalric et al. found two additional activations compared to the reflection period in their first experiment [AD16]: one in

the bilateral head of the caudate nucleus, indicating subjects' prefered topic (mathematics for mathematicians and general knowledge for non-mathematicians). The other, in the left angular gyrus, deactivated during the processing of meaningless statements – whether they be mathematical or non-mathematical in mathematicians but only for non-mathematical in non-mathematicians – compared with meaningful ones.

1.2.2 AI and mathematics

The use of AI for mathematics has been explored a lot and from various perspectives. In the literature, three different purposes emerge: conjectures making, theorem proving and numerical calculation. Most works focus on neural networks and in particular on the Transformer [Vas+17].

Conjectures making. One of the possible uses of AI in mathematics is to guide intuition. This approach, illustrated in [Dav+21] for topology and representation theory, consists in using AI to make conjectures thanks to machine learning techniques, which then only have to be proved by mathematicians. Several models have been developed for this purpose. For instance Lample and Charton [LC19] designed a method to generate datasets made up of couples of functions and their primitive and ordinary differential equations of first or second order and their solution. They then trained a Transformer's model on this dataset and showed that the model thus trained performs better than popular frameworks like `Maple` and `Matlab`. Another example is symbolic regression, that is the task of predicting a function from its values. D'Ascoli et al. [dAs+22] trained two models of the Transformer, one for integer sequences and one for float sequences and compared their ability to predict the next terms of a recurrent sequence. They showed not only that their models outperformed Wolfram `Mathematica`'s built-in functions, but also that they are robust enough to accurately predict sequences which use functions they were not familiar with (e.g. $u_n = \tanh(n)$ when the model is only trained with circular trigonometric functions) or whose input terms are noisy. In a later paper, Kamienny et al. [Kam+22] trained a Transformer's model to predict the full mathematical expression of a sequence given its first terms. While the traditional way of adressing this task is to determine a parametric function first (e.g. $x \mapsto \cos(ax+b)$ with a, b to be determined) and then to estimate the parameters, they proposed a so-called "end-to-end" approach which skips the parametric function step and directly predicts the function. They showed that their approach is more efficient that the parametric function one and that it is robust enough to make accurate predictions even with noisy input.

Theorem proving. Polu et al. [PS20] developed a proof assistant, GPT-f, based on the architercture of the Transformer and the `Metamath` meta-logic, which is a formal environment relying on unification – while most environments, like that of `Coq` or `Agda`, rely on natural deduction or sequent calculus – which makes it fast and low-level. They trained their model on a set of pairs $\langle$goal, proof step$\rangle$ and showed that pretraining the model on the `WebMath` dataset, made up of mathematical ressources like arXiv and StackExchange improved its performance as well. The GPT-f model thus trained, although it not always able to complete proofs on its own, is very efficient at automatically generating low level steps which are ubiquitous in proof assistants. In a later paper, Polu et al. [Pol+22] proposed a slightly different approach consisting in training the model on its own trajectories to improve performance (while the training dataset in their previous paper was pre-generated), provided the initial trajectories are obtained from a varied enough set of problems. They thereby trained a model which is able to solve challenging problems for high school olympiads.

Numerical calculation. While most works focus on using AI for symbolic mathematics (e.g. exact integral computation or proof generation) Charton et al. [CHL21] proved that the Transformer is also able to perform computations as well. Following Lample et al. [LC19], Charton et al. trained a

Transformer's model on a dataset consisting or pairs ⟨problem, solution⟩ where the problem was related to differential systems and their stability. Even though it did not have built-in mathematical knowledge, the model learnt from examples and was eventually able to predict both qualitative and quantitative properties of differential systems and partial differential equations whose behaviour cannot be described analytically. Charton et al. [Cha21] replicated this work with computational linear algebra, proving that the same method can be used to train a Transformer's model to compute not only 5×5 random matrices' transpose or product, but also eigenvalues and inverse, even when the characteristic polynomial is not analytically factorisable.

Refinements of the Transformer for mathematics. In addition to the ad hoc models above mentioned, pretrained models specifically designed for mathematics are becoming available on HuggingFace. For instance, MathBERT [Pen+21], a Transformer's model based on BERT shows good performance at understanding mathematical formulae typeset in LaTeX in their context (e.g. in the mass-energy equivalence $E = mc^2$, c stands for the speed of light in vacuum) and can be loaded and used as is. Similarly, Wu et al. [Wu+22] proposed a refinement of the architecture of the Transformer, the so-called "memorising Transformer", which is able to memorise new data at inference time, without the need to be trained, by adding an external memory to the original architecture. This new architecture allows for improvements in perplexity and does not even need to be trained from scratch: a trained Transformer's model can be adapted just by providing it with the memory mechanism.

One notable exception to the use of the Transformer for mathematics is the Ramanujan Machine [Raa+21]. The Ramanujan Machine uses optimisation-based gradient descent and meet-in-the-middle techniques to conjecture mathematical equations involving fundamental constants like $\pi, \zeta(3)$ or the Euler-Mascheroni constant γ. This method has proved very powerful and generated many conjectures of the form

$$\frac{\gamma(c)}{\delta(c)} = f_i\left(\alpha(0) + \cfrac{\beta(1)}{\alpha(1) + \cfrac{\beta(2)}{\alpha(2) + \ddots}}\right) \tag{1.1}$$

where c is a fundamental constant, $\{f_i, i \in \mathbf{N}\}$ is a family of functions and $\alpha, \beta, \gamma, \delta \in \mathbf{Z}[X]$ are integer polynomials.

The review of the literature, however, did not reveal any model which can process an advanced mathematical statement which does not necessarily contain equations and judge whether it is true or false. This kind of networks would have been particularly suitable for the project as this is the exact task subjects were given during the fMRI experiments that are to be analysed.

1.2.3 Use of AI in neuroscience

A lot of work has been done to understand the brain's processing of language using computational models [Hal+22]. In the past few years, the development of Natural Language Processing (NLP) has led to numerous advances. Two kinds of approaches are reviewed in this section: offshoots of Latent Semantic Analysis (LSA) [Dee+90], like GloVe [PSM14], and Artificial Neural Networks (ANNs) based on the architecture of the Transformer.

Approaches using LSA. Huth et al. [Hut+16] generated an atlas of the semantic regions of the brain. To do so, they used fMRI imaging to scan subjects while they were read a two-hour long story and used LSA to project each word of the story into a 985-dimensional embedding space. The embeddings of the words were derived from their cooccurrences with 985 common English words selected from Wikipedia. They used Ridge regression to try to predict the time serie of each

voxel using the story's words' embeddings, yielding a model weight vector $(\beta_1 \quad \cdots \quad \beta_{985})$ encoding the voxel's semantic tune. They preformed a principal component analysis on the weights of the 10,000 best predicted voxels and found that the first four principal components captured shared semantic features accross subjects. They then projected both vocabulary words' embeddings and voxels' semantic vectors onto the first four principal components and were thus able to associate each voxel with the set of words it is most sensitive to. To to so, they clustered words into twelve categories (using the k-means algorithm) which they labelled by hand, and used a RBG colour code to encode a vector's position in the PC1-PC2-PC3 space, thus bringing out proximity of each voxel with the twelve category labels. The final atlas[1] was obtained by using a probabilistic and generative model of areas tiling the cortex. The atlases for four different subjects are presented on figure 1.2.

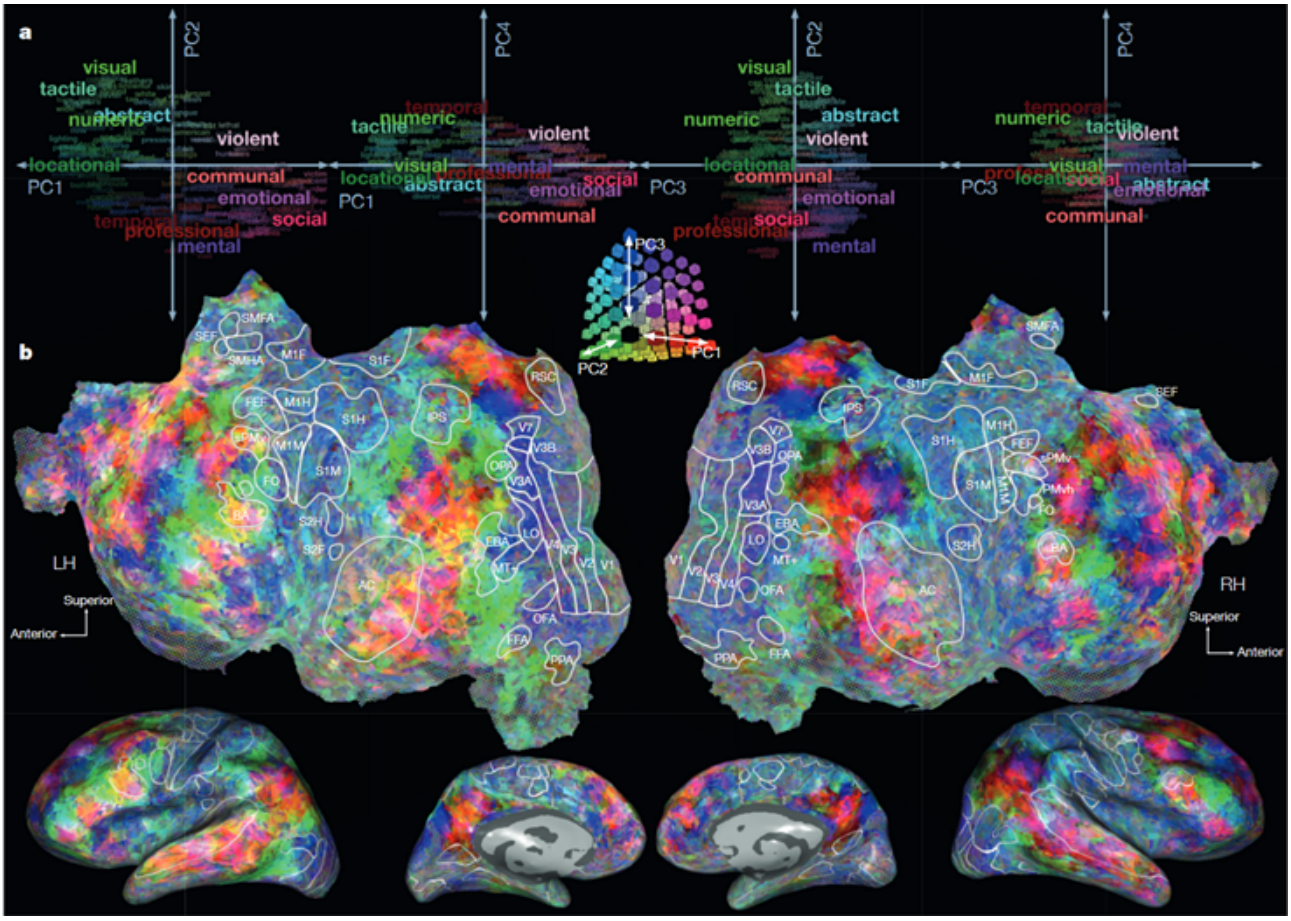


Fig. 1.2. Principal components of voxel-wise semantic models. **(b)** and **(c)** show semantic principal component flatmaps for four different subjects. Extract from [Hut+16].

Another use of LSA was made by Pereira et al. [Per+18], who built a decoder of linguistic meaning from brain activation. To do so, they sampled the semantic space of words using a 30,000-word vocabulary and obtained 300-dimensional GloVe vectors for these words. To visualise the semantic space, they built a two-dimensional clustered map: they clustered the vectors in two-hundred regions and used tSNE and Voronoi parcellation to visualise the clusters in two-dimensions. The semantic map thus obtained is presented on figure 1.3. Pereira et al. then trained a decoder on a set of stimuli that properly covered the semantic space they obtained and on the associated fMRI brain images. The stimuli consisted of one sentence describing one concept (e.g. "Arson is the criminal act of burning a building or wildland."). As the decoder performed well at decoding sentences which were not included in the training dataset, they also evaluated it on text passages consisting

[1] https://www.gallantlab.org/brainviewer/Deniz2019/

of three or four sentences about a particular concept (with no additional training) and observed above-chance performance, even at distinguishing between sentences within a same passage.

Fig. 1.3. Visualisation of the semantic space. Extract from [Per+18].

Approaches using the Transformer. Schrimpf et al. [Sch+21] studied how well NLP tools predict the brain's fMRI time series for language processing. They compared different models and found out that GloVe embeddings do not allow good prediction, whereas intermediate layers of the Transformer predict up to 100% of the signal up to noise ceil in language ROIs. They also found a positive correlation between a model's prediction score and its ability to predict the next word of a sentence, and that the more a model predicts activity, the better it predicts behavioural parameters like reading time. Interestingly, Schrimpf et al. also found that the Transformer performs above-chance even with no training: while the model's architecture (e.g. number of layer) is import, the size of its training dataset or vocabulary does not have much influence on the performance of the model. Similarily, Caucheteux et al. [CK22] compared several deep language models with the brain using MEG and fMRI imaging. They too found that the middle layers of compositional models like the Transformer predict brain activity with high accuracy and that a model's quality of prediction and ability to predict next word are positively correlated. They thus argue for a partial – there seems to be a plateau for high performance models – convergence of brains and algorithms in natural language processing: the higher a model's performance, the better its prediction of brain fMRI or MEG activity, although architecture plays an important role too.

On another note, Caucheteux et al. [CGK21] tried to replicate the hierarchy of language brought out by Lerner et al. [Ler+11] by using GPT-2 to emulate the brain response. Their results show strong similarities with those of Lerner et al., suggesting that the Transformer – and especially its intermediate layers – is a good model of the way language is processed in the brain. Millet et al. [Mil+22] obtained similar results with other models of the Transformer. In addition, they found that self-supervised learning models perform sightly better than supervised learning ones.

To compare the Transformer's and brain's activation, all the above works use the same method: they select a subset of the stimuli and fMRI data (e.g. 80%) to fit a GLM on of the brain's activity given the model's hidden states to obtain a weight matrix W and compute a *brain score* using the rest of the data

$$\text{corr}(Y, W \cdot X) \tag{1.2}$$

where Y is the true brain response and X the hidden states of the model for the held-out stimuli. The final brain score is obtained by averaging the brain scores obtained for each subset of a partition the dataset.

Interpretability of ANN models. A major drawback when using ANNs is that these models are often considered black boxes whose parameters are not easily interpretable. Kar et al. [KKF22] argue that the best neuroscientific interpretability is achieved when both the ANN aligns with the brain and the models' parameters are interpretable. Thus, a new prolific strand of literature should emerge from more interpretable NLP tools.

1.3 Conventions

Softwares. All analyses were coded in `Python` 3.9. Figures were plotted using `pandas` dataframes and either `matplotlib`'s `pyplot` or `seaborn`. Statistical tests and clusterings were performed using `scikit-learn` and `statsmodels`. NLP was done using `transformer` and `spacy`. fMRI data were analysed using `nilearn`.

Statistical tests. For all non fMRI-related statistical tests, a p-value was computed and the null hypothesis was rejected whenever $p < 0.05$.

fMRI data analyses. The hemodynamic response function (HRF) used is the `spm` HRF from `nilearn`. All brain activation results are reported unsing an uncorrected FPR $p < 0.001$ and cluster size thresholded to ten voxels.

Data availability. All material associated with the work presented in this thesis can be found online at `https://perso.crans.org/sdebray/projects/MathsNLP/MathsNLP.tar.gz`.

CHAPTER 2

Semantic Analysis of Mathematics

This chapter describes the creation of a mathematical vocabulary, the GloVe embeddings [PSM14] of which were used to analyse fMRI-acquired data. So as to make sure that the GloVe embedding vectors capture the mathematical semantics of words, several analyses were performed, namely principal component analysis (PCA), spectral clustering (see appendix A on page 43) and hierarchical clustering. These analyses are presented in section 2.2 on page 11.

2.1 Creation of a mathematical vocabulary

This section presents the method followed to create a mathematical vocabulary. The complete thousand-word vocabulary can be found in the tarball of the project, along with the code used to obtain it.

2.1.1 Mathematical corpus

A major difficulty encountered in this project was to find a suitable mathematical corpus in French. The need for French was due to the fact that the fMRI data were acquired on French subjects presented with stimuli in French. Most mathematical articles are written in English and access to French textbooks is mostly restricted. The solution that was found was to use mathematical articles from French Wikipedia[1].

Since Wikipedia does not allow for its pages to be scrapped by a bot, the only way to get all pages was to use a dump. The dump that was chosen is `wikipedia_fr_all_maxi_2022-04.zim`[2] (created in April 2022).

All pages of the dump were parsed and a bot decided whether each page was mathematical or not. To do so, it reached the bottom of the page and searched for an occurence of one of the following strings:

- `Portail des mathématiques`;
- `Portail de la géométrie`;
- `Portail de l'analyse`;
- `Portail de l'algèbre`;
- `Portail des probabilités et de la statistique`;

[1] `https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Accueil_principal`
[2] `https://download.kiwix.org/zim/wikipedia_fr_all_maxi.zim`

- `Arithmétique et théorie des nombres`;
- `Portail de la logique`;
- `Portail de l'informatique théorique`.

indicating that the page belonged to the portal of mathematics or theoretical computer science. These strings were manually selected to cover all mathematical pages. The dump also contained disambiguation pages that were discarded; they were detected as they are the only pages containing no portal banner at the bottom. In an early version, physics pages were also considered mathematical but this choice later proved questionable and physics pages were removed. Table 2.1 summarises some figures about the parsing procedure.

Size of the dump	30.3 GiB
Total number of pages	2,402,095
Number of mathematical pages	16,455
Number of non-mathematical pages	2,236,840
Number of disambiguation pages	148,800
Size of the mathematical corpus	66.8 MiB
Numer of words in the mathematical corpus	11,214,962

Table 2.1. Summary of the scrapping of Wikipedia.

2.1.2 Extraction of the mathematical vocabulary

The corpus obtained in section 2.1.1 on the previous page was already somewhat preprocessed. However, so as to ensure the vocabulary does not contain several similar occurences of the same word (e.g. singular and plural for a noun of infinitive and conjugate form for a verb), a lemmatisation step was deemed necessary. Two lemmatisation passes were carried out using `Python` `spacy`'s `fr_core_news_md` model. The lemmatised corpus was then provided as input to the GloVe pipeline[3]. The pipeline has three steps: it first creates a 50,000-word vocabulary sorted by decreasing number of occurences, then builds a coocurrence matrix on the vocabulary words in the corpus and finally runs GloVe on said matrix. Roughly, GloVe uses the cooccurrence matrix of a vocabulary in a corpus to derive semantic vectors for each words by taking into account ratios of cooccurrences, a detailed description of this algorithm is given in appendix A.1 on page 43. The window size for cooccurrence was set to fifteen words and the embedding vectors lied in fifty dimensions.

This procedure was also applied to the non-mathematical corpus obtained in 2.1.1 on the previous page, so as to provide an embedding for non-mathematical stimuli of experiments to be analysed in chapter 4 on page 29.

As it happens, the vocabulary output by GloVe only contained 39,345 words, because the GloVe pipeline did not count words with less than five occurrences in the corpus. It was pretty noisy (function words, pieces of LaTeX code for equations, `HTML` tokens, etc.) and much too large (most words had less than ten occurrences in the corpus) so it needed to be reduced in size. To do so, words of the vocabulary were reviewed manually in decreasing order of frequency by a person with an extensive mathematical training. Only the 1,000 first words which were deemed mathematical – no matter whether they are elementary or advanced concepts, or the area of mathematics they are related to – were kept, thereby constituting the final thousand-word mathematical vocabulary.

As show on figure 2.1 on the next page, the words of the final mathematical vocabulary often appear in the non-mathematical corpus as well but are overall more frequent in the mathematical corpus (log frequency [per million] in the mathematical corpus: $\mu = 3.94$, $\sigma = 1.36$; log frequency [per million, when non-zero] in the non-mathematical: $\mu = 2.76$, $\sigma = 1.72$)[4].

[3] `https://github.com/stanfordnlp/GloVe`

[4] A first criterion chosen to obtain the mathematical vocabulary was to keep those words of the 39,345-word

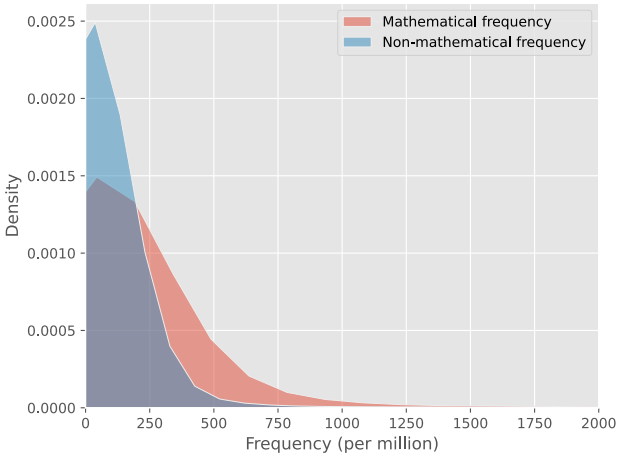


Fig. 2.1. Distribution of the frequency of words of the final thousand-word mathematical vocabulary in the mathematical and non-mathematical corpora.

2.2 Analysis of the mathematical vocabulary

Once the GloVe embeddings of the mathematical vocabulary was obtained, it seemed useful to visualise it so as to identify its main features. This section presents three different views of the same GloVe object, each highlighting different features. Ultimately, these embeddings are to be used for fMRI analysis in chapter 4 on page 29, and the only analysis which will be retained there is the PCA performed in section 2.2.1, as it allows to reduce the dimensionality of the embeddings while accounting for 19% of the observed variance.

2.2.1 Principal component analysis

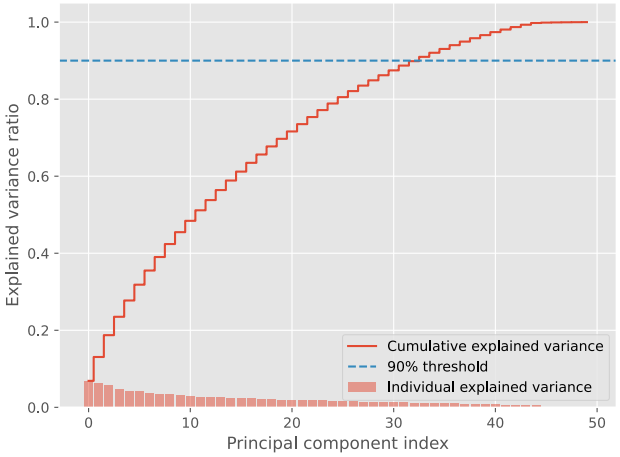
First, a Principal Component Analysis (PCA) [WEG87] was performed to try to see how much variance is explained by the first principal components and if they encode interesting features.

All vectors were submitted to PCA and the output vectors were normalised. The new vectors were then projected onto the $k = 34$ first principal components, where k was chosen such that the k first principal components explain at least 90% of the observed variance. The thirty-four-dimensional vectors were finally split into ten clusters using spectral clustering [Lux07] to better visualise the distribution of semantically consistent clusters in the planes PC1-PC2, PC1-PC3 and PC2-PC3. The clusters were labelled by hand, but it was sometimes difficult to find a salient cue common to all words within a cluster.

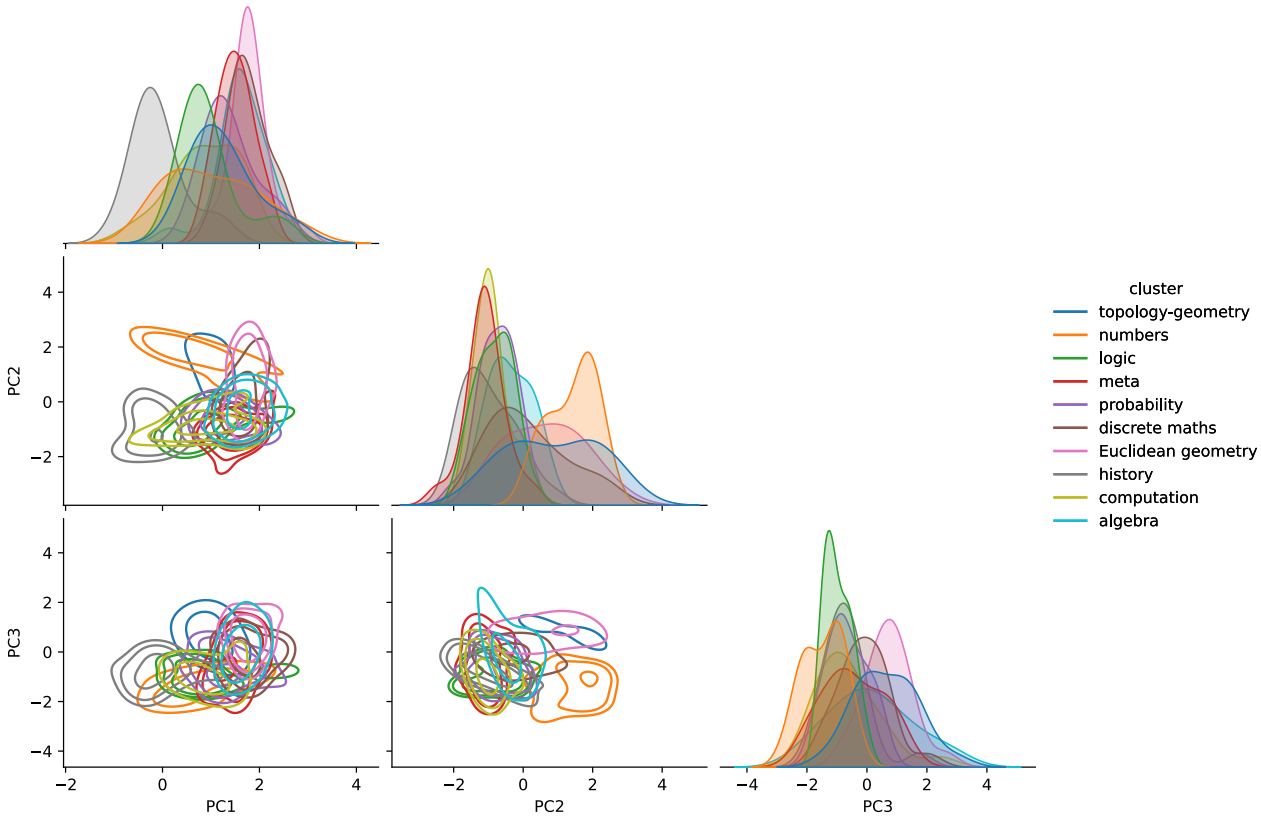
Figure 2.2 on the following page presents the results of the PCA. Figure 2.2a on the next page shows that the three first principal components explain about 19% of the observed variance. On figure 2.2b on the following page, one can see that the projection onto the three first principal components preserves the clustering, although the planes PC1-PC2, PC1-PC3 and PC2-PC3 do not allow to make a distinction between clusters. Two clusters, however, seem to stand out: history (that is names of famous mathematicians) on PC1 and numbers on PC2.

An attempt was made to interpret each of the three first principal components' feature, but no satisfactory interpretation emerged. Representation of words in the planes PC1-PC2, PC1-PC3 and PC2-PC3 can be found on figure 2.3 on page 15.

vocabulary which were ten times more frequent in the mathematical coprus than in the non-mathematical one. This criterion was nowhere near satisfying as only 60.2% of the final vocabulary respect it. If this criterion had been used, words like "un" or "ensemble" would not appear in the vocabulary.

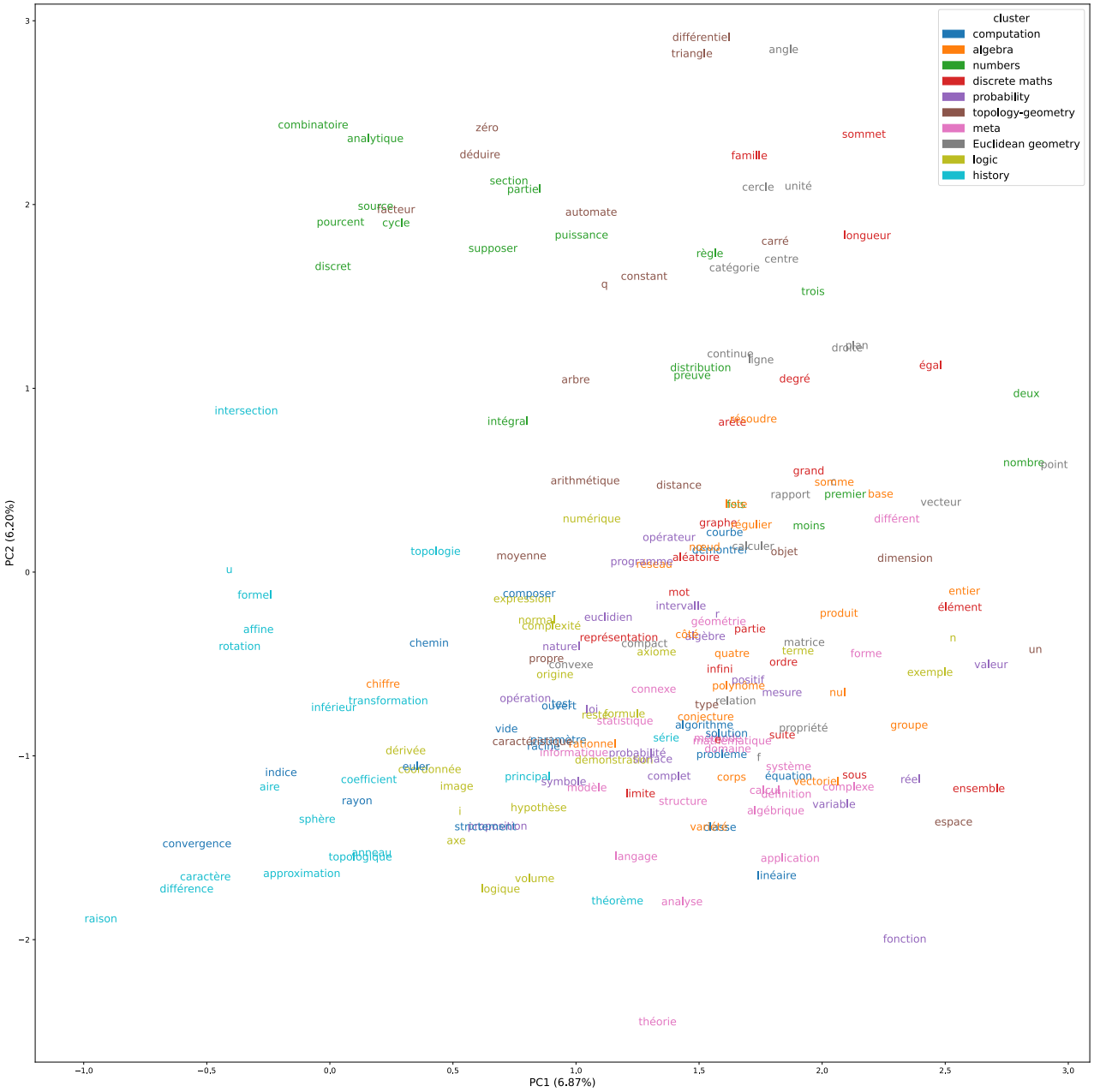


(a) Ratio of variance explained by each principal component.

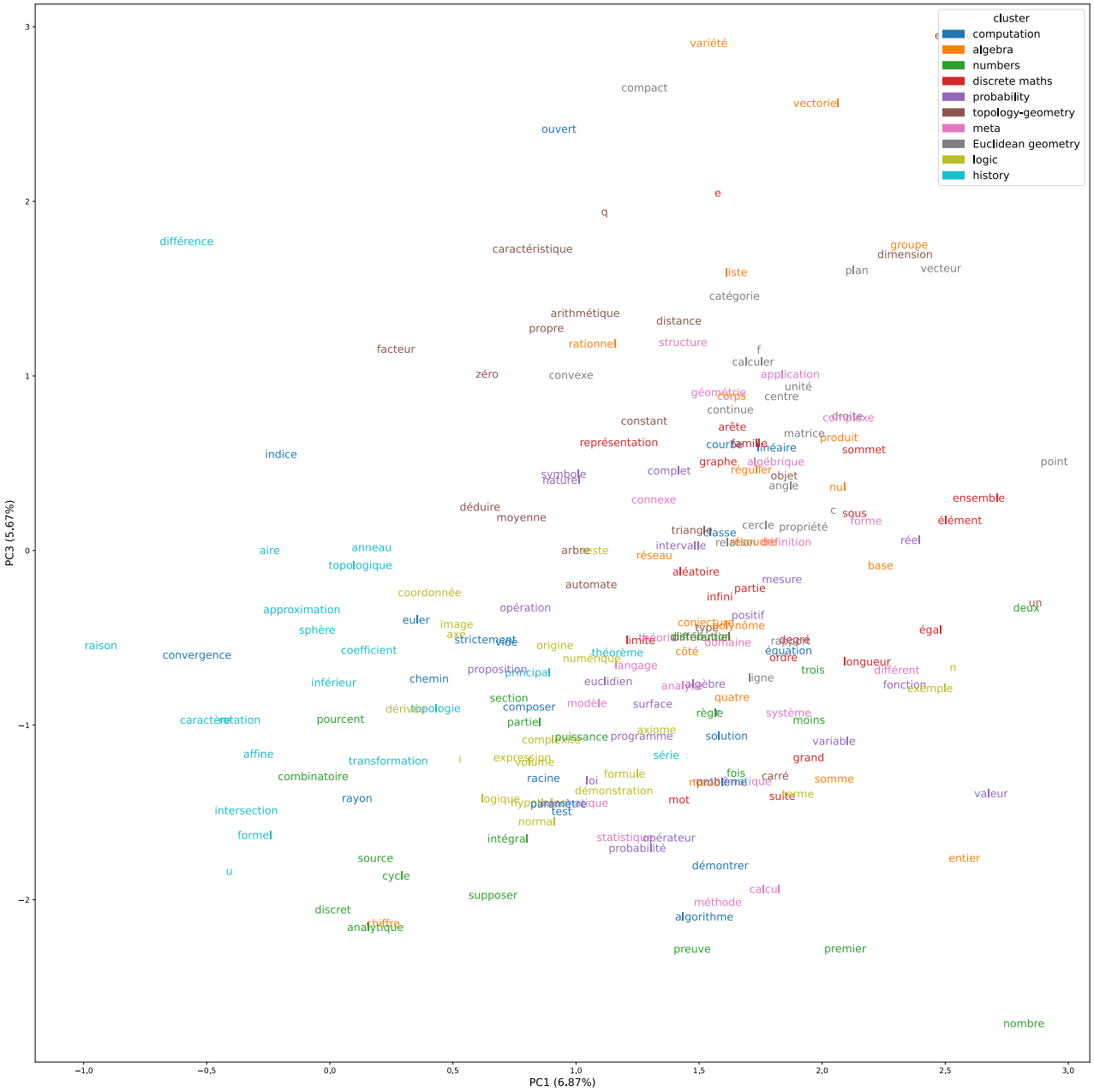


(b) Kernel density estimate plot of the twenty most frequent words of each cluster in the planes PC1-PC2, PC1-PC3 and PC2-PC3. Levels encompass 30%, 50% and 80% of the probability mass.

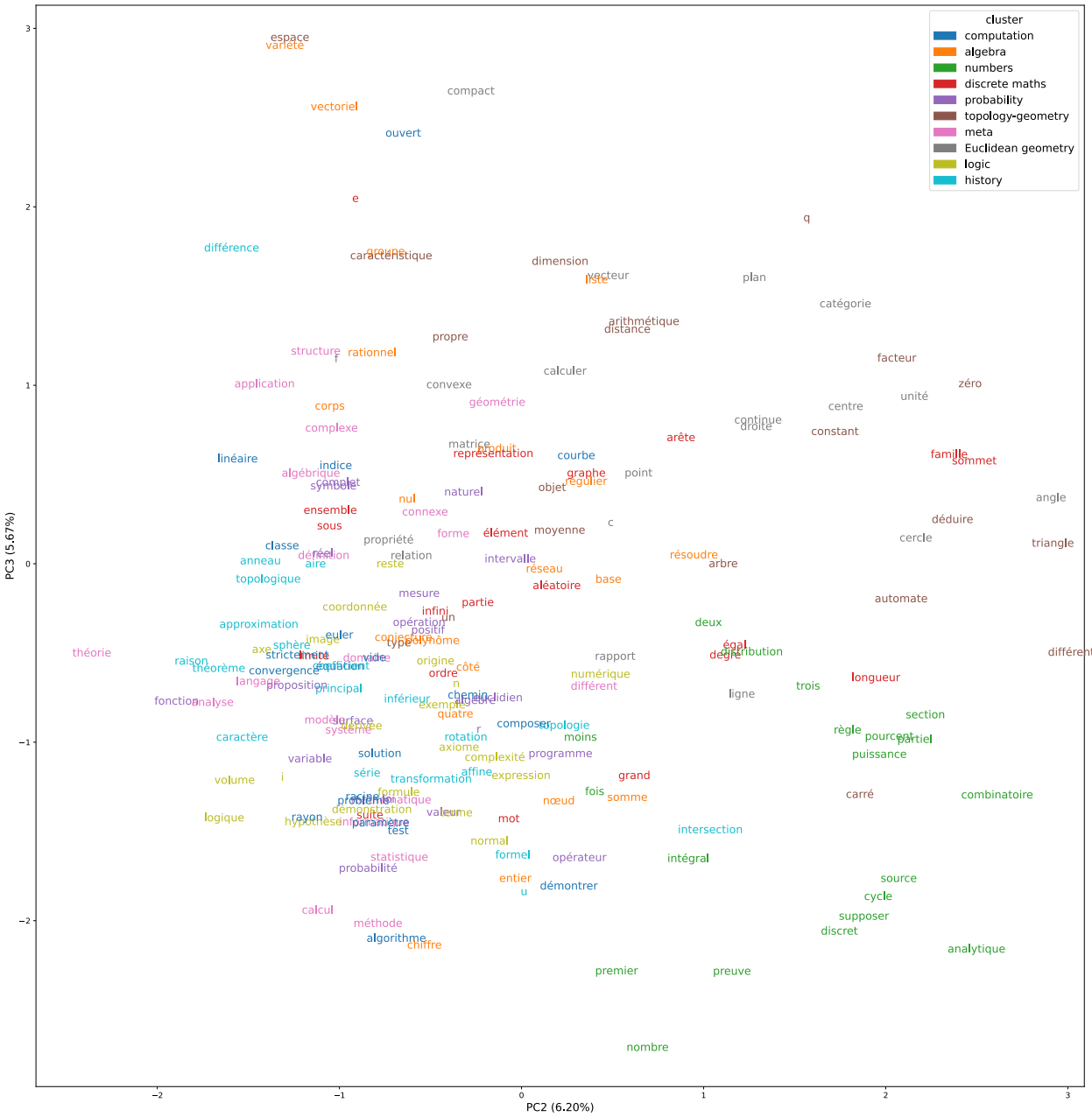
Fig. 2.2. Principal component analysis of the mathematical vocabulary.


cluster
computation
algebra
numbers
discrete maths
probability
topology-geometry
meta
Euclidean geometry
logic
history
PC1 (6.87%)
PC2 (6.20%)
différentiel
triangle
angle
combinatoire
analytique
zéro
déduire
sommet
famille
section
partiel
cercle
unité
source
facteur
pourcent
cycle
automate
puissance
supposer
discret
règle
carré
longueur
centre
catégorie
constant
q
trois
droite
plan
continue
ligne
distribution
preuve
égal
arbre
degré
deux
intersection
intégral
arête
résoudre
nombre
point
arithmétique
distance
grand
somme
rapport
premier
base
vecteur
numérique
graphe
régulier
moins
différent
opérateur
courbe
topologie
moyenne
calculer
objet
dimension
aléatoire
u
formel
composer
expression
mot
intervalle
entier
élément
produit
normal
euclidien
géométrie
affine
complexité
partie
rotation
chemin
représentation
algèbre
matrice
n
un
naturel
compact
axiome
quatre
terme
forme
propre
convexe
ordre
valeur
origine
infini
exemple
chiffre
positif
connexe
polynôme
mesure
nul
transformation
inférieur
opération
ouvert
loi
relation
type
conjecture
algorithme
propriété
groupe
vide
statistique
solution
suite
série
dérivée
euler
racine
informatique
probabilité
problème
indice
coefficient
coordonnée
principal
symbole
complet
corps
équation
sous
réel
aire
image
modèle
limite
vectoriel
ensemble
rayon
hypothèse
structure
algébrique
variable
sphère
espace
convergence
axe
langage
application
anneau
topologique
caractère
approximation
linéaire
différence
volume
logique
théorème
analyse
raison
fonction
théorie

(a) PC1-PC2


cluster
computation
algebra
numbers
discrete maths
probability
topology-geometry
meta
Euclidean geometry
logic
history
PC1 (6.87%)
PC3 (5.67%)
variété
compact
vectoriel
ouvert
e
q
différence
caractéristique
groupe
dimension
vecteur
plan
liste
catégorie
arithmétique
distance
propre
rationnel
structure
facteur
f
calculer
zéro
convexe
application
unité
géométrie
corps
centre
continue
constant
droite
arête
matrice
représentation
produit
courbe
famille
linéaire
sommet
indice
graphe
algébrique
point
symbole
naturel
complet
régulier
objet
angle
nul
ensemble
déduire
connexe
moyenne
sous
forme
élément
triangle
classe
cercle
propriété
réel
aire
anneau
intervalle
arbre
réseau
topologique
aléatoire
base
automate
mesure
coordonnée
partie
infini
un
deux
approximation
opération
euler
positif
sphère
image
égal
raison
convergence
coefficient
limité
côté
équation
ordre
longueur
trois
différent
n
inférieur
chemin
proposition
principal
euclidien
ligne
fonction
exemple
section
composer
modèle
surface
quatre
caractérisation
pourcent
partiel
règle
système
moins
puissance
programme
axiome
solution
variable
affine
transformation
expression
série
grand
combinatoire
racine
loi
formule
fois
carré
somme
rayon
démonstration
problème
suite
valeur
intersection
logique
paramètre
mot
test
normal
formel
intégral
probabilité
entier
source
démontrer
u
cycle
calcul
supposer
méthode
discret
algorithme
analytique
preuve
premier
nombre

(b) PC1-PC3



(c) PC2-PC3

Fig. 2.3. Representation of the twenty most frequent words of each cluster in the planes PC1-PC2, PC1-PC3, PC2-PC3.

2.2.2 Spectral clustering

Following Pereira et al. [Per+18], the GloVe vectors were also clusterised using spectral clustering [Lux07] (see appendix A.2 on page 44 for a description of the algorithm). Roughly, spectral clustering does the same as the k-means algorithm, but it first normalises the data to clusterise. The aim of the clustering is to see how GloVe embeddings distinguishe between concepts and what connections they make between the different branches of mathematics. The number of clusters was manually set to 100, as it was the largest number which allowed easy interpretation of clusters.

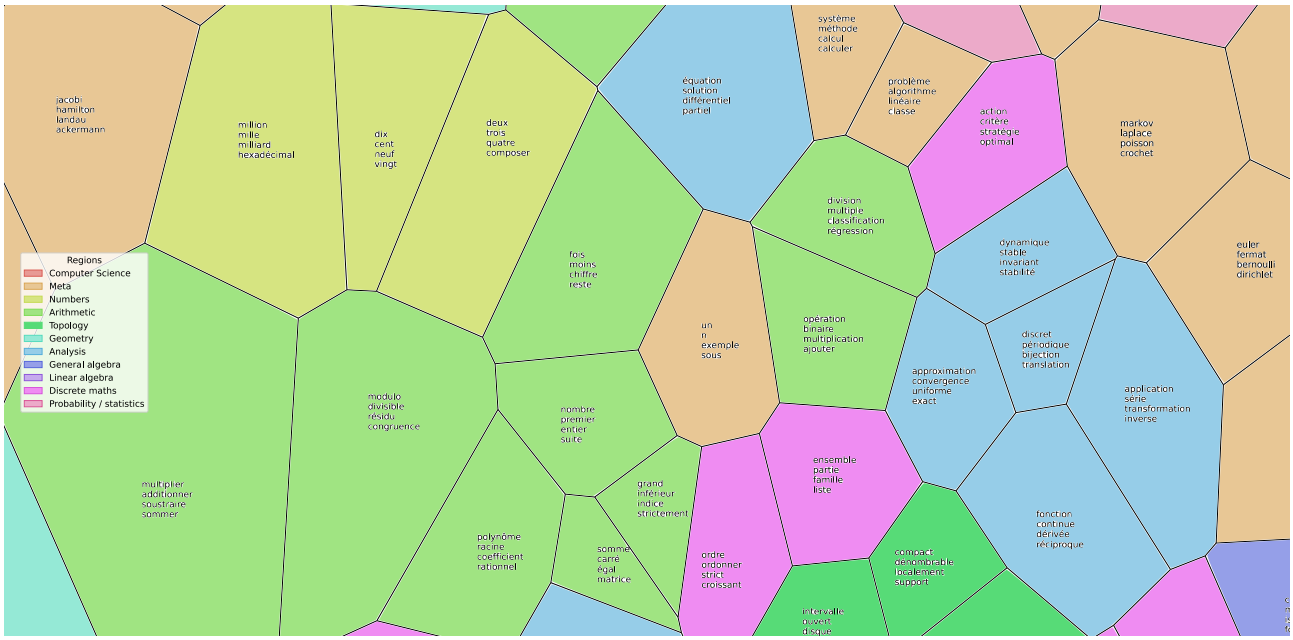


Fig. 2.4. View of a region of the clustered two-dimensional semantic map of mathematics.

To obtain a two-dimensional representation, the clusters' centroids along with the four most frequent words of each cluster were projected in two dimensions using tSNE (with random seed set to 42) and Voronoi boundaries were plotted around centroids. Clusters were then coloured by hand to show proximity between different branches of mathematics. A region of the final map is presented on figure 2.4 and the full map is available online at `https://perso.crans.org/sdebray/projects/MathsNLP/ClusteredMapMathematics.svg`.

Most clusters can be labelled easily and make good sense. However, the spacial distribution of clusters does not allow to indentify larger clusters dedicated to geometry or algebra for instance. Interestingly, names of mathematicians are not grouped with their field of interest but are all clustered together, thus creating "history of mathematics" clusters. Moreover, some basic parts of mathematics, like arithmetics and numbers, are split into several clusters of increasing level of elaboration (e.g. "very small numerals" < "larger numerals" < "powers of ten" or "basic operations" < "modular arithmetic").

2.2.3 Hierarchical clustering

A second kind of clustering was applied to the vectors: agglomerative clustering [Nie16]. This clustering allows to have a hint of the hierarchy of mathematical concepts rather than their geometrical proximity.

To obtain a visualisation of the agglomerative clustering, a dendrogram was plotted using Python's `scipy.cluster.hierarchy.dendrogram` function. The number of leaves is limited to 100 so as to get a clustering similar to that obtained in section 2.2.2 and each non-singleton

leaf l is labelled with the four most frequent words labelling the singleton leaves merged to obtain l. A subtree of the dendrogram is presented on figure 2.5 on the next page and the full dendrogram is available online at `https://perso.crans.org/sdebray/projects/MathsNLP/DendrogramMathematics.svg`.

One can see on figure 2.5 on the following page that clusters are overall easily interpretable and labellable. Clusters are quite specific: for instance, "dérivable, pôle, différentiable, holomorphe" stands for complex analysis. However, even though the difference between the two children of a node is generally easy to figure out, some intuitively close branches (like algorithms and theoretical computer science) are represented far from each other. This problem might also be due to the choice of the order in which leaves are plotted, but this choice is made by the plotting function based on the distance to the parent node, which seems sensible. Here also, names of mathematicians are grouped toghether rather than with the domain they worked on.

Interestingly, this representation seems nicer than that obtained using spectral clustering and tSNE following the method of Pereira et al. [Per+18]. The clusters are more specific and easily interpretable and and the proximity of clusters makes more sense. Agglomerative clustering thus seems suitable for visualising similarities in a vocabulary.

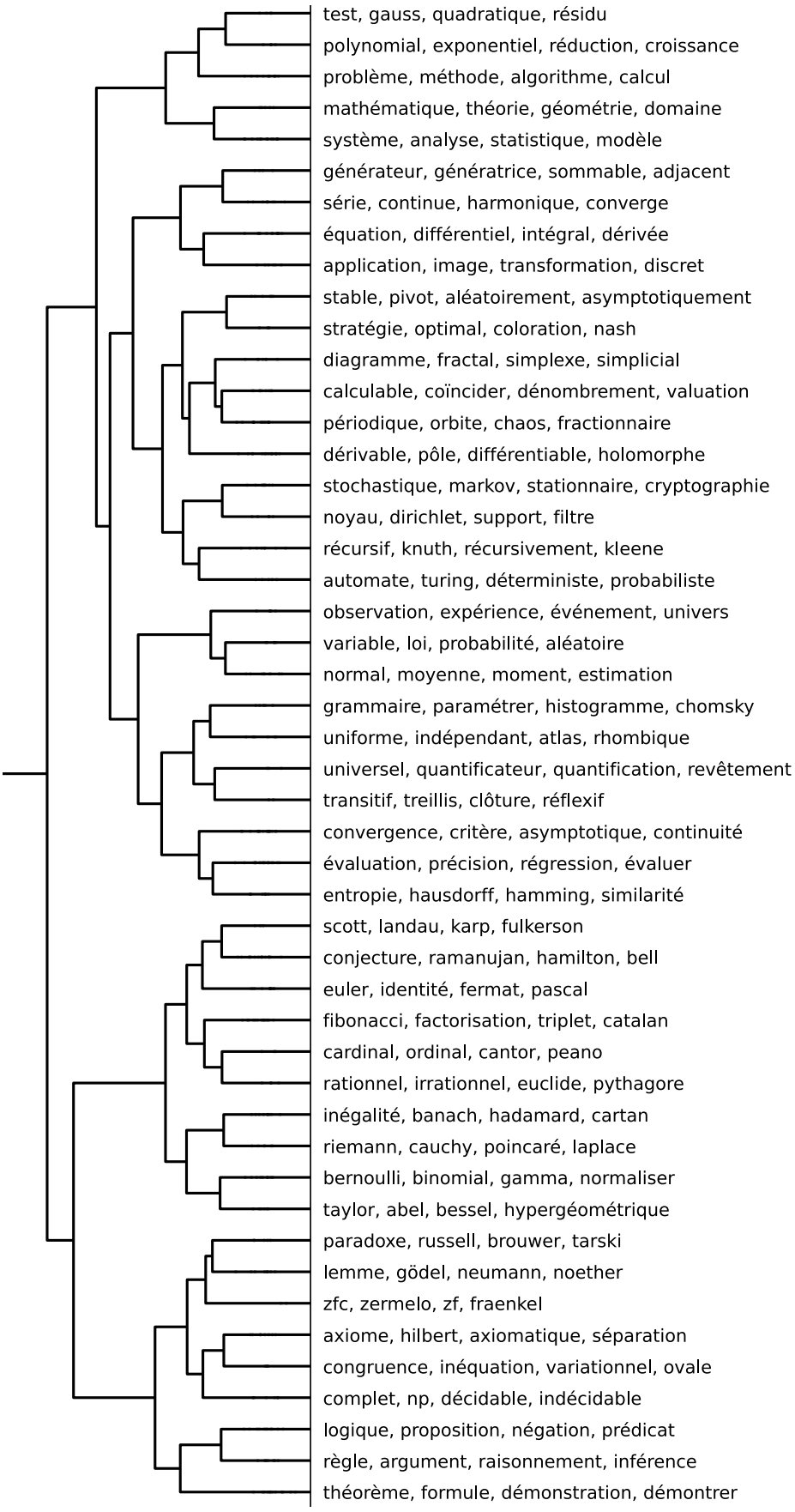


Fig. 2.5. View of a subtree of the dendrogram representing the agglomerative clustering of the GloVe embedding of the mathematical vocabulary.

CHAPTER 3

Mathematics and the Transformer

In chapter 2, GloVe was used to generate lexical semantic embeddings. Once computed, these embeddings were fixed, that is, context-independant. By contrast, the Transformer [Vas+17] computes, word by word, context senstive representations. This chapter is dedicated to the attempts that were made to model mathematics using the Transformer. The main conclusion of this chapter is that the default Transformer is not a good model to probe mathematics processing in the human brain, nor does it perform well at extracting mathematical knowledge.

The aim of this chapter is to compare humans' and the Transformer's performance on the same tasks. Since Amalric et al. [AD16; AD18; AD19] and Moreno and Dehaene have already designed and conducted behavioural (and fMRI) experiments on French speaker human subjects, it was decided to evaluate the Transformer on these tasks too.

3.1 Transformer's model selection

This section describes the method that was used to chose between GPT-fr [SC21] and CamemBERT [Mar+20].

In the first place, these two models were selected among all those presented in HuggingFace[1]. When making this choice, the main criterion was that the model used should be trained on a very large corpus in French, since the fMRI data from chapter 4 on page 29 were acquired on French speakers who were presented stimuli in French. However, HuggingFace references quite a few such models. To further restrict the choice, a review of the literature and of the different tasks typically assigned to the Transformer brought out two tasks that were relevant to the purpose of the study: text generation and mask filling; and, for each of these task, one family of models: GPT and BERT. GPT-fr and CamemBERT are a trade-off between training dataset size and recentness for the two identified families.

3.1.1 Tokenising

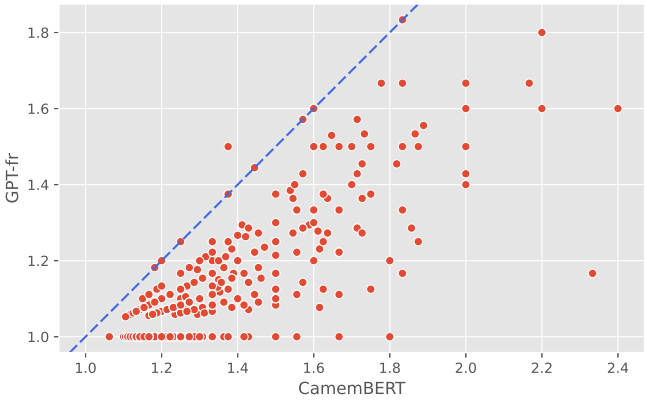
Since the aim of this project was to compare the Transformer with human subjects, the very first step was to make sure that the model tokenises words in a sensible way. The working hypothesis is that *human subjects perform as few sub-word splits as possible*, thus the ideal tokeniser would be

[1] `https://huggingface.co/`

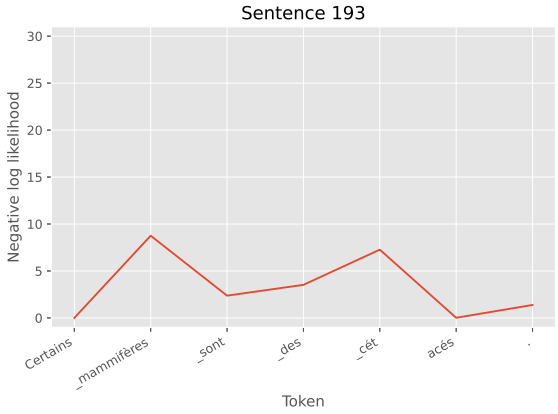
such that

$$\frac{\text{number of tokens}}{\text{number of words}} = 1. \tag{3.1}$$

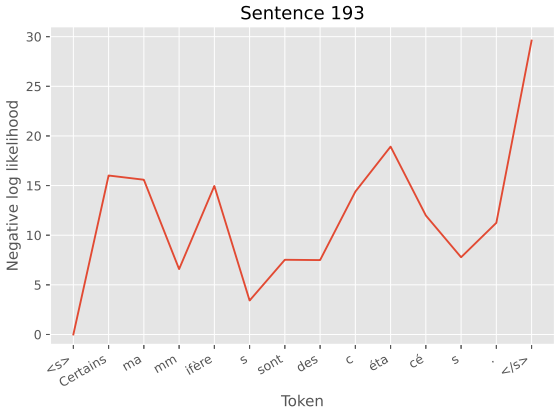
This requirement is all the more important since the Transformer computes token-wise negative log likelihood before averaging over tokens to obtain a text-wide negative log likelihood. The hypothesis of the project, on the other hand, is that humans only assess entire words, even though they might split words into more basic sub-words from time to time for understanding (as opposed to critical analysis) purposes, e.g. "uncomputable" = "un" + "compute" + "able".



(a) GPT-fr's number of tokens to number of words ratio against CamemBERT's (one point represents one stimulus).



(b) GPT-fr's tokenisation of the sentence "Certains mammifères sont des cétacés.".



(c) CamemBERT's tokenisation of the sentence "Certains mammifères sont des cétacés.".

Fig. 3.1. Comparison of GPT-fr's and CamemBERT's tokenisers.

Systematic analysis of the two models' tokeniser (see figure 3.1a) showed that GPT-fr's tokeniser has a much lower tokens to words ratio than CamemBERT's (for GPT-fr, $\mu = 1.12$, $\sigma = 0.15$ and for CamemBERT, $\mu = 1.34$, $\sigma = 0.21$). To illustrate this trend, figures 3.1b and 3.1c compares the two tokenisers' output on the same input. The need for equation 3.1 to be verified is made clear by CamemBERT's high negative log likelihood on nonsensical tokens like "mm" or "éta", which are not even French morphemes and thus, given the working hypothesis, are not assessed as is by the human brain. This fact alone justifies the choice of GPT-fr as the model to be compared with human subjects.

However, GPT-fr does not fully verify 3.1. It it thus important to verify that its tokeniser behaves uniformly across all categories of stimuli and all truth values. Figure 3.2 on the next page plots, for each experiment, the ratio of the number of tokens to number of words by sentence depending on

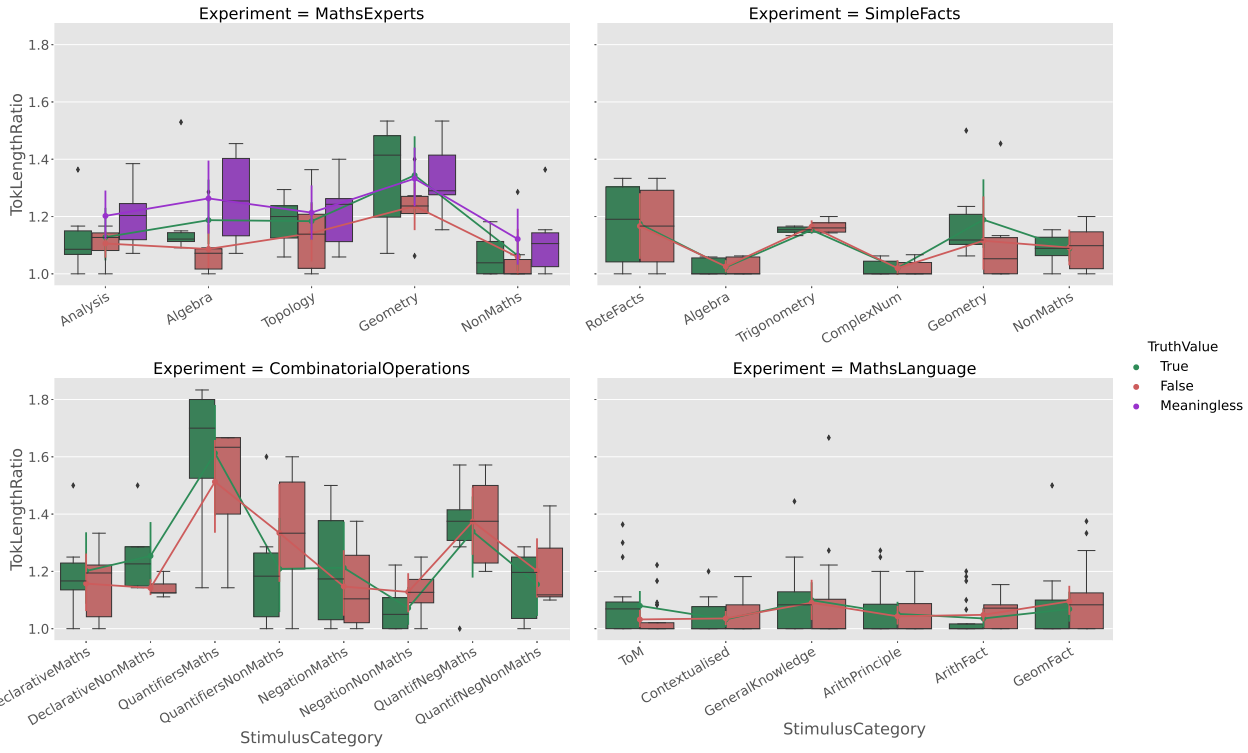


Fig. 3.2. Tokens to words ratio of GPT-fr's tokeniser as a function of the stimuli's category and truth value.

the category and truth value. For each experiment, the number of tokens to number of words ratio by stimulus was computed. These ratios were then submitted to a two-way ANOVA with the factors category (e.g. algebra, topology) and truth value; except for categories colorless green and word lists as all stimuli in these categories were false. An effect of category was found for every experiment, and an effect of truth value was found for MATHSEXPERTS (see table 3.1, below-threshold p-values are printed in bold).

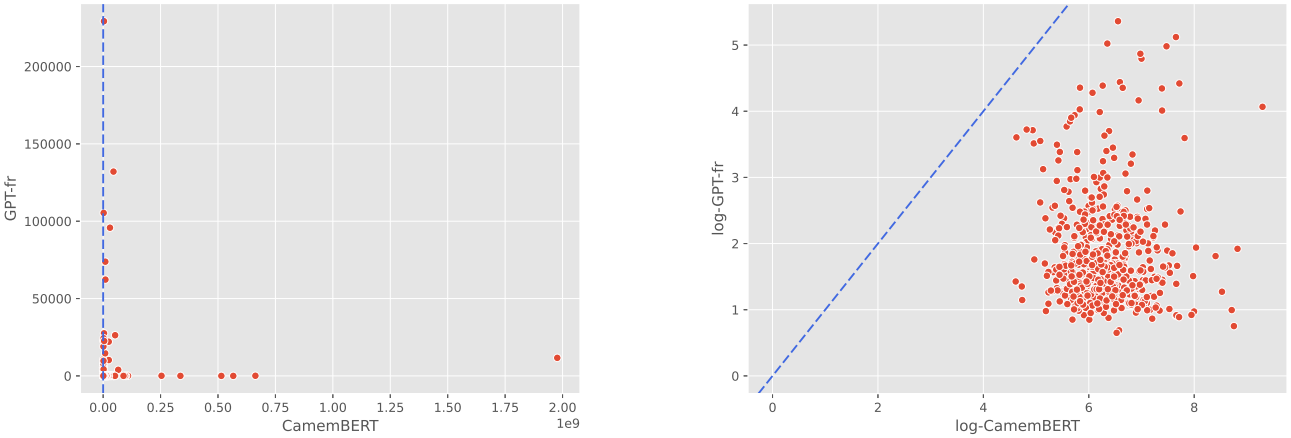
Experiment	Effect of category	Effect of truth value	Interaction
MATHSEXPERTS	$\mathbf{6.6 \times 10^{-5}}$	**0.013**	0.95
SIMPLEFACTS	$\mathbf{2.3 \times 10^{-4}}$	0.65	0.91
COMBINATORIALOPERATIONS	$\mathbf{7.5 \times 10^{-8}}$	0.84	0.65
MATHSLANGUAGE	**0.026**	0.76	0.60

Table 3.1. ANOVA of the number tokens to number of words per stimulus with stimulus category and truth value as independent variables for GPT-fr's tokeniser (in p-values).

3.1.2 Range of perplexity

Although CamemBERT's number of tokens to number of words ratio is prohibitively high, it seems relevant to compare GPT-fr's and CamemBERT's range of complexity to make the best choice. For this comparison, both models were used as Causal Language Modelling models, even though CamemBERT is best suited for Masked Language Modelling (see appendix A.3 on page 45 for an explanation of what CLM and MLM are). Figure 3.3 on the following page plots GPT-fr's preplexity against CamemBERT's for each stimulus.

Here again, GPT-fr seems to perform much better than CamemBERT and systematic comparison proved that CamemBERT's perplexity is always at least ten times larger than GPT-fr's,



(a) Linear-linear plot.

(b) Log-log plot.

Fig. 3.3. GPT-fr's perplexity against CamemBERT's (one point represents one stimulus). Both models used as CLM models.

which motivates the log-log plot in figure 3.3b. This may have to do with the fact that GPT-fr is a text-generation model whereas CamemBERT is a best suited for mask-filling, which is less relevant a task when it comes to causal language modeling. In any case, figures 3.1a on page 20 and 3.3 justify that *the model which should be used in the project is GPT-fr*.

3.2 Measure of surprisal

As explained in appendix A.3 on page 45, the Transformer returns different kinds of output. This section is a discussion as to which output is most relevant for comparing it with human subjects.

3.2.1 Mean vs max negative log likelihood

Given a tokenised input $X = (x_1, \ldots, x_t)$, GPT-fr's perplexity is defined as the exponential of the mean negative log likelihood of tokens $x_1, \ldots, x_t$ (see equation A.13 on page 46). That is to say that, given tokens $x_1, \ldots, x_{i-1}$, the model assigns a score to each token of its vocabulary – which captures the model's surprisal to see vocabulary token v occur after $x_1 \cdots x_{i-1}$ – and then averages its scores over all actual tokens tokens $(x_1, \ldots, x_t)$.

Humans, on the other hand, seem to perform the same token-wise (i.e. word-wise according to the working hypothesis) analysis but, intuitively, it seems unlikely that they average their surprisal over all words. For example, two very long and trivial sentences ending respectively with "[...] and I ate a steak with a crocodile." and "[...] and I ate a steak with my best friend." will roughly have the same mean surprisal, although the former sounds much more peculiar than the latter.

Therefore, considering the mean negative log likelihood seems to be inadequate with a view to comparing GPT-fr with human subjects. Instead, it seems more relevant that the surprisal on a sentence be given by the *maximum* of the negative log likelihood on the tokens[2] of said sentence. Besides, even in the purpose of using the Transformer as a classifier as in section 3.3.1 on page 24, taking the maximum of negative log likelihood rather than the average seems more appropriate so as to make sure that a high surprisal on a given token is not smoothened by a low surprisal on the rest of the tokens. Indeed, most false or nonsensical mathematical statements, could be turned into a true or sensical statement only by changing one word, e.g. in the following example:

(1) Over a **C**-vector space E of finite dimension, all matrices are continuous.

[2] Ideally, the maximum would be computed over the words rather than the tokens.

(2) Over a **C**-vector space E of finite dimension, all matrices are invertible.
(3) Over a **C**-vector space E of finite dimension, all matrices are trigonalisable.

sentences (1) on the preceding page, (2) and (3) are respectively meaningless, false and true yet they only differ by one word.

3.2.2 Negative log likelihood vs perplexity

Although negative log likelihood and perplexity are closely related – since the perplexity on an input X is the exponential of the (mean or maximum) negative log likelihood on X – the choice of scale is curcial both for ANOVAs and correlation tests. This section justifies the choice of *negative log likelihood* as the relevant metric to be considered.

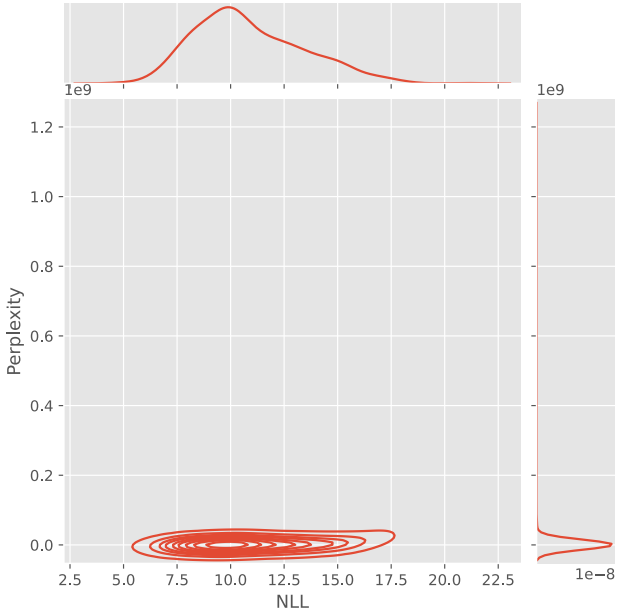


Fig. 3.4. Comparison of the distribution of GPT-fr's maximum negative log likelihood and perplexity on the stimuli.

First, by comparing the distributions of negative log likelihood and perplexity on the stimuli, one can see that negative log likelihood is much closer to a normal distribution that perplexity is. Figure 3.4 plots the bivariate distribution along with the two marginal distributions and also shows that the range of perplexity is much larger than that of negative log likelihood. Thus the distribution of negative log likelihood seems to have nicer properties. It is worth noting that this state of affairs remains true even if mean negative log likelihood were to be considered rather than maximum negative log likelihood (see figure D.1 on page 67).

The second reason to use negative log likelihood rather than perplexity is that is well established that the brain processes values logarithmically rather than linearly [Deh03; SL13]. It would therefore make no sense, with a view to comparing the score with humans' performance, to exponentiate negative log likelihood to obtain a linear scale.

3.3 The Transformer's performance

This section presents the assessment of the Transformer's mathematical abilities. 3.3.1 on the following page deals with the use of the Transformer to discriminate between sensical and nonsensical, true and false, and mathematical and non-mathematical statements. 3.3.2 on page 26 compares

the Transformer's performance with those of the subjects of the fMRI experiments described in B on page 47. Both using tools and conventions described in 3.2 on page 22.

3.3.1 Does the Transformer understand mathematics?

Using the metric defined in section 3.2 on page 22, an analysis of GPT-fr's surprisal was conducted in order to assess the network's ability to distinguish between true, false and meaningless statements.

Figure 3.5 plots the maximum negative log likelihood for each stimulus depending on the experiment, the stimulus category and truth value. The maximum negative log likelihood was then submitted to a two-way ANOVA with the factors category and truth value; except for categories colorless green and word lists as all stimuli in these categories were false. An effect of category was found in all experiments but MATHSEXPERTS, an effect of truth value was found in MATHSEXPERTS and an interaction was found in this experiment as well. The results of the ANOVA are presented in table 3.2.

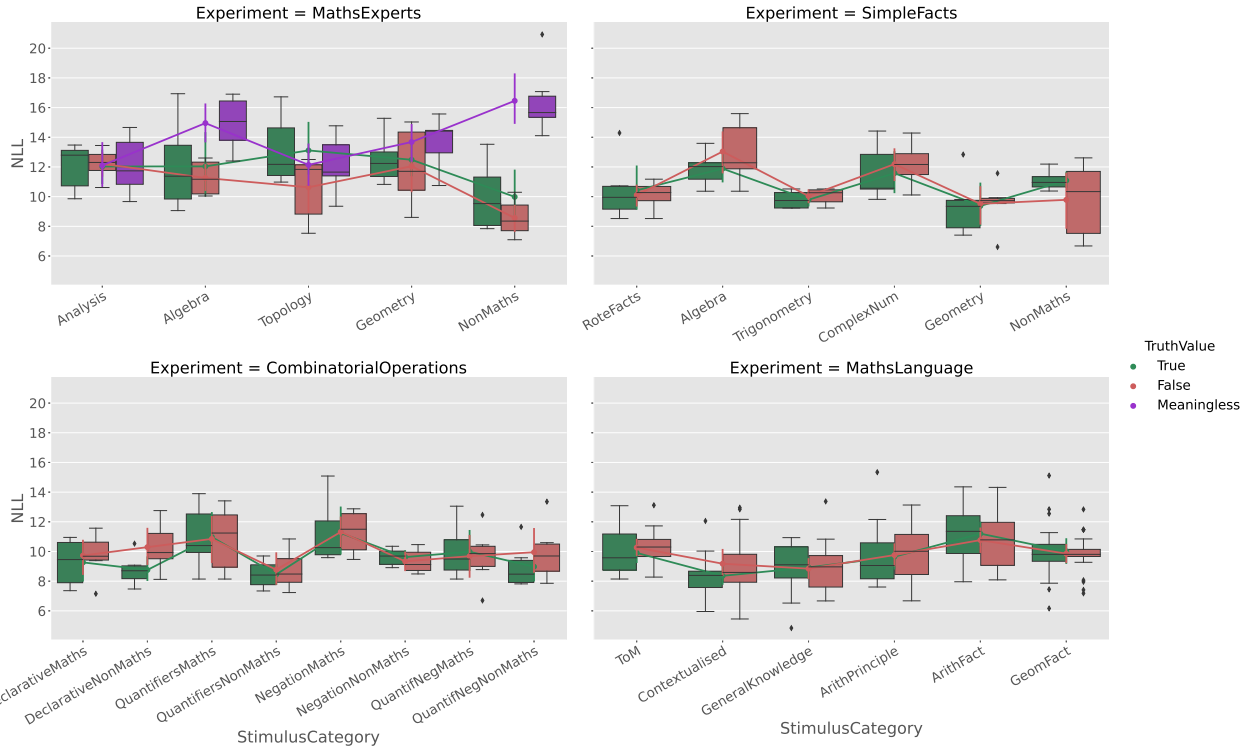


Fig. 3.5. GPT-fr's maximum negative log likelihood by stimulus as a function of stimuli's category and truth value.

Experiment	Effect of category	Effect of truth value	Interaction
MATHSEXPERTS	0.40	$\mathbf{1.3 \times 10^{-6}}$	$\mathbf{1.1 \times 10^{-4}}$
SIMPLEFACTS	$\mathbf{8.8 \times 10^{-5}}$	0.82	0.55
COMBINATORIALOPERATIONS	$\mathbf{3.0 \times 10^{-3}}$	0.32	0.85
MATHSLANGUAGE	$\mathbf{2.5 \times 10^{-8}}$	0.63	0.63

Table 3.2. ANOVA of GPT-fr's maximum negative log likelihood by stimulus with stimulus category and truth value as independent variables (in p-values).

Figure 3.5 suggests that the effect of truth value that was found in MATHSEXPERTS, as well as the huge interaction between category and truth value, is due to the presence of meaningless stimuli.

To confirm this hypothesis, the same ANOVA was performed on the meaningful stimuli only. This time, the only effect found was one of category for all experiments (see table 3.3). For the sake of readability, the plot associated with this ANOVA in only presented in appendix D.2 on page 68, since it is very close to figure 3.5 on the facing page. In addition, a one-way ANOVA of GPT-fr's maximum negative log likelihood by stimulus was run with meaning (i.e. meaningful or meaningless) as independent variable was run for MATHSEXPERTS and MATHSLANGUAGE. It revealed a very strong effect of meaning ($p = \mathbf{1.3 \times 10^{-5}}$ for MATHSEXPERTS and $p = \mathbf{9.2 \times 10^{-56}}$ for MATHSLANGUAGE).

Experiment	Effect of category	Effect of truth value	Interaction
MATHSEXPERTS	$\mathbf{3.8 \times 10^{-3}}$	0.065	0.56
SIMPLEFACTS	$\mathbf{8.8 \times 10^{-5}}$	0.82	0.55
COMBINATORIALOPERATIONS	$\mathbf{3.0 \times 10^{-3}}$	0.32	0.85
MATHSLANGUAGE	$\mathbf{2.5 \times 10^{-8}}$	0.63	0.63

Table 3.3. ANOVA of GPT-fr's maximum negative log likelihood by *meaningful* stimulus with stimulus category and truth value as independent variables (in p-values).

As one can see on figure D.2 on page 68, the effect of category seems to be due to the presence of both mathematical and non-mathematical statements. Indeed, as a rule of thumb, GPT-fr seems to perform much better on non-mathematical stimuli than on mathematical ones. To assess this impression, the same ANOVA was again performed to meaningful mathematical stimuli only. This time, an effect of category was found in SIMPLEFACTS and MATHSLANGUAGE and no effect of truth value or interaction were found at all (see table 3.4). In addition, a one-way ANOVA of GPT-fr's maximum negative log likelihood by stimulus was run with nature (i.e. mathematical or non-mathematical) as independent variable was run. It did not reveal an effect of nature in all experiments, but only in COMBINATORIALOPERATIONS and MATHSLANGUAGE ($p = 0.29$ for MATHSEXPERTS, $p = 0.53$ for SIMPLEFACTS, $p = \mathbf{1.3 \times 10^{-3}}$ for COMBINATORIALOPERATIONS and $p = \mathbf{5.8 \times 10^{-4}}$ for MATHSLANGUAGE). Since the strong effect for MATHSLANGUAGE might have been due to an interaction with meaning – as only non-mathematical stimuli contained meaningless statements – a second ANOVA was run on meaningful categories only and the effect persisted ($p = \mathbf{3.3 \times 10^{-5}}$) but this time, the maximum negative log likelihood was on average higher for non-mathematical than for mathematical stimuli.

Experiment	Effect of category	Effect of truth value	Interaction
MATHSEXPERTS	0.89	0.15	0.44
SIMPLEFACTS	$\mathbf{2.9 \times 10^{-5}}$	0.37	0.83
COMBINATORIALOPERATIONS	0.077	0.96	0.96
MATHSLANGUAGE	$\mathbf{3.8 \times 10^{-3}}$	0.62	0.79

Table 3.4. ANOVA of GPT-fr's maximum negative log likelihood by *meaningful mathematical* stimulus with stimulus category and truth value as independent variables (in p-values).

The results presented in table 3.4 indicate that GPT-fr is unable to distinguish between true and false mathematical statements, although it can most of the time detect nonsensical sentences as shown on figure 3.5 on the facing page. In addition, the effect of category found in experiments SIMPLEFACTS and MATHSLANGUAGE shows that GPT-fr does not perform equally well on all types of mathematics: it seems indeed to have overall a greater familiarity with simple geometry (in both experiments) than with elementary algebra for instance.

3.3.2 Comparison with human subjects

Given the limited ability of GPT-fr to discriminate between true and false mathematical statements (see section 3.3.1 on page 24), it is tempting to say that *GPT-fr behaves like a non-mathematician presented with mathematical stimuli.* This section assesses the extent to which this assertion holds.

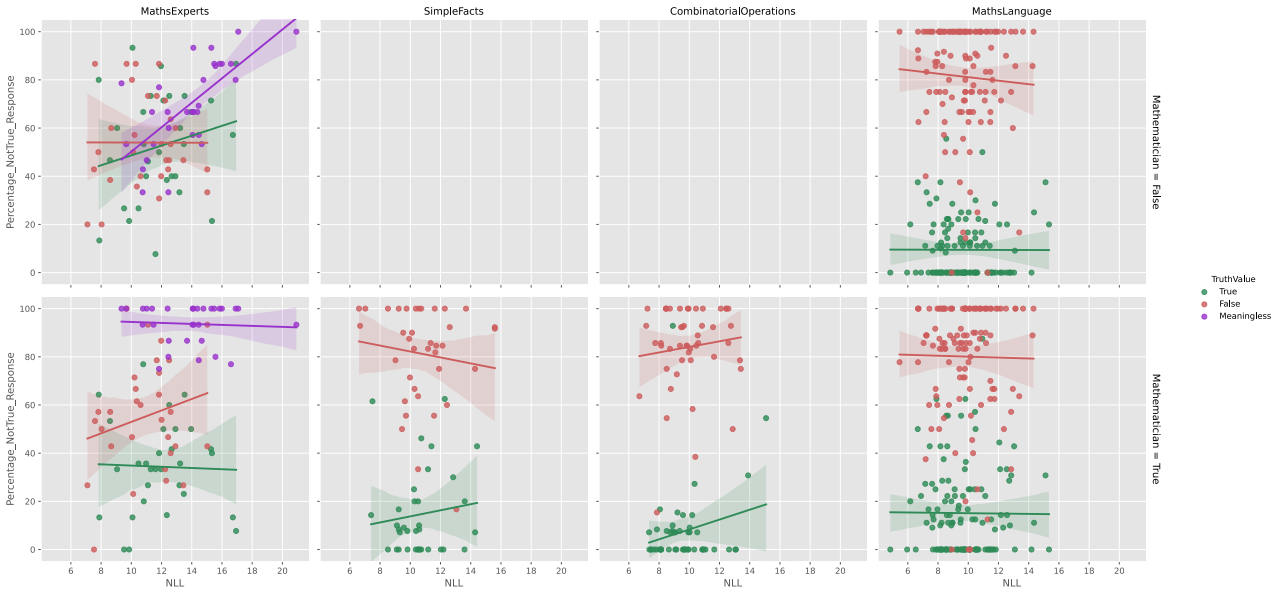


Fig. 3.6. Percentage of subjects judging the stimulus is not true against GPT-fr's maximum negative log likelihood (one point represents one stimulus).

Figure 3.6 plots, for each stimulus, the percentage of subjects who evaluted the stimulus was not true (i.e. false or meaningless when applicable) against GPT-fr's maximum negative log likelihood on said stimulus. The hypothesis that GPT-fr behaves like a non-mathematician would be confirmed if the correlation in the plots of the first row was high (typically $p < 0.05$). Indeed, the percentage of subjects who evaluated a stimulus was not true is arguably the most salient cue of subjects' surprisal and, if anything, should correlate well with GPT-fr's surprisal for the reasons discussed in section 3.2 on page 22.

	Experiment	True	False	Meaningless
Non-mathematicians	MATHSEXPERTS	0.27	0.99	$\mathbf{2.4 \times 10^{-5}}$
	MATHSLANGUAGE	0.97	0.54	
Mathematicians	MATHSEXPERTS	0.87	0.25	0.75
	SIMPLEFACTS	0.50	0.46	
	COMBINATORIALOPERATIONS	0.13	0.45	
	MATHSLANGUAGE	0.93	0.88	

Table 3.5. Significance of the correlation between the percentage of subjects evaluating a stimulus is not true and GPT-fr's maximum negative log likelihood on said stimulus (in p-values).

As shown in table 3.5, however, the two quantities are correlated only for the meaningless stimuli of MATHSEXPERTS. It therefore seems that the Transformer's ability for mathematics is nowhere near humans', whether they are mathematicians or non-mathematicians, except when human subjects do not understand the stimuli they are presented with.

As a sanity check of the hypothesis made in section 3.2.2 on page 23, the same ANOVA was also performed on perplexity rather than negative log likelihood and the results were no different, except for true statements presented to mathematicians in COMBINATORIALOPERATIONS ($p = 1.6 \times 10^{-3}$).

There seems to be no reason for this surprising state of affairs, which confirms the relevance of negative log likelihood over perplexity.

CHAPTER 4

Reanalysis of MATHSEXPERTS's fMRI Data

This chapter presents the reanalysis of the fMRI data of Amalric et al.'s MATHSEXPERTS experiment [AD16][1]. More precisely, the vocabulary created in chapter 2 on page 9 was used to assess whether the GloVe embedding of mathematical words helps predict brain activity depending on whether subjects are mathematicians or not.

4.1 Method: fitting the GLM

This section presents the method followed to analyse the fMRI data. All the first-level maps created can be found in the tarball of the project.

4.1.1 fMRI data acquisition and preprocessing

This section describes the settings in which the fMRI data were acquired and preprocessed by Marie Amalric. It is a summary of the method section of [AD16].

Runs. All statements were recorded by a female native French speaker who was familiar with mathematical concepts.

The experiment was divided into six runs of fifteen statements each, which included one exemplar of each subcategory of statements (e.g. analysis true, topology meaningless). On screen, the only display was a fixation cross on a black background. Each trial started with a beep and a colour change of the fixation cross (which turned to red), announcing the onset of the statement. After auditory presentation, a fixed-duration reflection period of 4 s allowed subjects to decide whether the statement was true, false, or meaningless. The end of the reflection period was signaled with a beep and the fixation cross turning to green. Only then, for 2 s, could subjects give their evaluation of the sentence (true, false, or meaningless) by pressing one of three corresponding buttons (held in the right hand). Each trial ended with a 4 s resting period.

Acquisition. Subjects were scanned using a 3 T whole body system (Siemens Trio) with a thirty-two-channel head-coil and high-resolution multiband imaging sequences developed by the Center for Magnetic Resonance Research (CMRR) [multiband factor: 4; Grappa factor: 2; interleaved axial

[1] The three other experiments SIMPLEFACTS, COMBINATORIALOPERATIONS and MATHSLANGUAGE were not reanalysed due to lack of time.

slices, 1.5 mm thickness and 1.5 mm isotropic in-plane resolution; matrix: 128×128; repetition time (RT): 1.500 ms, echo time (ET): 32 ms; scans per run: 200].

Preprocessing. Using `Matlab`'s `SPM8` software, functional images were first realigned, normalized to the standard Montreal Neurological Institute (MNI) brain space, and spatially smoothed with an isotropic Gaussian filter of 2 mm FWHM.

4.1.2 Stimuli onsets: words or sentences?

In Amalric et al.'s paradigm, only the onsets of sentences were used. So as to decide whether it was worth taking into account the precise words' onsets to generate the fMRI model regressors, a fake example was coded.

To estimate the difference between word or sentence onsets, five sentences of duration 4.6 s starting at 19 s intervals were considered. These parameters were chosen to fit with the design of MATHSEXPERTS. Sentences consisted of seven words each, each word starting 0.7 s after the previous one (the first word started at the beginning of the sentence). To emulate GloVe embeddings, a random weight in $[0,1]$ was assigned to each word, and an additional random weight in $[0,2]$ was added to each sentence so as to modulate signal intensity between sentences.

Two different GLMs were fitted: one with onsets for words and one with onsets for sentences. In the GLM with words onsets, the duration of each word was set to 0.7 s; in the GLM with sentences onsets, the duration of sentences was set to 4.6 s and the modulation of a sentence was defined as the average of its words' weights.

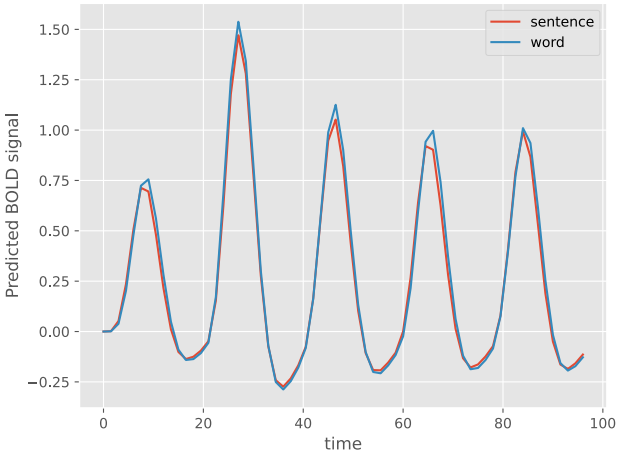


Fig. 4.1. Comparison of the predicted time series with onset for words or sentences after convolution with the HRF for random modulation.

On figure 4.1, the time series predicted by the two models are presented. One can see that there are very few differences between the red and blue plots (correlation: $r = 0.998$). This justifies that *the experiment will be analysed with onsets for sentences.*

4.1.3 GloVe embedding of stimuli

Before fitting a GLM, the GloVe embbeding of each stimuli was computed. To do so, stimuli were lemmatised twice using `spacy`'s `fr_core_news_md` in order to get only infinitive verbs and masculine singular adjectives and nouns, as was done with the mathematical corpus from which the vocabulary was extracted. For each lemma, an algorithm checked whether it belonged to the

vocabulary – the mathematical one for lemmas belonging to mathematical stimuli and the non-mathematical one for other lemmas – and, if not, used the `french_lefff_lemmatizer` lemmatiser[2] to check if the predicted noun, verb or adjective form of the lemma did belong to the vocabulary. For each of the three principal components, weights of the recognised lemmas were then averaged within stimuli, yielding the weight of each stimulus.

Regarding the mathematical vocabulary, an attempt was made to use the thousand-word vocabulary from chapter 2 on page 9 first. However, it turned out that this vocabulary was not large enough to contain all words of the stimuli: for mathematical stimuli, only 32.8% of non-punctuation lemmas belonged to the vocabulary on average. So as to obtain a more representative embedding of mathematical stimuli, it was decided to perform PCA on the whole 39,345-word vocabulary provided by the GloVe pipeline (see section 2.1.2 on page 10) and use the first three principal components of this vocabulary to compute the embedding of mathematical stimuli. This method allowed for stimuli's embeddings to be computed using on an average of 98.0% of their lemmas.

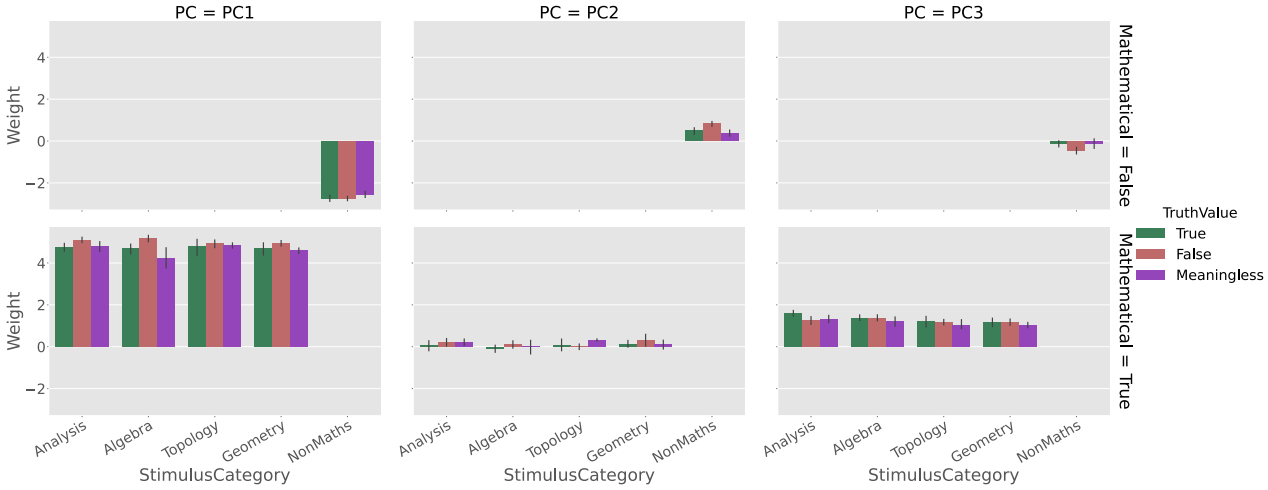


Fig. 4.2. Projection of the stimuli's embedding onto PC1, PC2 and PC3 as a function of their category and truth value (bars show 95% confidence intervals).

As explained in chapter 2 on page 9, the principal components of the mathematical vocabulary are not easily interpretable. However, they might be able to make a distinction between algebra, analysis, topology and geometry – the bars for non-mathematical stimuli do not come from the same GloVe embedding. Figure 4.2 shows that there is no reason to hope here either, even with the global vocabulary. This might either mean that the principal component of GloVe embeddings do not capture enough mathematical semantics to distinguish between types of mathematics, or that GloVe's representation of mathematics transcends ours.

4.1.4 Design matrices

In the first-level analysis, the data were predicted with three kinds of regressors. First, regressors on noninterest include the six movement parameters and two categorical regressors for alerting sound and motor response (when the subject presses a button to evaluate the truth value of the sentence). Regressors of interest are divided into two categories: four categorical regressors indicating whether a sentence is meaningful mathematical, meaningful non-mathematical, meaningless mathematical or meaningless non-mathematical – one during the reading of the sentence and one during the reflection period –, and, for each of these conditions, three parametric regressors for the three first principal components. All regressors of interest have kernel equal to sentence duration, except for

[2] `https://github.com/ClaudeCoulombe/FrenchLefffLemmatizer`

the categorical regressors during the reflection period which have kernel equal to 4 s. The design matrix for one subject is presented on figure 4.3.

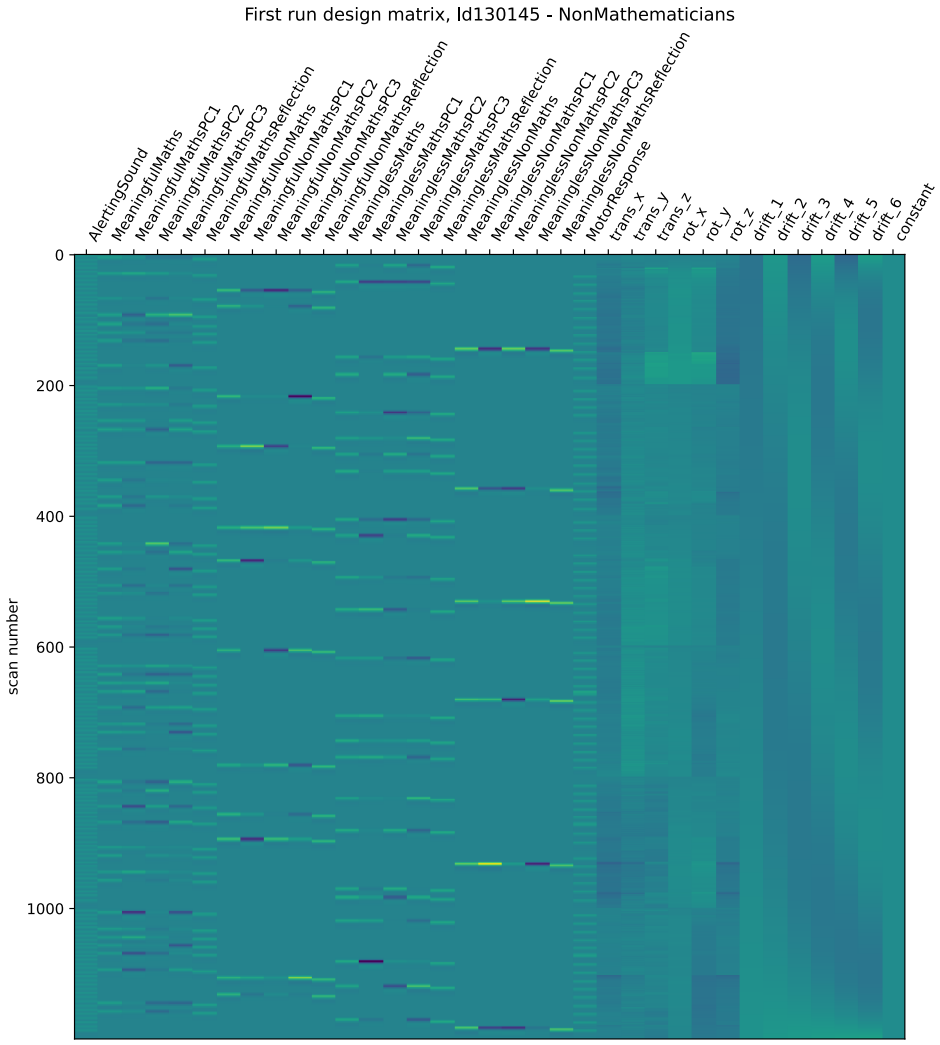


Fig. 4.3. Design matrix of a (non-mathematician) subject.

As it happens, the design matrices thus defined were problematic. Indeed, for each given meaning and type of stimuli (e.g. meaningful and mathematical), the predicted time series for the categorical and the three associated parametric regressors were highly correlated (for instance, in the first run of the first subject, twenty-three out of thirty-five variance inflation factors were greater than ten, and eight of them were infinite), even though the projections of vocabulary words onto the three first principal components were not correlated at all. This problem leads to uninterpretable maps and needed to be solved. It was thus decided to merge all runs, to standardise each parametric regressor and to orthogonalise them within each set:

- $\{\mathsf{MeaningfulMaths}\} \cup \{\mathsf{MeaningfulMathsPC}i, i \in [\![1,3]\!]\}$,
- $\{\mathsf{MeaninglessMaths}\} \cup \{\mathsf{MeaninglessMathsPC}i, i \in [\![1,3]\!]\}$,
- $\{\mathsf{MeaningfulNonMaths}\} \cup \{\mathsf{MeaningfulNonMathsPC}i, i \in [\![1,3]\!]\}$,
- $\{\mathsf{MeaninglessNonMaths}\} \cup \{\mathsf{MeaninglessNonMathsPC}i, i \in [\![1,3]\!]\}$,

using the Gram-Schmidt procedure:

$$\mathrm{MM} \leftarrow \mathrm{MM} \tag{4.1}$$

$$\mathrm{MMPC1} \leftarrow \mathrm{MMPC1} - \frac{\langle \mathrm{MM}|\mathrm{MMPC1}\rangle}{\|\mathrm{MM}\|^2}\mathrm{MM} \tag{4.2}$$

$$\mathrm{MMPC2} \leftarrow \mathrm{MMPC2} - \frac{\langle \mathrm{MM}|\mathrm{MMPC2}\rangle}{\|\mathrm{MM}\|^2}\mathrm{MM} - \frac{\langle \mathrm{MMPC1}|\mathrm{MMPC2}\rangle}{\|\mathrm{MMPC1}\|^2}\mathrm{MMPC1} \tag{4.3}$$

$$\mathrm{MMPC3} \leftarrow -\frac{\langle \mathrm{MM}|\mathrm{MMPC3}\rangle}{\|\mathrm{MM}\|^2}\mathrm{MM} - \frac{\langle \mathrm{MMPC1}|\mathrm{MMPC3}\rangle}{\|\mathrm{MMPC1}\|^2}\mathrm{MMPC1} - \frac{\langle \mathrm{MMPC2}|\mathrm{MMPC3}\rangle}{\|\mathrm{MMPC2}\|^2}\mathrm{MMPC2} \tag{4.4}$$

where MM is a condition (e.g. MeaninglessMaths or MeaningfulNonMaths) and $\langle\cdot|\cdot\rangle$ denotes the canonical inner product in $\mathbf{R}^{\text{total number of scans}}$. This procedure *impacts the interepretability of the weights the GLM assigns assign to each regressor in section 4.2*: for example, the weight associated with MeaningfulMathsPC1 will take into account the parallel components of MeaningfulMathsPC2 and MeaningfulMathsPC3.

Three second-level analyses were performed, each using a different design matrices: one with mathematicians only, one with all subjects, and one with a column for mathematicians and a column for non-mathematicians.

4.2 Contrasts, results & discussion

This section aims at answering two main questions:

(i) Where is there an effect, if any, of the meaningful maths principal components in mathematicians?

(ii) Where is there an effect, if any, of the meaningful non-maths principal components in all subjects?

4.2.1 Retrieving Amalric et al.'s mathematical network in mathematicians

As a sanity check, the correctness of the model was probed by trying to retrieve Amalric et al.'s mathematical network in mathematicans [AD16]. Since the same data are used here and in Amalric et al.'s paper, the replication should be no problem.

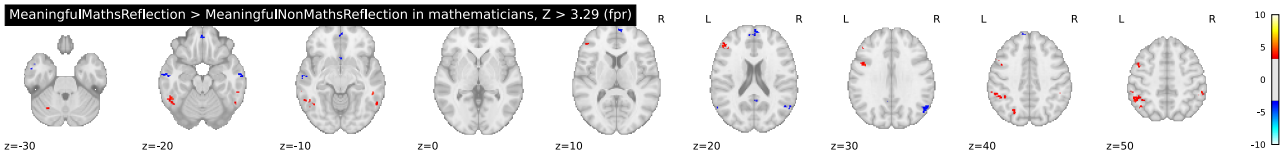
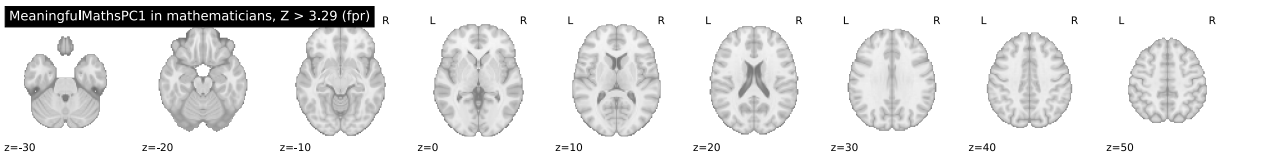
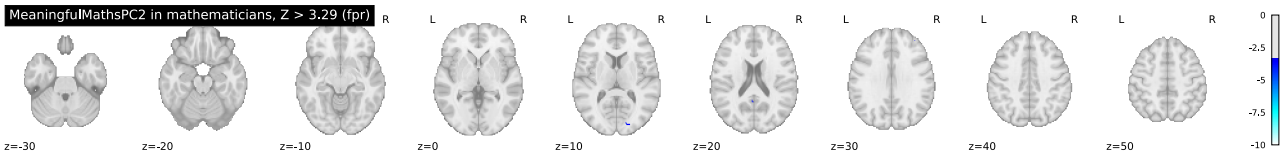


Fig. 4.4. Second level contrast MeaningfulMathsReflection > MeaningfulNonMathsReflection in mathematicians, FPR correction $q < 10^{-3}$.

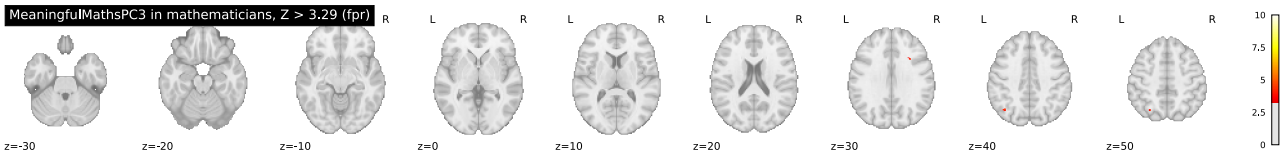
Figure 4.4 shows the contrast MeaningfulMathsReflection > MeaningfulNonMathsReflection in mathematicians only and is very similar to that reported by Amalric et al., with activations in the bilateral intraparietal sulci, the bilateral inferior temporal regions, the bilateral dorsolateral, superior, and mesial prefrontal cortex, and the cerebellum. It thus seems that the model defined in 4.1.4 on page 31 has no obvious flaws, which reinforces the credit that can be given to subsequent analyses.



(a) Effect of MeaningfulMathsPC1.



(b) Effect of MeaningfulMathsPC2.



(c) Effect of MeaningfulMathsPC3.

Fig. 4.5. Effect of MeaningfulMathsPCs in mathematicians, FPR correction $q < 10^{-3}$, no significant effect detected.

4.2.2 Effect of principal components for meaningful mathematical stimuli

This section is devoted to question (i) on the preceding page.

Figure 4.5 presents the group analysis of the effect of MeaningfulMathsPCs in mathematicians. One can see that there very few – if any – activations on the maps, even for the first principal component. In addition, thorough analysis of individual maps confirmed that the absence of effects at the group level does not result from significant effects in different regions for each subject.

So as to further investigate the effects of MeaningfulMathsPCs, a ROI analysis was performed in spheres of radius 1 cm around centres identified by Amalric et al. [AD16] (see table D.1 on page 69). In order not to look for signal outside of the brain, the spheres masker defined by the ROIs was intersected with the average mask of all subject (threshold: 50%).

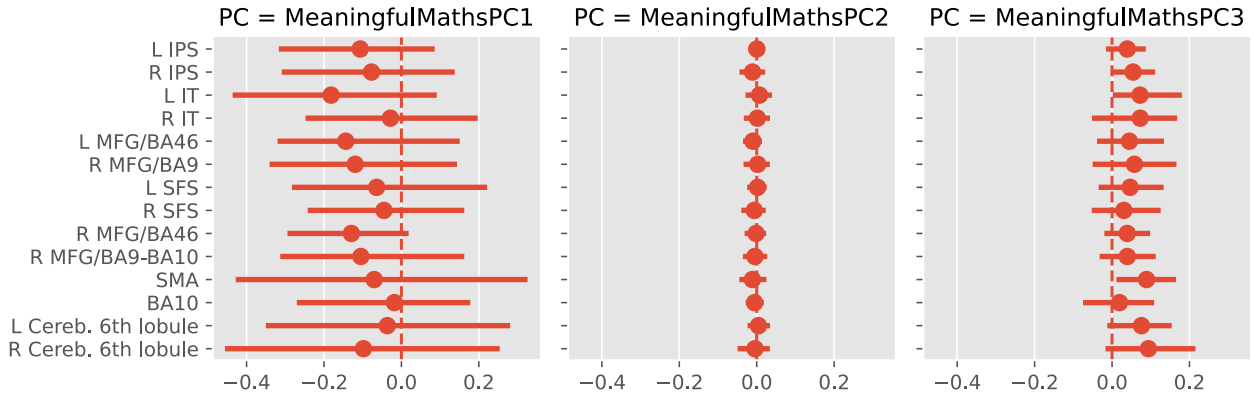
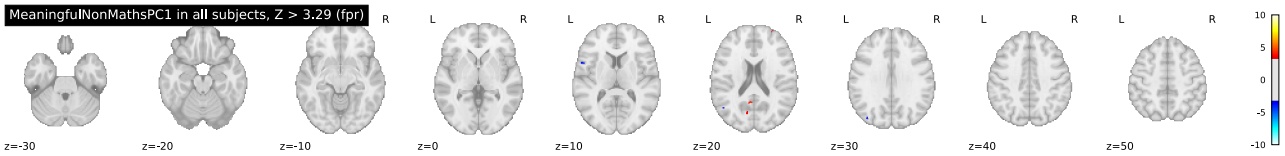


Fig. 4.6. ROI analysis (95% confidence intervals) of the MeaningfulMathsPCs in mathematicians, Bonferroni corrected.

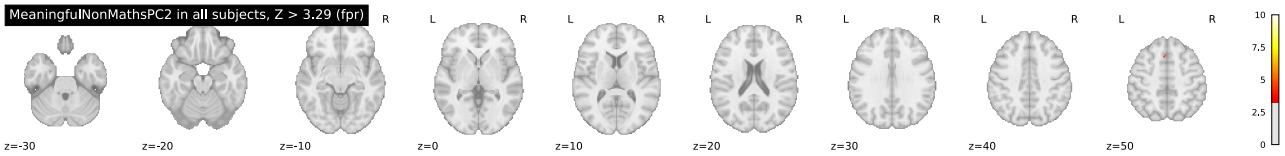
Figure 4.6 shows the 95% confidence intervals of the model's weight associated to each MeaningfulMathsPC in the ROIs defined in table D.1 on page 69 in mathematicians. One can see that, even in the ROIs, there is no significant effect of the MeaningfulMathematicalPCs, as all confidence intervals contain 0. It thus seem that the MeaningfulMathematicalPCs provide no additional information to that of a mere categorical regressor.

4.2.3 Effect of principal components for meaningful non-mathematical stimuli

This section is devoted to question (ii) on page 33.



(a) Effect of MeaningfulNonMathsPC1.



(b) Effect of MeaningfulNonMathsPC2.

(c) Effect of MeaningfulNonMathsPC3.

Fig. 4.7. Effect of MeaningfulNonMathsPCs in all subjects, FPR correction $q < 10^{-3}$, no significant effect detected.

Figure 4.7 presents the second level analysis of the effect of MeaningfulNonMathsPCs in all subjects. Here again, one can see on the brain maps that there is no effect of the MeaningfulNonMathsPCs; thus suggesting that the first principal components of GloVe embeddings carry no information, which was already expected from the analyses and discussions in sections 2.2.1 on page 11 and 4.1.3 on page 30.

4.3 Second model

As the first model presented in section 4.1 on page 29 did not enable the detection of any significant effect of the first principal components of GloVe mathematical embeddings on the fMRI activations, an attempt was made to fit a second GLM using a general GloVe model (trained on all corpora, mathematical and non-mathematical).

This time, the embedding was that of the global vocabulary obtained from the corpus made up of both mathematical and non-mathematical corpora. The spirit of the approach is different: the PCA was performed on the embedding of stimuli directly – as the aim is to analyse the embedding of stimuli rather than that of the vocabulary –, and no regressor was there to make the distinction between mathematical and non-mathematical statements: the principal components are here to make this distinction. Therefore, the question asked here is slightly different than that of section 4.2 on page 33 and can be stated as follows: *do the first principal components of the global GloVe embedding make a distinction between mathematical and non-mathematical stimuli, and do they make it possible to retrieve the mathematical network reported by Amalric et al. [AD16] in mathematicians?*

A first step toward an answer to the question is provided by figure 4.8 on the next page. One can see on this figure that the embeddings of algebraic, analytical, topological and geometrical statements are similar, but very different from those of non-mathematical statements, especially for the projection onto PC1. Thus, there are reasons to hope that the model will help predict the distinction between the mathematics and the language networks.

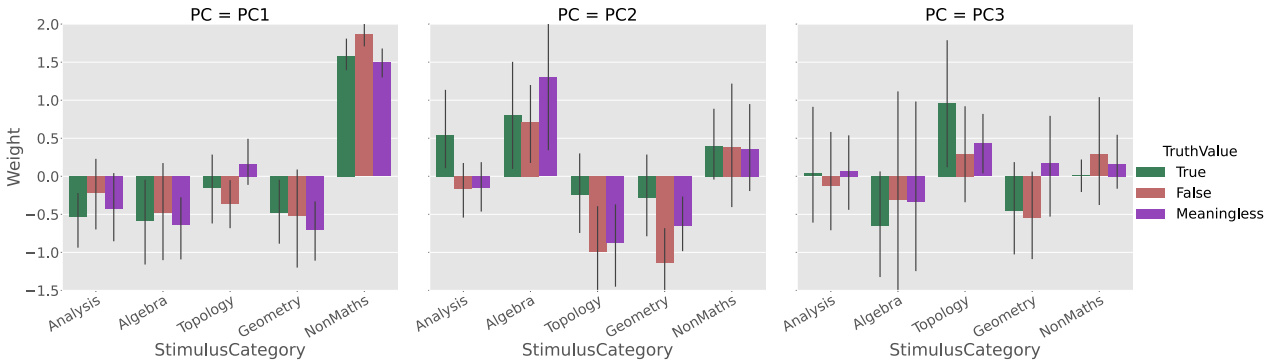


Fig. 4.8. Projection of the stimuli's embedding onto PC1, PC2 and PC3 as a function of their category and truth value (bars show 95% confidence intervals).

The design matrix of the model is quite similar to that of the model of section 4.1 on page 29. Runs have been merged, but this time, there was no need to orthogonalise regressors as the PCA already provides orthogonal regressors. Regressors of noninterest are the same as in section 4.1.4 on page 31, but this time only four categorical regressors (Meaningful and Meaningless twice: during the reading of the sentence [kernel: sentence duration] and during the reflection period [kernel: 4 s]); and three parametric regressors: the three first principal components of the embeddings for meaningful stimuli only. The design matrix for one subject is presented on figure 4.9 on the facing page.

As shown on figure 4.5 on page 34, MeaningfulPC1 has an effect in some of the mathematical regions identified by Amalric et al. [AD16] (see figure 4.4 on page 33). However, MeaningfulPC2 and MeaningfulPC3 have no significant effect. This, along with figure 4.8 prompts to hypothesise that MeaningfulPC1 carries information about whether a stimulus is mathematical or not.

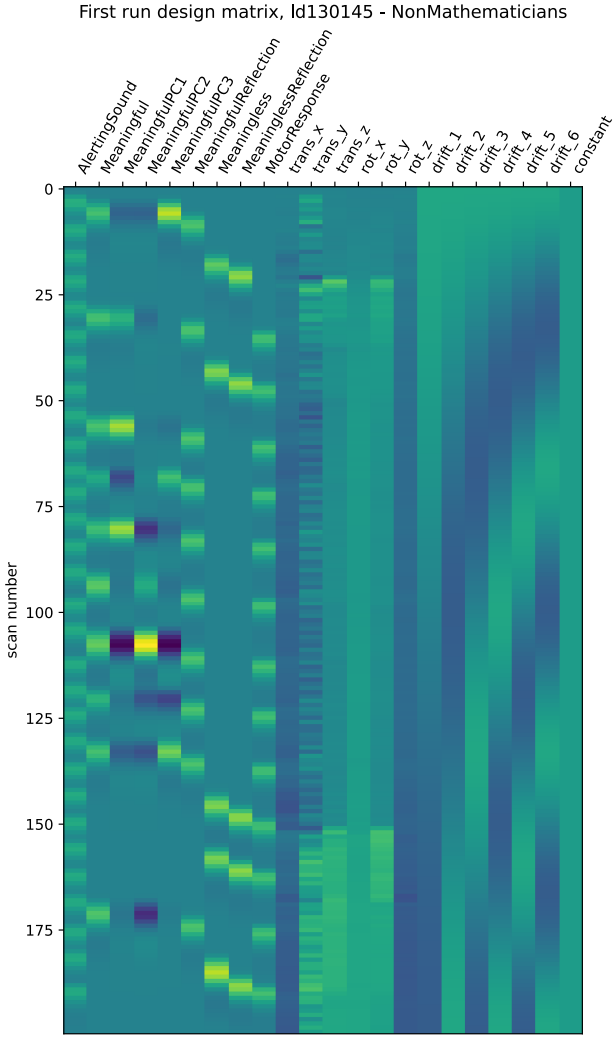
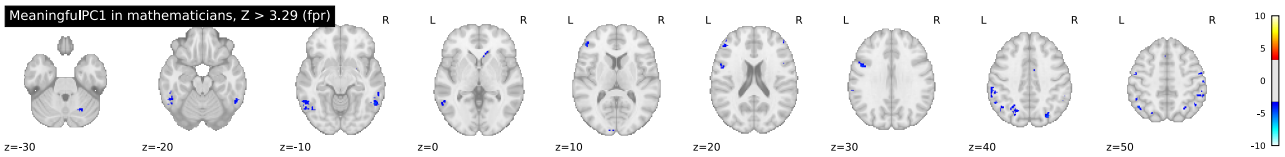
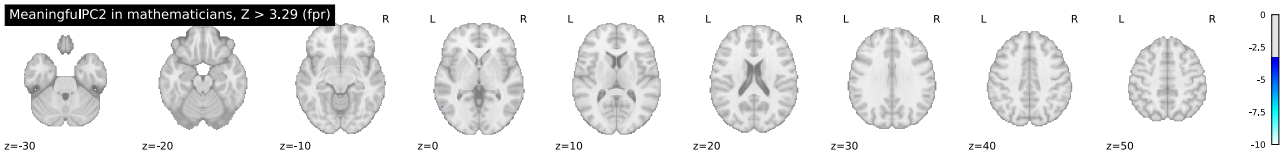


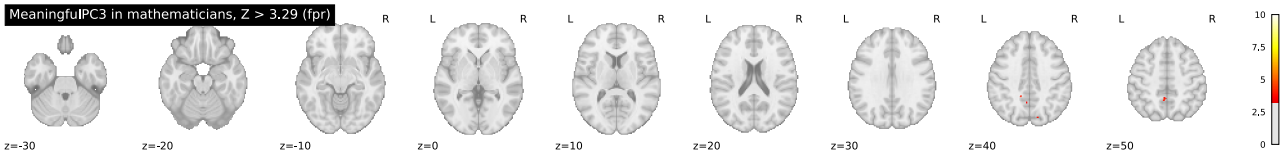
Fig. 4.9. Design matrix of a (non-mathematician) subject for the second model.



(a) Effect of MeaningfulPC1.



(b) Effect of MeaningfulPC2.



(c) Effect of MeaningfulPC3.

Fig. 4.10. Effect of MeaningfulPCs in mathematicians, FPR correction $q < 10^{-3}$.

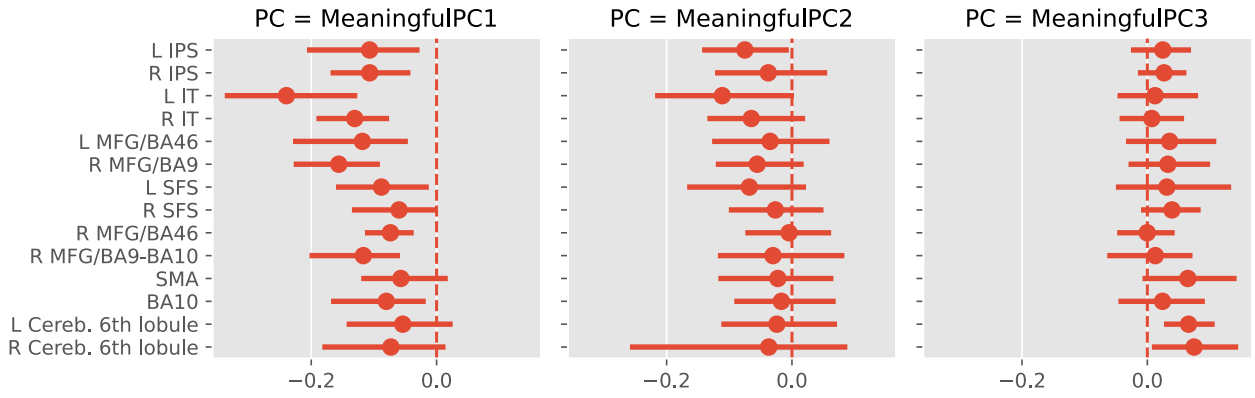


Fig. 4.11. ROI analysis (95% confidence intervals) of the MeaningfulPCs in mathematicians, Bonferroni corrected.

To probe this hypothesis, a ROI analysis was conducted like in section 4.2 on page 33 and all ROIs but R SFS, SMA and Cerebellum show a significant effect of MeaningfulPC1 (see figure 4.11 for the 95% confidence intervals of the model's weight associated to the ROIs defined in table D.1 on page 69).

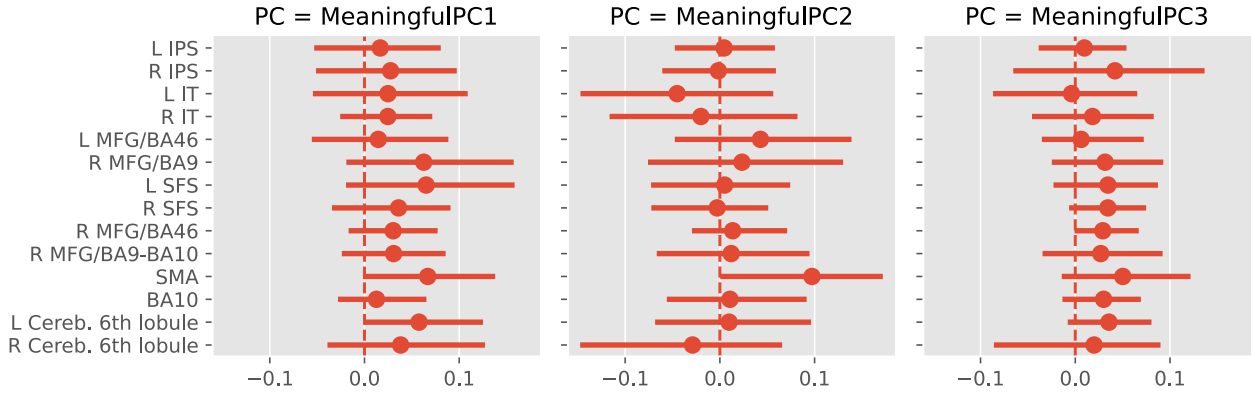


Fig. 4.12. ROI analysis (95% confidence intervals) of the MeaningfuPCs in controls, Bonferroni corrected.

By way of comparison, the same ROI analysis was performed on non-mathematicians, and no effect of MeaningfulPC1 was found. It therefore seems that MeaningfulPC1 does indeed carry information about mathematics. Indeed, if the piece of information carried by MeaningfulPC1 was orthogonal to both mathematics and non-mathematics, it would have a significant effect in both or neither controls and mathematicians.

CHAPTER 5

General Discussion

5.1 Semantic clustering of mathematics

Chapter 2 on page 9 evaluates how well GloVe captures the mathematical semantics of mathematical words. The clustering performed in sections 2.2.2 on page 16 and 2.2.3 on page 16 show that GloVe does capture a fair amount of mathematical semantics, as spectral and hierarchical clusterings bring out an interesting classification of mathematics. It also shows that the method used by Pereira et al. to sample the semantic space is quite stable, albeit not optimal since agglomerative clustering seems to provide a sharper clustering than spectral clustering.

However, the variance of the GloVe embedding of the vocabulary spans over many directions, which results in not very explanatory principal components, which, in addition, are poorly interpretable. There is certainly some work to be done to find a way to satisfactorily reduce the dimension of GloVe embeddings.

5.2 The Transformer's mathematical abilities

In chapter 3 on page 19, one proved that GPT-fr is able to make the distinction between meaningful and meaningless statements, and that its behaviour correlates with that of non-mathematicians when presented with meaningless statements. However, it seems that the mathematical abilities of the default GPT-fr model (i.e. with no additional training) of the Transformer do not go beyond this distinction and are thus very limited.

Theses results are a priori contrary to those presented in section 1.2.2 on page 3 [LC19; dAs+22; Kam+22; PS20; Pol+22; CHL21; Cha21]. However, these works and the present project differ in two ways. First, in all works presented in section 1.2.2 on page 3, the Transformer was retrained on additional mathematical data, which is not the case here. Furthermore, the assessment of the Transformer is different in both cases: in this project, the Transformer was fed with a statement and its perplexity was used to probe its ability to discriminate between true and false statements. In other projects, however, the Transformer was asked to make an actual prediction of the result of a computation or a proof tree, which are much guided tasks. These two observations might explain the poor performance of the Transformer found in chapter 3 on page 19.

As far as the Transformer's poor predictions of mathematical cognition are concerned, the findings of chapter 3 on page 19 go against those of Schrimpf et al. [Sch+21] for natural language processing. Indeed, Schrimpf et al. showed that the Transformer is good at predicting subjects'

behaviour (reading time in their case), no matter its training. These different results, however, are probably due to the fact that natural language processing is much more native a task for the Transformer than mathematics.

5.3 GloVe as a predictor of the brain's processing of mathematics

It seems that the GloVe embeddings of the mathematical vocabulary do not provide more information than a mere categorical regressor when it comes to analysing fMRI data. However, one proved in section 4.3 on page 35 that the first principal component of the GloVe embeddings of the global vocabulary makes a clear distinction between mathematical and non-mathematical stimuli, and enables to retrieve Amalric et al.'s [AD16] mathematical network.

Nevertheless, the approach adopted in chapter 4 on page 29 is less promising than that of Huth et al. [Hut+16], which consists in performing PCA across voxels. The bad performance of the first model presented in section 4.1 on page 29 may be imputed to two factors. First, Schrimpf et al. [Sch+21] argue that GloVe is not a good model per se to predict brain activity, as it does not perform well at next-word-prediction tasks. Second, it might be that PCA is too brutal a way to reduce the dimensionality of the GloVe embeddings, and a solution might be found just by using other methods (e.g. tSNE) or by not reducing the dimensionality and using ℓ_2-penalised (Ridge) regression rather than standard GLMs.

Finally, it should be noted that Huth et al. [Hut+16] report *intra-subject* semantic atlases obtained after hours of fMRI scanning, while chapter 4 on page 29 presents group analyses of subjects scanned on a limited number of stimuli. It thus seems reasonable that the results presented here are less impressive than those of Huth et al.

5.4 Conclusions, limitations and future directions

5.4.1 How useful is AI to understand our understanding of mathematics?

Although AI has made tremendous progress in the last few years, this project did not allow full use to be made of AI's mathematical and brain-predicitive abilities. However, it made it clear – if it were necessary – that AI can indeed capture a fair amount of mathematical semantics and even retrieve the difference between mathematical and non-mathematical statements. As it was very limited in duration, this project was only a first step toward a full seizure of AI's ability to predict mathematical cognitions and there are still many uncharted paths to explore.

5.4.2 Limitations & deviations from pre-registration

The results presented in the project should be interpreted in the light of a few remarks.

First, all analyses were performed on data in French. This has many undesirable consequences, the first one being that very few resources are available. In particular, the mathematical corpus extracted in chapter 2 is only 66.8 MiB large. Similarily, in addition to not being retrained, the Transformer model GPT-fr used in chapter 3 on page 19 is not as sophisticated as English state-of-the-art models like GPT-3.

Progress could also be made by changing the paradigm of behavioural and fMRI experiments. Indeed, models of the Transformer derived from BERT and GPT are often used as language generators rather than just perplexity calculators – and they have proved to perform very well at the former task. Thus, a paradigm in which subjects are asked to complete a statement or comment on it can be expected to enable greater convergence between models and subjects, although it would require more thinking regarding practical implementation and comparison criteria.

Besides, the brain maps presented in chapter 4 on page 29 should be read cautiously, as the regressors have been orthogonalised before fitting the GLM. As the interpretations that were made do not involve individual weights of the model for principal components, there however seems to be no problem with the discussion of the maps presented.

The project has diverged from the pre-registered one at several moments and for various reasons. First, the poorly interpretable results and clusters in 2 on page 9 lead to reducing the dimension of GloVe embeddings from 300 to 50; and to modify the way the mathematical vocabulary is extracted (not by considering only the words which are ten times more frequent in the mathematical corpus than in the non-mathematical one, as discussed in section 2.1.2 on page 10). It was also decided to perform agglomerative clustering in addition to spectral clustering in chapter 2 on page 9, as this representation seemed attractive and unknowned from us at the time the preregistration was submitted. Another flaw in the initial project that was corrected on-the-fly was that it was not planed to make a distinction between meaningful and meaningless statements in the analysis of fMRI data. This might have been problematic as Amalric et al.'s [AD16] results suggest big differences in the treatment of meaningful and meaningless statement. Finally, as an internship is time-limited, modifications had to be made to the project so as to finish on schedule:

- only the mathematical vocabulary was analysed in chapter 2 on page 9, as the two other vocabularies were of lesser importance for the analysis of fMRI data;
- only one fMRI experiment was analysed and the contributions of principal components were then assessed using contrast, this approach is less easily interpretable and less common in the literature than the one planed initially with the brain-score-like approach;
- due to poor sampling of stimuli with the thousand-word vocabulary presented in chapter 2 on page 9, embedding of stimuli for fMRI analysis were computed using the whole (noisy) mathematical vocabulary output by the GloVe pipeline;
- due to unexpected correlation problem, runs had to be merged and the design matrix of fMRI data analysis had to be orthogonalised.

5.4.3 Possible continuations

The first obvious continuation of this internship is the reanalyse the other fMRI experiments SIMPLEFACTS, COMBINATORIALOPERATIONS and MATHSLANGUAGE. All the work presented in this thesis could also be redone in English, where much more ressources are available, with for instance more sophisticated models of the Transformer.

Besides, the principal components of the GloVe embedding of the global vocabulary do not enable a more fine-grained analysis of fMRI data than a mere categorical model. This might be due to the dimension reduction performed by the PCA, as the principal components are not interpretable (see section 2.2.1 on page 11). This problem might be solved by using other methods to lower the dimensionality of the GloVe embeddings (e.g. tSNE or maybe even asking for three-dimensional GloVe vectors directly), or by performing PCA across voxels like Huth et al. [Hut+16].

In the longer term, it could be interesting to perform a second analysis of the fMRI data making the distinction between mathematicians and non-mathematicians. A GloVe embedding for the global (mathematical and non-mathematical) corpus could be computed and used for mathematicians, while only the non-mathematical embedding would be used for non-mathematicians, with random weights for unknown mathematical words. The comparison of models using these two embeddings would yield a comparison between the mathematical networks of mathematically-educated and uneducated people, also probing the differences in GloVe embeddings for ambiguous words such as "group".

Finally, training a Transformer's model instead of using a pretrained model could help improve performance at predicting subjects' behaviour. If so, the hidden states of the model should better predict brain activity on mathematical tasks. In the spirit of Pereira et al. [Per+18] the analysis of

fMRI data thanks to these hidden states – or maybe GloVe embeddings, although is seems less likely to succeed – could also lead to the training of a decoder able to predict subjects' behaviour (i.e. their evaluation of a statement as true, false or meaningless) from brain activation.

APPENDIX A

Mathematical Toolbox

So as to make this thesis somewhat self-contained, this appendix explains briefly the mathematical and computational tools that were used throughout the project. The aim is not at all to make a technical presentation of these tools or their mathematical foundations, but rather to provide a working knowledge of them and, incidentally, to justify their use.

A.1 GloVe vectors for word representation

This section briefly explains how GloVe vectors are built, see [PSM14] for more details.

GloVe is a model designed to capture words' semantic features in a given corpus, inspired by latent semantic analysis [Dee+90] and local context window methods [MYZ19]. Roughly speaking, given the cooccurrence matrix of a vocabulary in a corpus, it extracts relevant cues of the matrix and outputs global vectors – "GloVe" actually stands for *Glo*bal *Vec*tors – for words of the vocabulary in an embedding space of any desired dimension.

More precisely, consider a vocabulary $\{x_1, \dots, x_n\}$ with cooccurrence matrix X where $X_{i,j}$ is the number of times x_j appears in the context (for a definition of "context" chosen by the user) of x_i. Then the probability of x_j appearing in the context of x_i is given by

$$P_{i,j} := \Pr(j \mid i) = \frac{X_{i,j}}{X_i}$$

where $X_i := \sum_{j=1}^{n} X_{i,j}$. The main idea behind GloVe is that *ratios of cooccurrence probabilities are more informative than cooccurrence probabilities themselves.* For example, when both or neither x_i and x_j are of interest in the context of a word x_k, $P_{i,k}/P_{j,k}$ will be close to 1, whereas it will be either very close to 0 or to $+\infty$ if only x_i or x_j is of interest. Therefore, by chosing wisely the word x_k, one can extract semantic information about x_i and x_j.

Now, since $P_{i,k}/P_{j,k}$ only depends on x_i, x_j and x_k, there must exist a function F such that

$$\frac{P_{i,k}}{P_{j,k}} = F(w_i, w_j, \tilde{w}_k) \tag{A.1}$$

where $w_i, w_j \in \mathbf{R}^d$ are word vectors encoding x_i and x_j and $\tilde{w}_k \in \mathbf{R}^d$ is a vector encoding the context word x_k. The difference of notation between $\{w_i, w_j\}$ and $\tilde{w}_k$ is here to highlight the fact that $\{x_i, x_j\}$ and x_k do not play the same role in $P_{i,k}/P_{j,k}$. To determine F uniquely, it has been decided to impose the following constraints:

1. F must encode $P_{i,k}/P_{j,k}$ in the word vector space $\mathbf{R}^d$, thus it seems only natural to expect that

$$\frac{P_{i,k}}{P_{j,k}} = F(w_i - w_j, \tilde{w}_k), \tag{A.2}$$

   linearity plays an important role for instance when comparing "early" and "late", and "breakfast" and "supper": one could expect $w_{\text{late}} - w_{\text{early}}$ and $w_{\text{supper}} - w_{\text{breakfast}}$ to be pretty similar;
2. since F is a linear form, it would be more convenient if its domain was also $\mathbf{R}$, which can be achieved by considering $F \leftarrow F \circ \langle \cdot | \cdot \rangle$, that is to say

$$\frac{P_{i,k}}{P_{j,k}} = F\left((w_i - w_j)^\top \tilde{w}_k\right); \tag{A.3}$$

3. finally, since the cooccurrence matrix X makes no difference between words and context words, F must be invariant under $w \leftrightarrow \tilde{w}$ and also under $X \leftrightarrow X^\top$.

Pennington et al. [PSM14] proved that, under the above conditions, equation A.1 on the previous page rewrites as

$$w_i^\top \tilde{w}_k + b_i + \tilde{b}_k = \log(X_{i,k}) \tag{A.4}$$

where b_i and $\tilde{b}_k$ are bias terms for w_i and $\tilde{w}_k$. The aim is thus to find vectors w_i and $\tilde{w}_k$ which make the difference of the terms in equation A.4 as small as possible.

To solve the ill-defined logarithm problem (it may be that $X_{i,k} = 0$) and disadvantage rare noisy cooccurrences (i.e. small $X_{i,k}$) over more common ones, the authors propose to transform A.4 into the following weighted least squares problem

$$J = \sum_{i=1}^{n} \sum_{j=1}^{n} f(X_{i,j}) \left(w_i^\top \tilde{w}_j + b_i + \tilde{b}_j - \log(X_{i,j})\right)^2 \tag{A.5}$$

where f is a weighting function empirically optimised

$$f(x) = \begin{cases} \left(\frac{x}{100}\right)^{3/4} & \text{if } x < 100 \\ 1 & \text{otherwise} \end{cases}. \tag{A.6}$$

GloVe implementations use unsupervised learning to solve the weighted least square problem defined by equation A.5.

A.2 Spectral clustering

One of the most common and intuitive ways of clustering data into a given number k of clusters is to use the k-means algorithm. Given a set of observations $D := \{x_1, \dots, x_n\} \subseteq \mathbf{R}^d$ and a target number $k \leq n$ of clusters, this procedure aims at finding a partition $S := \amalg_{i=1}^{k} S_i$ of D which satisfies

$$S = \arg \min_{\substack{X = \amalg_{i=1}^{k} X_i \\ X \text{ partition of } D}} \sum_{i=1}^{k} \sum_{x \in X_i} \|x - \mu(X_i)\|^2 \tag{A.7}$$

where $\mu(X_i)$ denotes the centroid of X_i and $\|\cdot\|$ can be any norm (usually the Euclidean one).

However, Pereira et al. [Per+18] argue that the assumptions of the k-means algorithm are not realistic in the context of lexical features and suggest spectral clustering as a workaround. There are quite a few possible implementions of spectral clustering, all presented in [Lux07]. The one presented here is that used by Pereira et al. in their 2018 paper [Per+18], which resorted to clustering in conditions similar to those of this project.

Spectral clustering basically consists in preprocessing the observations so that the assumption required by k-means is indeed verified. The preprocessing consists of the following steps:

(1) Compute the cosine similarity $C \in \mathscr{M}_{n\times n}(\mathbf{R})$ such that

$$C_{i,j} = \frac{\langle x_i | x_j \rangle}{\|x_i\| \|x_j\|} \tag{A.8}$$

where $\langle \cdot | \cdot \rangle$ denote the canonical inner product in $\mathbf{R}^d$.

(2) For every $(i,j) \in [\![1,n]\!]^2$, do $C_{i,j} \leftarrow \frac{C_{i,j}+1}{2}$ to make the cells of C range over $[0,1]$ rather than $[-1,1]$.

(3) Zero out the diagonal of C: $C_{i,i} \leftarrow 0$ for every $i \in [\![1,n]\!]$.

(4) Normalise each row of C: for every $(i,j) \in [\![1,n]\!]^2$ do

$$C_{i,j} \leftarrow \frac{C_{i,j}}{\sum_{j'=1}^{n} C_{i,j'}}. \tag{A.9}$$

(5) Compute the d eigenvalues with largest magnitude $\lambda_1, \ldots, \lambda_d$ of C and associated unitary eigenvectors $X_1, \ldots, X_d$.

(6) Compute the number p of eigenvectors to keep.

i. Compute the normalised running totals $(S_i)_{1\le i\le d}$ of $(\lambda_i)_{1\le i\le d}$:

$$S_i := \frac{1}{\sum_{j=1}^{d} \lambda_j} \sum_{j=1}^{i} \lambda_j. \tag{A.10}$$

ii. Select enough eigenvectors to account for at least 99% of the observed variance:

$$p := \min\{i \in [\![1,d]\!] \mid S_i \ge 0.99\}. \tag{A.11}$$

(7) Create the matrix $M := \begin{pmatrix} X_1 & \cdots & X_p \end{pmatrix} \in \mathscr{M}_{n\times p}(\mathbf{R})$ and perform k-means on the lines of M.

So as to make the k-means step reproducible, the random seed was (arbitrarily) set to 32. As for the choice of eigenvectors (which are not unique), it has no influence since changing the i-th eigenvector amounts to changing the i-coordinate of all vectors to clusterise, which should not change the clustering.

A.3 The Transformer

This section briefly presents the Transformer, see [Vas+17] for an in-depth description of this architecture and HuggingFace[1] for an online documentation.

The Transformer is a class of neural networks relying solely on a *self-attention* function, which enables it to do computations in parallel – whereas most state-of-the-art neural networks only perform sequential computations because they read their input linearly and adapt their behaviour at time t according to their state at time $t-1$.

The Transformer consists of two stacks: one for encoding and one for decoding. Each stack consists of N (typically $N = 6$) layers equiped with self-attention mechanisms, and the decoder also performs attention on the output of the encoder stack. The architecture of the Transformer is depicted on figure A.1 on the next page.

The Transformer is suited for two kinds of language modelling. Causal Language Modelling (CLM), first, consists in analysing a sentence from the left to the right: the model tries to predict the first token, then the second token using the actual first token, and so on. In other words, CLM processeses inputs undirictionally. Masked Language Modelling, on the other hand (MLM), predicts each token given all of the others, whether they be located before or after the token to be predicted.

[1] `https://huggingface.co/`

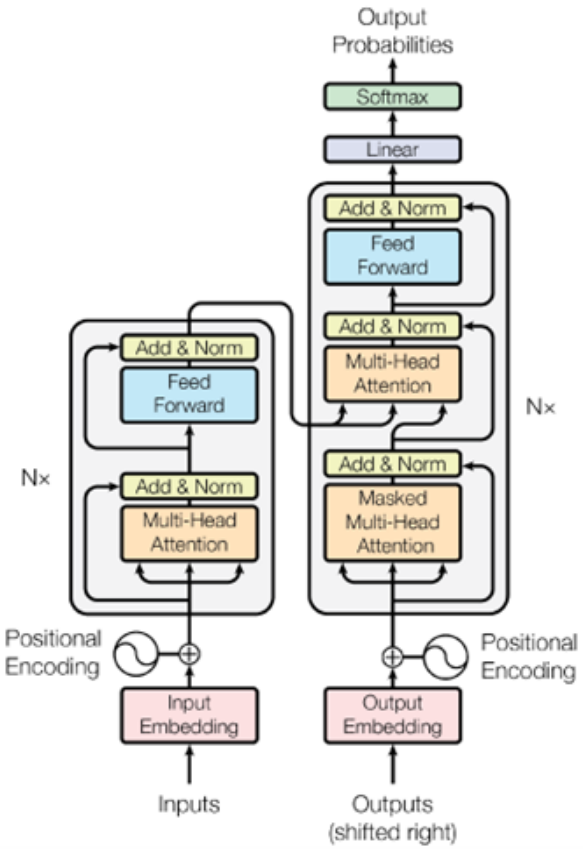


Fig. A.1. Architecture of the Transformer. Extract from [Vas+17].

This latter approach is bidirectional: it consider the input as a whole and only masks the token that has to be predicted. GPT-fr is a CLM model and CamemBERT is a MLM model – in fact, all BERT-based models are MLM models and all GPT-based models are CLM models.

As GPT-fr and CamemBERT were used in the project as CLM models, the rest of this section only applies to such models. Given a training vocabulary $(w_1, \dots, w_n)$ and a tokenised sequence $X = (x_1, \dots, x_t)$, the output layer of a CLM model returns a score matrix S of size $t \times n$ where $S_{i,j}$ is the score (in an arbitrary metric) assigns to the i-th token being w_j. The model then applies the softmax function to each line of S to transform the scores into a conditional probability distribution $t \mapsto \Pr_\theta(t \mid x_{<i})$. Finally, it computes, for each token x_i, a score $\mathrm{NLL}(x_i)$ called *negative log likelihood* which is defined as

$$\mathrm{NLL}(x_i) := -\log(\Pr_\theta(x_i \mid x_{<i})); \tag{A.12}$$

this score is obtained by applying the negative log loss function $\mathscr{L} : x \mapsto -\log(x)$ to the inner product $\left\langle S_i^\top \middle| (\delta_{(w_k = x_i)})_{1 \le k \le n}^\top \right\rangle$ where S_i is the i-th line of S. The final score output by the Transformer, *perplexity*, is defined as the mean of the negative log likelihood over all tokens of the input

$$\mathrm{Perplexity}(X) := \exp\left(-\frac{1}{t}\sum_{i=1}^{t} \mathrm{NLL}(x_i)\right). \tag{A.13}$$

Note that this approach is equivalent to taking the exponentiation of the cross-entropy between the data and model predictions.

APPENDIX B

Description of Amalric's and Moreno's experimental designs

This appendix presents the designs of the four experiments of which behavioural data were used to compare humans and the Transformer.

The MATHSEXPERTS paradigm. This experiment was published in [AD16] and aimed at comparing mathematicians and non-mathematics on high level mathematical stimuli (see appendix C.1 on page 49 for list of stimuli). Each subject was asked to assess every stimuli (presented in a randomised order) and to determine whether it was true, false or meaningless. One stimulus was read and then subjects were given a 4 s reflection and a 2 s response time, followed by a 7 s resting period. Subjects were asked always to give an answer, even when they were unsure or had no idea. Thirty subjects participated in this experiment, fifteen professional mathematicians (who held at least a Master's degree in mathematics) and fifteen controls (who also held a Master's degree but in humanities). There were ninety stimuli, equally divided into five categories (analysis, algebra, geometry, topology and non-maths) and three truth values (true, false and meaningless). Stimuli lengths and word frequencies across stimuli were matched.

The SIMPLEFACTS paradigm. This experiment was published in [AD19] and aimed replicating Amalric et al.'s MATHSEXPERTS experiment with much simpler stimuli (some even called upon rote memory, see appendix C.2 on page 52 for list of stimuli). Each subject was asked to assess every stimuli (presented in a randomised order) and to determine whether it was true, false or meaningless. One stimulus was read and the subjects were given a 2.5 s response time, followed by a 7 s resting period. Subjects were asked always to give an answer, even when they were unsure or had no idea. Fourteen subjects participated in this experiment, all professional mathematicians (holding at least a Master's degree in mathematics). There were seventy-two stimuli, equally divided into six categories (rote facts [well-known formulae like $(a+b)^2 = a^2 + b^2 + 2ab$], algebra, trigonometry, complex numbers, geometry and non-maths) and two truth values (true and false) – except for algebra for which there are five true and seven false stimuli (due to a misreading during the recording).

The COMBINATORIALOPERATIONS paradigm. This experiment was publised in [AD19]. It replicates the deisgn of Amalric et al.'s SIMPLEFACTS on the same subjects. Here, the aim was to determine whether the identified mathematical network responded to the logic structure of a sentence, e.g. with quantifiers and negations, independently of its mathematical or non-mathematical content (see appendix C.3 on page 54 for list of stimuli).There were ninety-six stimuli, equally

divided into mathematics and non-mathematics, four categories (declarative, quantifiers, negation and quantifiers plus negation) and two truth values (true and false). Stimuli lengths were matched across categories and stimuli grammar and syntax were matched within categories.

The MATHSLANGUAGE paradigm. This experiment has not been publised yet. It aims at at comparing teenagers and adults with high-level ASD with control teenagers and adults on various linguistic and mathematical tasks (see appendix C.4 on page 58 for list of stimuli). Each subject was asked to assess every stimuli (presented in a randomised order) and to determine whether it was true, false or meaningless. One stimulus was read every 8 s and subjects were asked to give their answer – even if they were unsure – after the end of the stimulus. At the time analyses were run, only one ASD (teenager) subject had been scanned, as well as sixteen control adults and fifteen control teenagers. Adults were only required to have graduted from high school but teenagers were explicitely required to be good at mathematics. There were three hundred and twenty-two stimuli, equally divided into eight categories (colorless green [grammatically correct but nonsensical sentences], word lists, theory of mind, contextualised, general knowledge, arithmetic principles, arithmetic facts and geometric facts) and two truth values (true and false) – except for geometric facts for which there were two additional stimuli. Stimuli lengths (number of words, syllables, characters, reading duration) and words frequencies across stimuli were matched.

Experiment	Categories of stimuli	Truth values	Number of subjects	Number of stimuli
MATHSEXPERTS	analysis algebra geometry topology non-maths	true false meaningless	15 mathematicians 15 controls	90 6 by condition
SIMPLEFACTS	rote facts algebra trigonometry complex numbers geometry non-maths	true false	14 mathematicans	72 6 by condition (5 and 7 for algebra)
COMBINATORIALOPERATIONS	declarative maths declarative non-maths quantifiers maths quantifiers non-maths negation maths negation non-maths quantifiers & negation maths quantifiers & negation non-maths	true false	14 mathematicians	96 6 by condition
MATHSLANGUAGE	colorless green word lists theory of mind contextualised general knowledge arithmetic principles arithmetic facts geometric facts	true false	16 adults (non-mathematicans) 15 teenagers (good at maths) 1 teenager with ASD	322 20 by condition (21 for geometry)

Table B.1. Summary of experimental designs.

APPENDIX C

List of Stimuli

C.1 MATHSEXPERTS

C.1.1 Analysis

True

- La série de Fourier d'une fonction f continue et $\mathscr{C}^1$ par morceaux converge simplement vers f.
- Toute fonction de $\mathbb{R}$ dans $\mathbb{R}$ localement polynomiale est polynomiale.
- La fonction $\frac{1}{\Gamma(z)}$ admet un prolongement analytique sur $\mathbb{C}$.
- Tout groupe topologique compact admet une mesure finie et invariante par le groupe des translations à gauche.
- Les fonctions test sont denses dans L^p pour tout réel $p \geq 1$.
- Une fonction $\mathscr{C}^\infty(\mathbb{R})$ dont toutes les dérivées sont positives est analytique.

False

- Les espaces L^p sont séparables.
- La transformée de Fourier est une isométrie de $\mathrm{L}^1(\mathbb{R}^n)$ dans lui-même.
- Le dual topologique de $\mathrm{L}^\infty(\mathbb{R})$ est $\mathrm{L}^1(\mathbb{R})$.
- Une inégalité sur des fonctions se vérifie également sur leurs primitives.
- Il existe une application continue de la boule unité dans elle même sans point fixe.
- La dérivée au sens des distributions d'un échelon est un échelon.

Meaningless

- Tout échelon de Dirac admet un développement limité à valeurs dans L^p.
- Il existe sur $\mathrm{L}^1(\mathbb{R}^n)$ une mesure localement polynomiale, séparable et analytique.
- En mesure finie, la série des zéros d'une fonction holomorphe est réflexive.
- Le dual topologique d'une série de Fourier est prolongeable analytiquement.
- La trace de la boule unité est divergente pour une dimension $p \neq 1$ et ∞.
- Tout espace polynomial compact est isométrique à un unique espace L^p.

C.1.2 Algebra

True

- Une matrice carrée à coefficients dans un anneau principal est inversible si et seulement si son déterminant est inversible.
- Si n est pair, toute sous-algèbre de $\mathscr{M}_n(\mathbb{C})$ de dimension inférieure à 4 admet un commutant non trivial.
- Les matrices carrées sur un corps équivalentes à une matrice nilpotente sont les matrices non inversibles.
- Il n'existe que cinq groupes de pavage direct du plan, à une conjugaison près.
- Il existe une sous-algèbre de dimension 13 dans les matrices complexes de taille 4×4.
- $\mathbb{Q}$ se plonge canoniquement dans tout corps de caractéristique nulle.

False

- Il existe un groupe d'ordre 169 dont le centre est réduit à un élément.
- Toute matrice à coefficients dans un anneau principal est équivalente à une matrice compagnon.
- Un groupe dont tous les sous-groupes propres sont abéliens est abélien.
- Dans l'algèbre $\mathscr{M}_n(\mathbb{C})$, si deux sous-algèbres commutent, la somme de leurs dimensions est inférieure à n^2.
- Toute matrice carrée est équivalente à une matrice de permutation.
- Il existe un groupe d'ordre infini admettant un nombre fini de sous-groupes.

Meaningless

- Tout anneau inversible carré admet un développement hexadécimal.
- Toute matrice de cardinal supérieur à 3 est factorielle.
- Le corps des fractions d'une algèbre immatricielle se plonge dans l'espace des projecteurs.
- Toute algèbre de dimension inférieure à 4 est combinaison linéaire de trois projecteurs.
- Il n'existe que cinq groupes nilpotents canoniquement additifs.
- Le corps $\mathbb{R}$[i] admet un commutant noethérien libre sur $\mathbb{Q}$.

C.1.3 Topology

True

- Une mesure finie sur un groupe compact invariante à gauche est bi-invariante.
- L'ensemble triadique de Cantor est de frontière égale à lui-même.
- Il existe des espaces non discrets dont toutes les composantes connexes sont réduites à un point.
- Une partie de $\mathbb{C}$, union d'ensembles connexes deux à deux non disjoints, est connexe.
- Toute partie bornée de $\mathbb{R}$ localement finie est finie.
- Un groupe topologique quotienté par sa composante neutre est totalement discontinu.

False

- Toute bijection continue entre deux espaces séparés est un homéomorphisme.
- Il existe une fonction continue de la sphère unité vers elle-même et sans point fixe.
- Tout compact convexe d'un espace euclidien est l'intersection d'une famille de boules fermées.
- Dans un espace topologique, tout sous-espace homéomorphe à un ouvert est lui-même ouvert.
- Tout graphe complet est plongeable dans la sphère unité de $\mathbb{R}^3$.
- Tout ensemble de nombres réels infini admet au moins un point d'accumulation.

Meaningless

- Tout morphisme croissant de l'ensemble triadique de Cantor est conjugué à un homéomorphisme de la boule unité.
- Toute mesure finie sur une algèbre de Hopf est localement modelée sur la mesure de Haar.
- La frontière d'un homéomorphisme est d'intérieur vide.
- Une partie de $\mathbb{C}$ est toujours invariante à gauche et continue à droite.
- Le graphe du complété d'un groupe compact est dense dans un ouvert partiellement connexe.
- Toute mesure indénombrable est l'intersection d'une famille de groupes compacts.

C.1.4 Geometry

True

- Tout champ de vecteur sur une sphère de dimension paire s'annule.
- Une hyperbole équilatère admet $\sqrt{2}$ pour excentricité.
- Pour une ellipse, la distance du centre à une directrice est égale au demi-grand axe sur l'excentricité.
- L'ensemble des points équidistants de deux droites non concourantes données de $\mathbb{R}^3$ euclidien est un paraboloïde hyperbolique.
- Un fibré vectoriel dont la base est contractile (par exemple une boule) est trivialisable.
- Le groupe orthogonal euclidien a exactement deux composantes connexes.

False

- Le projeté stéréographique de la sphère privée d'un point dans l'espace euclidien est borné.
- Une fonction holomorphe sur une surface de Riemann est constante.
- Toute surface compacte est difféomorphe à une surface algébrique.
- Par un point P d'une directrice d'une hyperbole passent deux tangentes à l'hyperbole.
- Le projeté orthogonal du foyer d'une parabole sur une de ses tangentes est sur la directrice.
- Tout champ de vecteurs de classe $\mathscr{C}^1$ sur le tore simple admet une singularité.

Meaningless

- Toute métrique riemannienne est conjuguée à la mesure de Haar.
- La projection stéréographique admet pour caractéristique d'Euler $\sqrt{2}$.
- L'ensemble des points équidistants de deux surfaces de Riemann est compatible avec un paraboloïde.
- Tout fibré holomorphe compact est un cas particulier de sphère.
- Toute variété algébrique localement contractile est contenue dans un hyperboloïde à deux nappes.
- Toute submersion localement ellipsoïdale est l'exponentielle d'une surface de Riemann.

C.1.5 Non-maths

True

- Toutes les cultures de la Méditerranée antique considéraient le taureau comme l'une des divinités.
- En Grèce antique, un citoyen incapable de payer ses dettes devenait esclave.
- Invention française, la TVA est un impôt direct sur la consommation.
- Le drapeau de la communauté esperanto est à dominante verte.
- Si l'on ne prend pas en compte le Vatican, Gibraltar est le plus petit pays du monde.

- L'idée des robots et des avatars était déjà présente dans la mythologie grecque.

False

- La construction du métro parisien est antérieure à celle du métro d'Istanbul.
- Les frontières de tous les pays d'Europe, sauf la Yougoslavie, ont été fixées au sortir de la seconde guerre mondiale.
- Le poète Aragon n'a jamais adhéré au parti communiste.
- La fin du concile de Trente coïncide avec la chute de l'empire romain d'Occident.
- Tous les membres du club des Cordeliers ont été guillotinés pendant la Terreur.
- Toutes les sociétés élèvent le marché au rang d'institution fondamentale et fondatrice.

Meaningless

- Le drapeau de la pomme de terre a été guillotiné à la fin du concile de Trente.
- Le marché institutionnalisé boit des avatars romains d'Occident.
- Tous les haricots verts endettés ont un bagage scientifique.
- La mythologie grecque est le plus petit alcool issu de la TVA.
- La plupart des taureaux robotisés n'ont jamais rencontré la Yougoslavie.
- Un poète est un impôt à dominante verte sur le métro.

C.2 SIMPLEFACTS

C.2.1 Rote facts

True

- $(a+b)^2 = a^2 + b^2 + 2ab$.
- $(a+b)(a-b) = a^2 - b^2$.
- $\cos(a+b) = \cos(a)\cos(b) - \sin(a)\sin(b)$.
- $\sin(a-b) = \sin(a)\cos(b) - \cos(a)\sin(b)$.
- $\cos(2a) = \cos^2(a) - \sin^2(a)$.
- $\cos^2(a) + \sin^2(a) = 1$.

False

- $(a-b)^2 = 2(a^2 + b^2)$.
- $(a+b)^3 = a^3 + b^3 + 1$.
- $\cos(a-b) = \sin(a)\cos(b) + \cos(a)\sin(b)$.
- $\sin(a+b) = \cos(a)\cos(b) + \sin(a)\sin(b)$.
- $\cos^2(a) + \sin^2(a) = \pi$.
- $\sin(2a) = 1 - \cos^2(a)$.

C.2.2 Algebra

True

- $(x+2)^2 = x^2 + 4 + 4x$.
- $(3-x)^2 = 9 + x^2 - 6x$.
- $(x-1)(x+1) = x^2 - 1$.
- $(z+1)^2 + (z-1)^2 = 2z^2 + 2$.
- $(3z+1)(3z-1) = 9z^2 - 1$.

False

- $(1+2z)^2 = 1+4z^2-4z$.
- $(x+4)^2-(x-4)^2 = x^2+16$.
- $(3x+1)^2 = 9x^2+1+3x$.
- $(2x-3)(2x+3) = 2x^2-3$.
- $(x+1)^2 = z^2+z+1$.
- $(z-2)^2 = z^2+4+4z$.
- $(z-3)(z+3) = z^2+9$.

C.2.3 Trigonometry

True

- $\sin\left(x+\frac{3\pi}{2}\right) = -\cos(x)$.
- $\cos(x+3\pi) = -\cos(x)$.
- $\cos(x+\pi) = -\cos(x)$.
- $\cos\left(x-\frac{\pi}{2}\right) = \sin(x)$.
- $\cos\left(x+\frac{\pi}{2}\right) = -\sin(x)$.
- $\sin(x-\pi) = -\sin(x)$.

False

- $\cos\left(x-\frac{3\pi}{2}\right) = \sin(x)$.
- $\cos(x+3\pi) = -\cos(x)$.
- $\sin(x-3\pi) = \cos(x)$.
- $\sin(x+\pi) = \sin(x)$.
- $\sin\left(x-\frac{\pi}{2}\right) = \cos(x)$.
- $\cos(x-\pi) = \cos(x)$.

C.2.4 Complex numbers

True

- $\Re\left(\mathrm{e}^{\frac{\mathrm{i}\pi}{2}}\right) = 0$.
- $\Im\left(\mathrm{e}^{\frac{\mathrm{i}\pi}{4}}\right) = \frac{\sqrt{2}}{2}$.
- $\Re\left(\mathrm{e}^{\frac{\mathrm{i}\pi}{4}}\right) = \Im\left(\mathrm{e}^{\frac{\mathrm{i}\pi}{4}}\right)$.
- $\sqrt{2}\cdot\mathrm{e}^{\frac{\mathrm{i}\pi}{4}} = 1+\mathrm{i}$.
- L'angle entre 1 et $1+\mathrm{i}$ est égal à $\frac{\pi}{4}$.
- L'angle entre 1 et i est égal à $\frac{\pi}{2}$.

False

- $\Re\left(\mathrm{e}^{\frac{3\mathrm{i}\pi}{4}}\right) = 1$.
- $\Im\left(\mathrm{e}^{\frac{\mathrm{i}\pi}{3}}\right) = \frac{1}{3}$.
- $\Re\left(\mathrm{e}^{\frac{\mathrm{i}\pi}{3}}\right) = \Im\left(\mathrm{e}^{\frac{2\mathrm{i}\pi}{3}}\right)$.
- $\Im\left(\mathrm{e}^{\frac{\mathrm{i}\pi}{4}}\right) = \sqrt{2}$.
- L'angle entre $1+\mathrm{i}$ et -1 est égal à $\frac{\pi}{4}$.
- L'angle entre $1-\mathrm{i}$ et i est égal à $\frac{3\pi}{2}$.

C.2.5 Geometry

True

- Un losange qui n'est pas un carré ne possède pas de cercle circonscrit.
- Dans un triangle équilatéral, l'ellipse tangente au milieu des côtés est un cercle.
- La rotation d'une droite autour d'un axe non-parallèle engendre un cône.
- Les points dont la somme des distances à deux points est constante forment une ellipse.
- Un triangle équilatéral se divise en deux triangles rectangles.
- La section d'un cône par un plan parallèle à l'axe du cône est une hyperbole.

False

- L'intersection d'une sphère et d'un plan est toujours un point.
- On peut paver un hexagone régulier par quatre triangles équilatéraux.
- Pour un cube, il existe exactement trois patrons différents.
- La révolution d'un cercle autour d'une droite engendre un hyperboloïde.
- La section d'un cylindre par un plan est toujours une droite.
- Les points équidistants d'un cercle et d'un point extérieur au cercle forment une droite.

C.2.6 Non-maths

True

- Le pointillisme est un style qui utilise des traits de pinceau visibles.
- Une acropole est une citadelle construite sur les hauteurs d'une cité grecque antique.
- Le blues est un genre musical dérivé des chants de travail des esclaves noirs.
- Le vibrato est une modulation périodique du son d'une note de musique.
- La pantomime est fondée sur l'attitude et le geste, sans recours à la parole.
- Un texte burlesque est caractérisé par l'emploi de termes familiers voire vulgaires.

False

- La Nouvelle Vague est un mouvement du cinéma des années deux-mille.
- L'Oulipo est un genre musical qui date de la Renaissance.
- Un roman épistolaire est un récit qu'une personne fait de sa propre existence.
- Le rock'n'roll est un genre musical caractérisé par un tempo lent.
- Un Harpagon est une personne caractérisée par sa bienveillance.
- Une vanité est un type de nature morte évoquant la beauté de la nature.

C.3 COMBINATORIALOPERATIONS

C.3.1 Declarative maths

True

- La fonction sinus est périodique.
- L'ensemble $\mathbb{R}$ est un corps.
- L'ensemble des entiers est dénombrable.
- Les relations d'équivalence sont symétriques.
- Les parties compactes sont fermées.
- Les cercles sont des ellipses.

False

- La fonction logarithme est définie sur tout $\mathbb{R}$.
- Le nombre d'or est un entier.
- La somme des $\frac{1}{n}$ est convergente.
- Les fonctions en escalier sont continues.
- Les rotations de l'espace sont commutatives.
- Les ensembles discrets sont connexes.

C.3.2 Declarative non-maths

True

- Le fruit du figuier est sucré.
- La montagne Pelée est un volcan.
- La grotte de Lascaux est préhistorique.
- Les épines des cactus sont douloureuses.
- Les bus londoniens sont rouges.
- Les babouins sont des primates.

False

- Le fruit du fraisier est récolté en hiver.
- La noix de coco est une épice.
- L'eau des lagons est jaune.
- Les fables de La Fontaine sont post-modernes.
- Les fruits de la passion sont salés.
- Les bois fossiles sont flexibles.

C.3.3 Quantifiers maths

True

- Certaines quadriques sont des cônes.
- Certains nombres réels sont des fractions.
- Certaines médianes sont des bissectrices.
- Certaines matrices sont diagonalisables.
- Certains polynômes sont quadratiques.
- Certains polygones sont convexes.

False

- Certains pentagones sont des rectangles.
- Certains plans sont de courbure positive.
- Certaines fonctions affines sont hyperboliques.
- Certains nombres rationnels sont transcendants.
- Certains hyperboloïdes sont bornés.
- Certains triangles sont des parallélogrammes.

C.3.4 Quantifiers non-maths

True

- Certains mammifères sont des cétacés.
- Certains récits antiques sont des épopées.
- Certains courants marins sont chauds.
- Certains romans sont autobiographiques.
- Certains volcans sont explosifs.
- Certains reptiles sont insectivores.

False

- Certains scarabées sont des crustacés.
- Certains rosiers sont des plantes bulbeuses.
- Certains éléphants sont des félins.
- Certaines îles tropicales sont glaciales.
- Certains écureuils sont carnivores.
- Certains chiens sont des rongeurs.

C.3.5 Negation maths

True

- La fonction cosinus n'est pas monotone.
- Le nombre π n'est pas rationnel.
- La fonction exponentielle n'est pas constante.
- Les fonctions quadratiques ne sont pas bijectives.
- Les carrés ne sont pas des coniques.
- Les hyperboloïdes ne sont pas connexes.

False

- Le nombre i n'est pas imaginaire pur.
- L'ensemble des translations n'est pas un groupe.
- Le groupe des translations n'est pas commutatif.
- Les tangentes ne sont pas des droites.
- Les suites convergentes ne sont pas bornées.
- Le bord d'un fermé n'est pas vide.

C.3.6 Negation non-maths

True

- Le bois mouillé n'est pas inflammable.
- La mer Baltique n'est pas chaude.
- La forêt amazonienne n'est pas désertique.
- Le lapin de garenne n'est pas agressif.
- Les grenouilles ne sont pas des insectes.
- Les algues vertes ne sont pas des plantes.

False

- La fleur d'oranger n'est pas parfumée.
- Le système républicain n'est pas une démocratie.
- L'atome d'uranium n'est pas radioactif.
- Les tigres ne sont pas carnivores.
- les cigognes ne sont pas des échassiers.
- Le vin rouge n'est pas alcoolisé.

C.3.7 Quantifiers & negation maths

True

- Certaines fonctions ne sont pas dérivables.
- Certains nombres entiers ne sont pas premiers.
- Certaines matrices ne sont pas inversibles.
- Certaines fonctions ne sont pas mesurables.
- Certaines ensembles infinis ne sont pas dénombrables.
- Certaines suites géométriques ne sont pas divergentes.

False

- Certains losanges ne sont pas des quadrilatères.
- Certaines fonctions dérivables ne sont pas continues.
- Certaines relations d'ordre ne sont pas transitives.
- Certaines boules ne sont pas convexes.
- Certaines séries de Fourier ne sont pas périodiques.
- Certaines fonctions bijectives ne sont pas injectives.

C.3.8 Quantifiers & negation non-maths

True

- Certaines plantes ne sont pas grasses.
- Certains légumes verts ne sont pas bons à manger.
- Certains romans ne sont pas publiés.
- Certains fruits ne sont pas comestibles.
- Certains contes philosophiques ne sont pas enfantins.
- Certaines plantes vertes ne sont pas grimpantes.

False

- Certains poissons pingouins ne sont pas des oiseaux.
- Certains poissons de rivière ne sont pas vertébrés.
- Certains nénuphars ne sont pas des plantes aquatiques.
- Certains ruminants ne sont pas herbivores.
- Certains fruits exotiques ne sont pas sucrés.
- Certains volcans d'Auvergne ne sont pas éteints.

C.4 MATHSLANGUAGE

C.4.1 Colourless green

- Le drapeau est une vieille analyse à froid de façon.
- La grimace est un arbre parallèle sans doigts.
- La sensation est un allumeur qui mange la douleur.
- Selon le bois, la plume est le choix des disques saisonniers.
- Pour charger, la décision pluvieuse est un inconvénient sans fauteuil.
- Pour chaque assoiffé, s'évader est une majorité.
- D'après les compagnies, la collectivité a la vache d'une terre.
- Pour les lions, la voile est une bouteille pensive.
- A chaque montagne, les importances sont de molles stalactites.
- En vérité, tous les gouvernements doivent retrouver à l'année.
- Aux champs bleus, les couvertures inversibles sont brutaux.
- Selon la voiture, l'amitié a été pliée sur cette application.
- Dans les chiens imperceptibles, le poivre est laid.
- Aux télégrammes, les pompiers sont crachés par la blessure fine.
- Les vieux sont complexes par les feuilles dans la pierre.
- L'artiste est l'album inévitable le plus fou de circonstances.
- L'intermédiaire de l'armée d'oxygène est doublé d'un programme véritable.
- La décoration se déduit grâce à l'espoir moléculaire.
- Un poulet est un cercle de chiffres internationales.
- Une mélancolie offre de piloter le pouvoir à ses savons chauds.
- Toute prière géométrique est attachée sur un mélange de vêtements.
- Le colis de l'ombre matérielle est toujours gratuit.
- Tous les cafés appris sont des moments bavardes.
- Un train auquel on prend la fête reste commun de soi.
- Une femme de traction est toujours un rond-point sale.
- On tue un astre en le grossissant avec des convictions.
- Toute base jeune peut être chère à une girafe.
- Le pompier d'ailes se brise lentement sur les chiens.
- Le lapin de mal va vite au bal du camion doux.
- Un bol de pain vaut son loup de jaune contraire.
- L'espoir d'oignon lisse et brun fait joli et gras.
- Un sol de vieux ciel chante bêtement un verre d'énergie.
- La solution de bureaux des heures est banale à finir.
- Les populations d'un meublé sont à belle gueule de la vitre.
- Le costume d'un melon est le frère de son gazon.
- Des bibliothèques de chaque cible ont le mauvais squelette.
- Les livres fondus se gavent de liberté aimablement.
- La pitié est débitée au roi du canard en petites montres.
- Les attributions abstraites se plient près de l'idéal.
- La réaction brosse par elle-même de longs zoos barbus.

C.4.2 Word list

- Est le une tandis comme sans te pour mais dans cependant.
- Un me pendant est que quand ainsi sans aux plus de.
- Qui le est un la selon pourtant se lorsque d'une avec.
- Le selon la est des néanmoins le avec pour très chaque.

- Sans pour est dont un lorsque de dedans celle ou puisque.
- Lui eux puis chaque pour si d'après pourtant d'un sitôt est.
- D'après où dans ils les d'une toute cependant et sans.
- La toutefois pour les est avec selon de une chacun plus.
- Chaque dès duquel de en tous eux mais ou ont celle.
- Aux les aussitôt sont un de encore une mon le auxquels.
- Selon alors auxquelles aux sur cette dans les est car ses.
- Aux les qui duquel les sont que selon ce le dans.
- Plus ensuite est de comme d'un les se d'après or celle toutefois.
- La est alors les par puisque de se près même je eux me.
- La de avant par même son comme donc en le tu vous si.
- La après est de des le d'un mon le de pendant il outre.
- On que toute sur est une afin le de tandis en pourquoi.
- Le de est tous auquel les d'ailleurs qui des chacun.
- Cependant à auquel elle la une de est un enfin soit.
- Être sont toute le une les quand qui sur afin très et.
- Sombre manteau titille cité chaud mégot trop chat.
- Devant décide lumière saule rouge lapin pouce.
- Douceur poisson cannelle grimace pôle mire mer.
- Poile choisis forêt vertes saison canapé doivent.
- Tourner décide journal antilope pluie effacer assis.
- Soif échapper concombre vite verre souterrain petit.
- Pigmées boeuf lac anticipe bouchon gazon chemise.
- Voilier tigres carafe pensif câlin boit ère blessé.
- Insigne doux colline biberon écrit blanc couper.
- Faux geler obligés semaine proposer doux carré fois.
- Période histoire visage pleine côté conditions jamais.
- Souvent été car double esprit dos jeune relations doute.
- Rouge doigts caractère occasion regard étoiles pièces.
- Membres intérieur venir gros faisant papier dents.
- Demain manger solution guerre couleur boire chambre bruit.
- Histoire jours frère société café armée saint états.
- Nombre voie assemblée parfois épaules peinture avions.
- Lettres grave économie création partie mesure mouvement.
- Compte production éléments demander courant amour fruits.
- Campagne langage phénomènes général français monsieur minutes.

C.4.3 Theory of mind

True

- Selon Darwin, l'homme est le cousin des singes.
- D'après Galilée, la terre a la forme d'une sphère.
- Pour La Fontaine, les fables sont de courtes leçons de vie.
- D'après Newton, les objets sont soumis à l'attraction terrestre.
- Selon les écrivains, l'Odyssée est un chef d'œuvre littéraire.
- D'après les climatologues, la Terre est en réchauffement.
- Selon les vulcanologues, l'Etna est toujours en activité.
- Selon Pasteur, la rage est une maladie mortelle.
- Selon Nelson Mandela, l'apartheid est un système raciste.
- Selon Greenpeace, la déforestation en Amazonie est une catastrophe.

- Pour les grecs, le mont Olympe est une montagne sacrée.
- Selon le pape, la mort est le début d'une vie éternelle.
- Selon les chrétiens fondamentalistes, Adam est notre ancêtre.
- Pour le président, l'Elysée est un lieu de tous les jours.
- Pour Ronaldo, le football est un art.
- Pour les créationnistes, Dieu est à l'origine de la vie.
- Pour Mowgli, la panthère noire est un animal affectueux.
- Pour Marie Antoinette, la monarchie est le meilleur régime politique.
- Pour les indiens d'Amérique, les armes à feu sont magiques.
- Selon Tarzan, nos parents sont des singes.

False

- Pour Mozart, la batterie donne le rythme de la musique.
- Pour Johnny Hallyday, chanter est un hobby.
- Selon Jules Verne, voyager au centre de la terre est impossible.
- Pour Hitler, les juifs sont des hommes comme les autres.
- Pour Donald Trump, les mexicains sont bienvenus aux États-Unis.
- D'après les musulmans, la viande de porc est bonne à manger.
- Selon Donald Trump, le changement climatique est inquiétant.
- D'après Astérix, les romains dominent tous les gaulois.
- Pour Napoléon, Austerlitz est le nom d'une gare.
- Pour Céline Dion, jouer de la musique est un hobby.
- Selon Zidane, la pétanque est un sport olympique.
- Selon Christophe Colomb, la terre est plate.
- Pour Nietzsche, Dieu est plus vivant que jamais.
- Selon Vladimir Poutine, la Russie est un tout petit pays.
- Pour Martin Luther King, la ségrégation raciale est légitime.
- Selon les physiciens, les atomes sont des particules gigantesques.
- Selon Gandhi, la violence résout tous les problèmes.
- D'après Mère Teresa, la haine est une émotion positive.
- D'après les humoristes, rire est mauvais pour la santé.
- Pour Einstein, la bombe atomique est un objet sans danger.

C.4.4 Contextualised

True

- A Pékin, la pollution est un problème majeur.
- Aux Pays Bas, les moulins fonctionnent avec le vent.
- En France, le homard est un aliment raffiné.
- A New York, les immeubles sont de hauts bâtiments modernes.
- En Allemagne, tous les enfants doivent aller à l'école.
- A San Francisco, les températures sont clémentes.
- Dans la mythologie grecque, le cheval de Troie est un piège.
- En Australie, le kangourou est un animal répandu.
- Dans l'Himalaya, les montagnes sont de hauts sommets enneigés.
- Aux Antilles, les restaurants servent souvent du poisson frais.
- A Londres, les autobus ont des étages.
- Selon la Bible, l'univers a été créé en quelques jours.
- Dans les îles volcaniques, le sable est noir.

- Dans la mythologie égyptienne, le chat incarne un dieu.
- En Chine, les baguettes servent de couverts à table.
- Au Japon, manger bruyamment est une forme de politesse.
- A Londres, les voitures roulent à gauche de la chaussée.
- Dans les contes pour enfants, les animaux parlent couramment.
- En Irlande, avoir les cheveux roux est assez courant.
- Dans la mer, les mammifères ont des nageoires et un évent.

False

- En Inde, les vaches sont des animaux de boucherie.
- En Amazonie, les araignées sont des animaux sans danger.
- Au Vatican, le pape est une personne comme les autres.
- Au Burundi, les médecins sont remboursés par la sécurité sociale.
- Sur la planète Mars, la lune entraîne des marées.
- En Equateur, la durée du jour est courte en hiver.
- A Paris, la Voie Lactée est souvent visible la nuit.
- Dans le métro parisien, les trains transportent des céréales.
- En Corée du Nord, les gens sont libres de s'exprimer.
- Pendant les cours, les étudiants peuvent parler entre eux.
- Aux Etats Unis, les études universitaires sont gratuites.
- En Allemagne, les automobilistes roulent à gauche.
- En Afrique du Sud, le diamant est une pierre sans valeur.
- Au Pakistan, tous les enfants sont scolarisés.
- En Bretagne, la température atteint souvent des sommets.
- Dans les plaines des Pays Bas, marcher est une activité épuisante.
- Sur Mars, l'atmosphère est composée principalement d'oxygène.
- Au Kenya, les lacs des plaines gèlent régulièrement.
- En Inde, le dinar est la monnaie courante.
- Au Japon, la mortalité infantile est importante.

C.4.5 General knowledge

True

- Le chameau est un grand mammifère à bosse.
- La tulipe est une plante multicolore à bulbe.
- Le caméléon est un animal qui change de couleur.
- L'eau gèle à la température de zéro degrés centigrades.
- La baleine est un énorme mammifère qui vit en mer.
- Le canari est un petit oiseau chanteur jaune.
- Le soleil se lève à l'est tous les matins.
- Les perles sont fabriquées par les huîtres dans la mer.
- Les flocons de neige ont des formes très variées.
- Le noyau de l'atome d'hydrogène est composé d'un seul proton.
- La photosynthèse se produit grâce à l'énergie lumineuse.
- Une encyclopédie permet de diffuser le savoir à de nombreux lecteurs.
- La tragédie classique est fondée sur une unité de lieu.
- Les chenilles sont des insectes qui se transforment en papillon.
- Les arbres perdent leurs feuilles au moment de l'automne.
- La rose est la fleur d'une plante épineuse.

- Février est le mois de l'année le plus court.
- Le calendrier grégorien est utilisé dans la plupart des pays occidentaux.
- Le corps humain est formé par des milliards de cellules.
- La mitose est le processus par lequel les cellules se divisent.

False

- Les papillons ont des ailes minuscules et transparentes.
- Le petit de la vache s'appelle un poulain.
- Les algues vertes sont des plantes aquatiques.
- Les poissons de rivière ne sont pas des vertébrés.
- La primevère est une plante vénéneuse et épineuse.
- Jupiter est la plus petite planète du système solaire.
- La terre tourne autour de la planète Venus.
- Les dinosaures sont des mammifères qui ont disparu.
- L'écureuil est l'animal terrestre le plus grand d'Amérique.
- Un pamphlet est un ensemble de textes mélancoliques.
- Le veau est le petit de la chèvre.
- La vapeur d'eau se condense uniquement sur les arbres.
- La pleine lune a lieu plusieurs fois chaque mois.
- L'hiver est la saison la plus chaude de l'année.
- La force de gravitation protège la terre des radiations solaires.
- La planète Terre possède plusieurs satellites naturels.
- Les phrases françaises se lisent de la droite vers la gauche.
- Le Vietnam est appelé le pays du soleil levant.
- Les aurores boréales ont lieu près de l'équateur.
- La Terre tourne sur elle-même en un mois.

C.4.6 Arithmetic principles

True

- Un nombre multiplié par zéro est égal à zéro.
- Un nombre multiplié par un est égal à lui-même.
- Ajouter zéro ne change pas un nombre.
- Un nombre positif est plus grand qu'un nombre négatif.
- La soustraction est le contraire de l'addition.
- Un nombre positif fois un nombre négatif donne un nombre négatif.
- Un nombre négatif fois un nombre négatif donne un nombre positif.
- Un nombre divisé par un est égal à lui-même.
- La somme de plusieurs nombres pairs est un nombre pair.
- L'ordre d'addition des nombres est sans importance.
- L'ordre de multiplication des nombres est sans importance.
- La moitié d'un nombre est plus grande que le tiers.
- Tous les nombres entiers sont divisibles par un.
- Un nombre entier est divisible par un et par lui-même.
- L'addition de nombres positifs est supérieure ou égale à ces nombres.
- La multiplication est le contraire de la division.
- Pour additionner des fractions, les dénominateurs doivent être égaux.
- Une fraction est la division d'un nombre entier par un autre.
- Un nombre négatif élevé au carré devient positif.

- Un nombre élevé à la puissance un est égal à lui-même.

False

- L'ordre de soustraction des nombres est sans importance.
- Tous les entiers sont des nombres pairs.
- Le carré d'un nombre impair est un nombre pair.
- Soustraire un à un nombre le laisse inchangé.
- Un nombre multiplié par zéro est égal à lui-même.
- Un nombre multiplié par un est égal à zéro.
- Le résultat d'une addition est toujours positif.
- Le résultat d'une soustraction est toujours négatif.
- Le résultat d'une division est toujours un nombre entier.
- Une somme de nombres négatifs peut être positive.
- La différence de nombres négatifs est toujours positive.
- Le produit de nombres négatifs est toujours négatif.
- Tous les nombres négatifs sont des nombres impairs.
- Un nombre auquel on soustrait zéro reste égal à lui-même.
- Une somme de fractions est toujours un nombre entier.
- On double un nombre en le multipliant par lui-même.
- Tout nombre réel peut être écrit comme une fraction.
- Le carré d'un nombre négatif est négatif.
- L'ordre selon lequel on divise des nombres est sans importance.
- Un nombre non nul multiplié par lui-même donne toujours un.

C.4.7 Arithmetic facts

True

- Le résultat de vingt-sept moins trois est le nombre vingt-quatre.
- Le résultat de soixante-dix-neuf moins cinq est soixante-quatorze.
- La somme de cinquante-cinq et deux donne cinquante-sept.
- Le produit de trois par sept a comme résultat vingt et un.
- Le nombre vingt-sept divisé par trois donne comme résultat neuf.
- Le nombre trois multiplié par douze fait trente-six.
- Deux fois vingt-quatre a comme résultat quarante-huit.
- La racine carrée de trente-six est plus grand que cinq.
- Quarante-cinq est le résultat de quinze fois trois.
- Le résultat de la multiplication de six par dix est soixante.
- Soustraire trois du nombre vingt donne dix-sept.
- Le produit de vingt et un par trois donne comme résultat soixante-trois.
- Un quart du nombre cent vaut vingt-cinq.
- L'addition de dix-huit et trois fait vingt et un.
- Un tiers de quarante-cinq a pour valeur quinze.
- La division de seize par quatre est égale à quatre.
- Cinquante pour cent de trente a comme valeur quinze.
- Le double du nombre neuf est le nombre dix-huit.
- Quarante-deux moins un donne le nombre quarante et un.
- Le nombre douze multiplié par dix fait cent vingt.

False

- Quarante divisé par deux fait quatre-vingt-dix-huit.
- Vingt-deux est le carré du chiffre dix-sept.
- Dix-sept est la racine carrée de cent quatre-vingt-dix.
- La division de trente-trois par deux donne le chiffre douze.
- Le triple de quinze est plus grand que deux cent soixante.
- Trente-sept est la moitié du nombre cinquante-quatre.
- La multiplication de quatre par trois donne dix.
- Le résultat de vingt-trois fois deux est le chiffre neuf.
- Cinq plus seize est plus grand que le nombre vingt-huit.
- Trente-cinq moins six fait cent soixante-dix-neuf.
- Deux fois soixante-dix font trois cent cinquante-quatre.
- Vingt-sept moins deux moins quatre fait dix-huit.
- Quatre-vingt-cinq est plus petit que quarante-quatre.
- Le nombre soixante-quatorze est plus grand que cent.
- Dix mille est le produit de cent par cent soixante.
- Seize est le résultat de deux multiplié par dix-huit.
- La moitié de deux cents vaut trois cent cinquante-cinq.
- Le résultat de la division de quatorze par sept est trois.
- Le nombre dix-neuf est compris entre un et dix-sept.
- Sept fois sept est plus grand que le chiffre cinquante-quatre.

C.4.8 Geometric facts

True

- Un carré est une figure dont tous les côtés sont égaux.
- Les côtés d'un triangle équilatéral ont la même longueur.
- Tous les angles d'un carré sont des angles droits.
- Les angles d'un polygone régulier sont de la même taille.
- Les lignes parallèles ne se croisent pas.
- Un point peut être traversé par un nombre infini de lignes.
- Les losanges et les carrés ont des côtés égaux.
- Un angle droit est plus grand qu'un angle aigu.
- Tous les points d'un cercle sont à égale distance du centre.
- Un triangle équilatéral se divise en deux triangles rectangles.
- Le périmètre d'un carré est la somme des longueurs des côtés.
- La surface d'un rectangle est le produit de ses côtés.
- La diagonale d'un carré est plus grande que son côté.
- Tout polygone peut entrer dans un cercle.
- Le diamètre d'un cercle est le double de son rayon.
- Un cercle a un nombre infini d'axes de symétrie.
- Les diagonales d'un losange sont ses axes de symétrie.
- Les faces d'un cube sont des carrés.
- Un carré est un rectangle aux côtés égaux.
- Un carré possède des axes de symétrie.
- Un cube a six faces carrées.

False

- Un cercle possède de nombreux côtés rectilignes.
- Un hexagone est une figure avec des angles droits.
- Un rectangle coupé par une ligne droite donne des carrés égaux.
- Un triangle isocèle a plusieurs axes de symétrie.
- Plus un triangle est grand, plus ses angles sont grands.
- Des lignes sont perpendiculaires quand elles forment un angle aigu.
- Les points d'un carré sont à égale distance de son centre.
- Le périmètre d'un cercle est le triple de son rayon.
- Des figures de même périmètre ont la même surface.
- Un triangle est une figure dont tous les côtés sont égaux.
- Un rectangle a un seul axe de symétrie.
- Un rectangle ne comprend que des angles aigus.
- Les côtés d'un losange peuvent être de différentes longueurs.
- Un triangle équilatéral est formé uniquement d'angles droits.
- Un triangle peut avoir plusieurs angles droits.
- Un triangle a plus d'angles qu'un carré.
- Des lignes perpendiculaires forment un angle aigu.
- Un losange peut se diviser en carrés.
- Un carré peut se diviser en triangles équilatéraux.
- Un angle droit est plus grand qu'un angle obtus.
- Un triangle peut avoir deux angles droits.

APPENDIX D

Supplementary Material

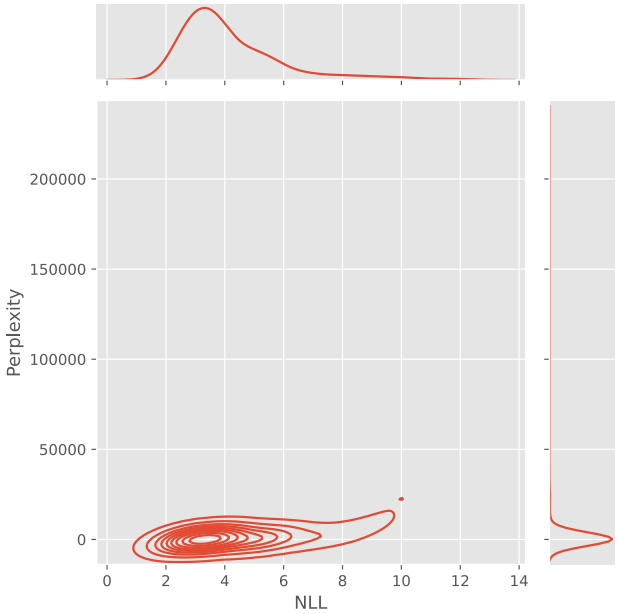


Fig. D.1. Comparison of the distribution of GPT-fr's mean negative log likelihood and perplexity on the stimuli.

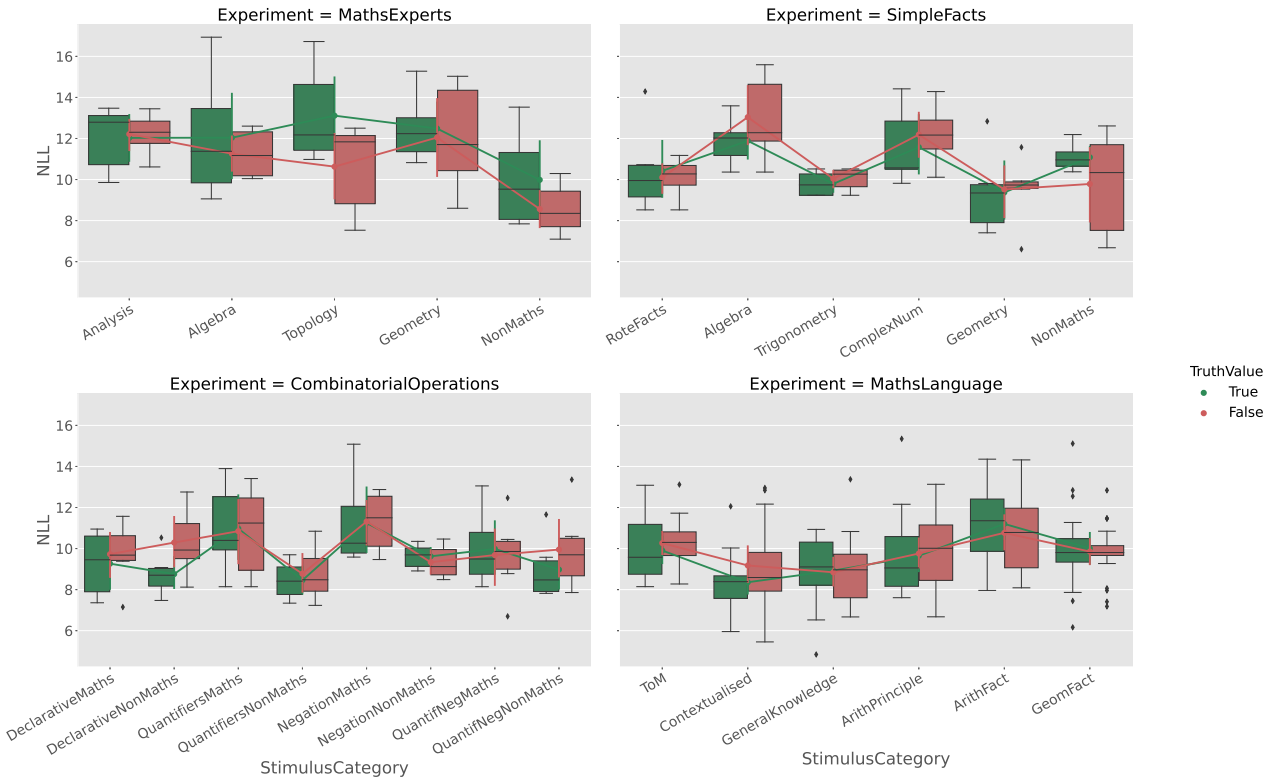


Fig. D.2. GPT-fr's maximum negative log likelihood by *meaningful* stimulus depending on stimulus category and truth value.

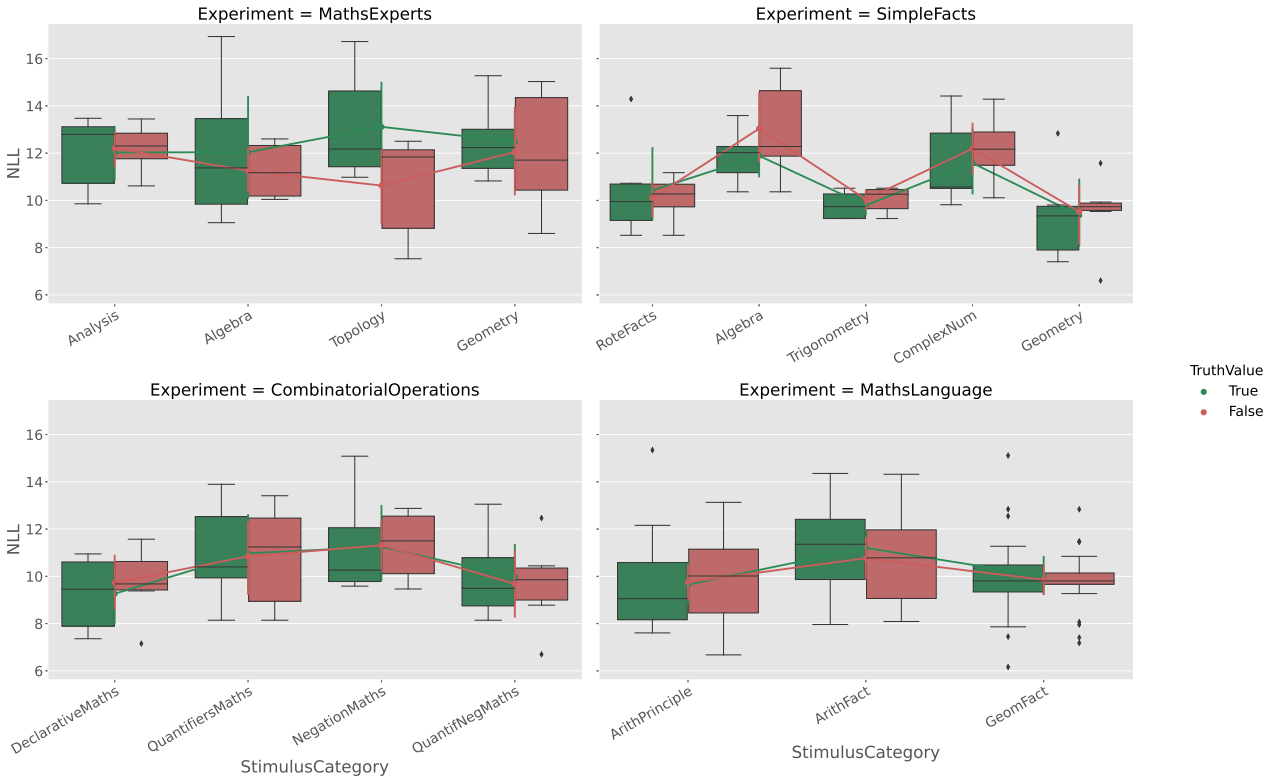


Fig. D.3. GPT-fr's maximum negative log likelihood by *meaningful mathematical* stimulus depending on stimulus category and truth value.

Region	x	y	z
L IPS	−53	−43	57
R IPS	50	−36	56
L IT	−53	−57	−16
R IT	52	−52	−19
L MFG/BA46	−44	31	27
R MFG/BA9	−47	7	31
L SFS	−24	8	64
R SFS	32	5	56
R MFG/BA46	50	47	16
R MFG/BA9-BA10	50	10	21
SMA	−2	23	51
BA10	−20	47	−16
L Cereb. 6th lobule	−29	−66	−29
R Cereb. 6th lobule	39	−73	−26

Table D.1. Coordinates of the mathematical ROIs defined by Amalric et al. [AD16].

List of Tables

List of Figures

Bibliography

[AD16] Marie Amalric and Stanislas Dehaene. “Origins of the brain networks for advanced mathematics in expert mathematicians”. en. In: *Proceedings of the National Academy of Sciences* 113.18 (May 2016), pp. 4909–4917. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1603205113. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1603205113.

[AD18] Marie Amalric and Stanislas Dehaene. “Cortical circuits for mathematical knowledge: evidence for a major subdivision within the brain’s semantic networks”. en. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 373.1740 (Feb. 2018), p. 20160515. ISSN: 0962-8436, 1471-2970. DOI: 10.1098/rstb.2016.0515. URL: https://royalsocietypublishing.org/doi/10.1098/rstb.2016.0515.

[AD19] Marie Amalric and Stanislas Dehaene. “A distinct cortical network for mathematical knowledge in the human brain”. en. In: *NeuroImage* 189 (Apr. 2019), pp. 19–31. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2019.01.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S1053811919300011.

[CGK21] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. “Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects”. en. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 3635–3644. DOI: 10.18653/v1/2021.findings-emnlp.308. URL: https://aclanthology.org/2021.findings-emnlp.308.

[Cha21] François Charton. “Linear algebra with transformers”. en. In: (Dec. 2021). DOI: 10.48550/arXiv.2112.01898. URL: https://arxiv.org/abs/2112.01898v1.

[CHL21] François Charton, Amaury Hayat, and Guillaume Lample. *Learning advanced mathematical computations from examples*. arXiv:2006.06462 [cs]. Mar. 2021. DOI: 10.48550/arXiv.2006.06462. URL: http://arxiv.org/abs/2006.06462.

[CK22] Charlotte Caucheteux and Jean-Rémi King. “Brains and algorithms partially converge in natural language processing”. en. In: *Communications Biology* 5.1 (Feb. 2022). Number: 1 Publisher: Nature Publishing Group, pp. 1–10. ISSN: 2399-3642. DOI: 10.1038/s42003-022-03036-1. URL: https://www.nature.com/articles/s42003-022-03036-1.

[Dai+16] Amy L. Daitch et al. “Mapping human temporal and parietal neuronal population activity and functional coupling during mathematical cognition”. en. In: *Proceedings of the National Academy of Sciences* 113.46 (Nov. 2016). ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1608434113. URL: https://pnas.org/doi/full/10.1073/pnas.1608434113.

[Das+13] Mohammad Dastjerdi et al. “Numerical processing in the human parietal cortex during experimental and natural conditions”. en. In: *Nature Communications* 4.1 (Oct. 2013). Number: 1 Publisher: Nature Publishing Group, p. 2528. ISSN: 2041-1723. DOI: 10.1038/ncomms3528. URL: https://www.nature.com/articles/ncomms3528.

[dAs+22] Stéphane d’Ascoli et al. *Deep Symbolic Regression for Recurrent Sequences*. arXiv:2201.04600 [cs]. Jan. 2022. DOI: 10.48550/arXiv.2201.04600. URL: http://arxiv.org/abs/2201.04600.

[Dav+21] Alex Davies et al. “Advancing mathematics by guiding human intuition with AI”. en. In: *Nature* 600.7887 (Dec. 2021). Number: 7887 Publisher: Nature Publishing Group, pp. 70–74. ISSN: 1476-4687. DOI: 10.1038/s41586-021-04086-x. URL: https://www.nature.com/articles/s41586-021-04086-x.

[Dee+90] Scott Deerwester et al. “Indexing by latent semantic analysis”. en. In: *Journal of the American Society for Information Science* 41.6 (1990), pp. 391–407. ISSN: 1097-4571. DOI: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28199009%2941%3A6%3C391%3A%3AAID-ASI1%3E3.0.CO%3B2-9.

[Deh03] Stanislas Dehaene. “The neural basis of the Weber–Fechner law: a logarithmic mental number line”. en. In: *Trends in Cognitive Sciences* 7.4 (Apr. 2003), pp. 145–147. ISSN: 1364-6613. DOI: 10.1016/S1364-6613(03)00055-X. URL: https://www.sciencedirect.com/science/article/pii/S136466130300055X.

[Ege16] E. Eger. “Neuronal foundations of human numerical representations”. en. In: *Progress in Brain Research*. Vol. 227. Elsevier, 2016, pp. 1–27. ISBN: 978-0-444-63698-0. DOI: 10.1016/bs.pbr.2016.04.015. URL: https://linkinghub.elsevier.com/retrieve/pii/S0079612316300413.

[Hal+22] John T. Hale et al. “Neurocomputational Models of Language Processing”. en. In: *Annual Review of Linguistics* 8.1 (Jan. 2022), pp. 427–446. ISSN: 2333-9683, 2333-9691. DOI: 10.1146/annurev-linguistics-051421-020803. URL: https://www.annualreviews.org/doi/10.1146/annurev-linguistics-051421-020803.

[Hut+16] Alexander G. Huth et al. “Natural speech reveals the semantic maps that tile human cerebral cortex”. en. In: *Nature* 532.7600 (Apr. 2016), pp. 453–458. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature17637. URL: http://www.nature.com/articles/nature17637.

[Kam+22] Pierre-Alexandre Kamienny et al. *End-to-end symbolic regression with transformers*. arXiv:2204.10532 [cs]. Apr. 2022. DOI: 10.48550/arXiv.2204.10532. URL: http://arxiv.org/abs/2204.10532.

[KKF22] Kohitij Kar, Simon Kornblith, and Evelina Fedorenko. *Interpretability of artificial neural network models in artificial Intelligence vs. neuroscience*. Number: arXiv:2206.03951 arXiv:2206.03951 [q-bio]. June 2022. DOI: 10.48550/arXiv.2206.03951. URL: http://arxiv.org/abs/2206.03951.

[LC19] Guillaume Lample and François Charton. “Deep Learning for Symbolic Mathematics”. en. In: (Dec. 2019). DOI: 10.48550/arXiv.1912.01412. URL: https://arxiv.org/abs/1912.01412v1.

[Ler+11] Yulia Lerner et al. “Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story”. en. In: *Journal of Neuroscience* 31.8 (Feb. 2011). Publisher: Society for Neuroscience Section: Articles, pp. 2906–2915. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.3684-10.2011. URL: https://www.jneurosci.org/content/31/8/2906.

[Lux07] Ulrike von Luxburg. “A Tutorial on Spectral Clustering”. en. In: *arXiv:0711.0189 [cs]* (Nov. 2007). arXiv: 0711.0189. URL: http://arxiv.org/abs/0711.0189.

[Mar+12] Masaki Maruyama et al. “The cortical representation of simple mathematical expressions”. en. In: *NeuroImage* 61.4 (July 2012), pp. 1444–1460. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2012.04.020. URL: https://www.sciencedirect.com/science/article/pii/S1053811912004016.

[Mar+20] Louis Martin et al. “CamemBERT: a Tasty French Language Model”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020). arXiv: 1911.03894, pp. 7203–7219. DOI: 10.18653/v1/2020.acl-main.645. URL: http://arxiv.org/abs/1911.03894.

[Mil+22] Juliette Millet et al. *Toward a realistic model of speech processing in the brain with self-supervised learning*. June 3, 2022. DOI: 10.48550/arXiv.2206.01685. arXiv: 2206.01685[cs,q-bio]. URL: http://arxiv.org/abs/2206.01685.

[MPO12] Martin M. Monti, Lawrence M. Parsons, and Daniel N. Osherson. “Thought Beyond Language: Neural Dissociation of Algebra and Natural Language”. en. In: *Psychological Science* 23.8 (Aug. 2012), pp. 914–922. ISSN: 0956-7976, 1467-9280. DOI: 10.1177/0956797612437427. URL: http://journals.sagepub.com/doi/10.1177/0956797612437427.

[MYZ19] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. “Linguistic Regularities in Continuous Space Word Representations”. en. In: July 2019. URL: https://openreview.net/forum?id=BybT3m-ubH (visited on 06/29/2022).

[Nie16] Frank Nielsen. “Hierarchical Clustering”. In: Feb. 2016, pp. 195–211. ISBN: 978-3-319-21902-8. DOI: 10.1007/978-3-319-21903-5_8.

[Par+12] Joonkoo Park et al. “Neural Dissociation of Number from Letter Recognition and Its Relationship to Parietal Numerical Processing”. In: *Journal of Cognitive Neuroscience* 24.1 (Jan. 2012), pp. 39–50. ISSN: 0898-929X. DOI: 10.1162/jocn_a_00085. URL: https://doi.org/10.1162/jocn_a_00085.

[Pen+21] Shuai Peng et al. “MathBERT: A Pre-Trained Model for Mathematical Formula Understanding”. en. In: *arXiv:2105.00377 [cs]* (May 2021). arXiv: 2105.00377. URL: http://arxiv.org/abs/2105.00377.

[Per+18] Francisco Pereira et al. “Toward a universal decoder of linguistic meaning from brain activation”. en. In: *Nature Communications* 9.1 (Dec. 2018), p. 963. ISSN: 2041-1723. DOI: 10.1038/s41467-018-03068-4. URL: http://www.nature.com/articles/s41467-018-03068-4.

[Pol+22] Stanislas Polu et al. *Formal Mathematics Statement Curriculum Learning*. Number: arXiv:2202.01344 arXiv:2202.01344 [cs]. Feb. 2022. DOI: 10.48550/arXiv.2202.01344. URL: http://arxiv.org/abs/2202.01344.

[PS20] Stanislas Polu and Ilya Sutskever. “Generative Language Modeling for Automated Theorem Proving”. en. In: *arXiv:2009.03393 [cs, stat]* (Sept. 2020). arXiv: 2009.03393. URL: http://arxiv.org/abs/2009.03393.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher Manning. “Glove: Global Vectors for Word Representation”. en. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: http://aclweb.org/anthology/D14-1162.

[Raa+21] Gal Raayoni et al. “Generating conjectures on fundamental constants with the Ramanujan Machine”. en. In: *Nature* 590.7844 (Feb. 2021). Number: 7844 Publisher: Nature Publishing Group, pp. 67–73. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03229-4. URL: https://www.nature.com/articles/s41586-021-03229-4.

[SC21] Antoine Simoulin and Benoit Crabbé. “Un modèle Transformer Génératif Pré-entrainé pour le _____ français”. In: *Traitement Automatique des Langues Naturelles*. Ed. by Pascal Denis et al. Lille, France: ATALA, 2021, pp. 246–255. URL: https://hal.archives-ouvertes.fr/hal-03265900.

[Sch+21] Martin Schrimpf et al. “The neural architecture of language: Integrative modeling converges on predictive processing”. en. In: *Proceedings of the National Academy of Sciences* 118.45 (Nov. 2021), e2105646118. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2105646118. URL: https://pnas.org/doi/full/10.1073/pnas.2105646118.

[Shu+13] Jennifer Shum et al. “A Brain Area for Visual Numerals”. en. In: *Journal of Neuroscience* 33.16 (Apr. 2013). Publisher: Society for Neuroscience Section: Articles, pp. 6709–6715. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.4558-12.2013. URL: https://www.jneurosci.org/content/33/16/6709.

[SL13] Nathaniel J. Smith and Roger Levy. “The effect of word predictability on reading time is logarithmic”. en. In: *Cognition* 128.3 (Sept. 2013), pp. 302–319. ISSN: 0010-0277. DOI: 10.1016/j.cognition.2013.02.013. URL: https://www.sciencedirect.com/science/article/pii/S0010027713000413.

[Vas+17] Ashish Vaswani et al. “Attention Is All You Need”. en. In: *arXiv:1706.03762 [cs]* (Dec. 2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[WEG87] Svante Wold, Kim Esbensen, and Paul Geladi. “Principal component analysis”. en. In: *Chemometrics and Intelligent Laboratory Systems*. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists 2.1 (Aug. 1987), pp. 37–52. ISSN: 0169-7439. DOI: 10.1016/0169-7439(87)80084-9. URL: https://www.sciencedirect.com/science/article/pii/0169743987800849.

[Wu+22] Yuhuai Wu et al. *Memorizing Transformers*. Number: arXiv:2203.08913 arXiv:2203.08913 [cs]. Mar. 2022. DOI: 10.48550/arXiv.2203.08913. URL: http://arxiv.org/abs/2203.08913.