CAMS - EHESS
Séminaire Systèmes complexes en sciences sociales
2 Juin 2026 – Paris, France

Explicabilité formelle et interprétabilité théorique des modèles neuronaux

Juan Luis (Gianni) Gastaldi

ETH zürich

www.giannigastaldi.com

Outline

Intro: IA et sciences humaines

Explicabilité formelle

Interprétabilité théorique

Conclusion

Outline

Behavior Revealed in Mobile Phone Usage Predicts Credit Repayment

Daniel Björkegren and Darrell Grissen

On Inferring User Socioeconomic Status with Mobility Records

Zheng Wang*†, Mingrui Liu*, Cheng Long*¶, Qianru Zhang‡, Jiangneng Li*, Chunyan Miao*§
*School of Computer Science and Engineering, Nanyang Technological University, Singapore,
†Huawei Singapore Research Center, Singapore,
‡Department of Computer Science, The University of Hong Kong, Hong Kong SAR
§China-Singapore International Joint Research Institute (CSIJRI), China

Towards Deep Learning Models for Psychological State Prediction using Smartphone Data: Challenges and Opportunities

Gatis Mikelsons
University of Oxford and
The Alan Turing Institute
gatis.mikelsons@jesus.ox.ac.uk

Matthew Smith
University of Warwick and
The Alan Turing Institute
m.d.smith@warwick.ac.uk

IA et données sociales

Behavior Revealed in Mobile Phone Usage Predicts Credit Repayment

Daniel Björkegren and Darrell Grissen

On Inferring User Socioeconomic Status with Mobility Records

Zheng Wang*†, Mingrui Liu*, Cheng Long*¶, Qianru Zhang‡, Jiangneng Li*, Chunyan Miao*§
*School of Computer Science and Engineering, Nanyang Technological University, Singapore,
†Huawei Singapore Research Center, Singapore,
‡Department of Computer Science, The University of Hong Kong, Hong Kong SAR
§China-Singapore International Joint Research Institute (CSIJRI), China

Towards Deep Learning Models for Psychological State Prediction using Smartphone Data: Challenges and Opportunities

Gatis Mikelsons
University of Oxford and
The Alan Turing Institute
gatis.mikelsons@jesus.ox.ac.uk

Matthew Smith
University of Warwick and
The Alan Turing Institute
m.d.smith@warwick.ac.uk

The Potential and Challenges of Evaluating Attitudes, Opinions, and Values in Large Language Models

Bolei Ma* [LMU, MCML] **Xinpeng Wang*** [LMU, MCML] **Tiancheng Hu** [Cambridge] **Anna-Carolina Haensch** [LMU, Maryland]
Michael A. Hedderich [LMU, MCML] **Barbara Plank** [LMU, MCML, ITU] **Frauke Kreuter** [LMU, MCML, Maryland]
[LMU] LMU Munich [MCML] Munich Center for Machine Learning [Cambridge] University of Cambridge
[Maryland] University of Maryland, College Park [ITU] ITU Copenhagen

Multilingual Political Views of Large Language Models: Identification and Steering

Daniil Gurgurov[1,5] **Katharina Trinley**[1] **Ivan Vykopal**[3,4]
Josef van Genabith[1,5] **Simon Ostermann**[1,5] **Roberto Zamparelli**[2]

[1]Saarland University [2]University of Trento
[3]Brno University of Technology [4]Kempelen Institute of Intelligent Technologies
[5]German Research Center for AI (DFKI)

The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation

Jochen Hartmann[a,1,2⊠], **Jasper Schwenzow**[b,1], **and Maximilian Witte**[b,1]

[a]Technical University of Munich, TUM School of Management, Arcisstr. 21, 80333 Munich, Germany
[b]University of Hamburg, Hamburg Business School, Moorweidenstrasse 18, 20148 Hamburg, Germany
[1]All authors contributed equally to this work.

Exemple: life2vec

nature computational science

Article

https://doi.org/10.1038/s43588-023-00573-5

Using sequences of life-events to predict human lives

Received: 6 June 2023

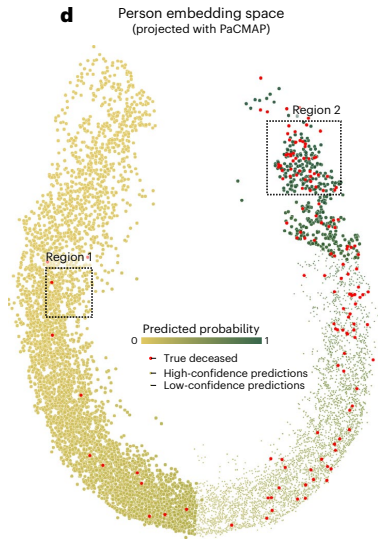
Accepted: 15 November 2023

Published online: 18 December 2023

Check for updates

Germans Savcisens[1], Tina Eliassi-Rad[2,3], Lars Kai Hansen[1], Laust Hvas Mortensen[4,5], Lau Lilleholt[6,7], Anna Rogers[8], Ingo Zettler[6,7] & Sune Lehmann[1,7]

Here we represent human lives in a way that shares structural similarity to language, and we exploit this similarity to adapt natural language processing techniques to examine the evolution and predictability of human lives based on detailed event sequences. We do this by drawing on a comprehensive registry dataset, which is available for Denmark across several years, and that includes information about life-events related to health, education, occupation, income, address and working hours, recorded with day-to-day resolution. We create embeddings of life-events in a single vector space, showing that this embedding space is robust and highly structured. Our models allow us to predict diverse outcomes ranging from early mortality to personality nuances, outperforming state-of-the-art models by a wide margin. Using methods for interpreting deep learning models, we probe the algorithm to understand the factors that enable our predictions. Our framework allows researchers to discover potential mechanisms that impact life outcomes as well as the associated possibilities for personalized interventions.



Méthodes d'interprétabilité mécanistique (Geiger et al., 2025)

Behavioral Methods

- Feature attribution
- Integrated gradients
- Effects of real-world concepts on models

Patching Activations with Interchange Interventions

- Interchange interventions
- Path patching
- Causal mediation analysis

Ablation-Based Analysis

- Concept erasure
- Sub-circuit analysis
- Causal scrubbing

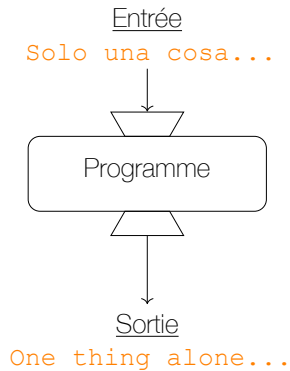
Modular Feature Learning

- Probing
- Sparse autoencoders
- PCA
- Differential Binary Masking
- Difference of means
- Distributed Alignment Search

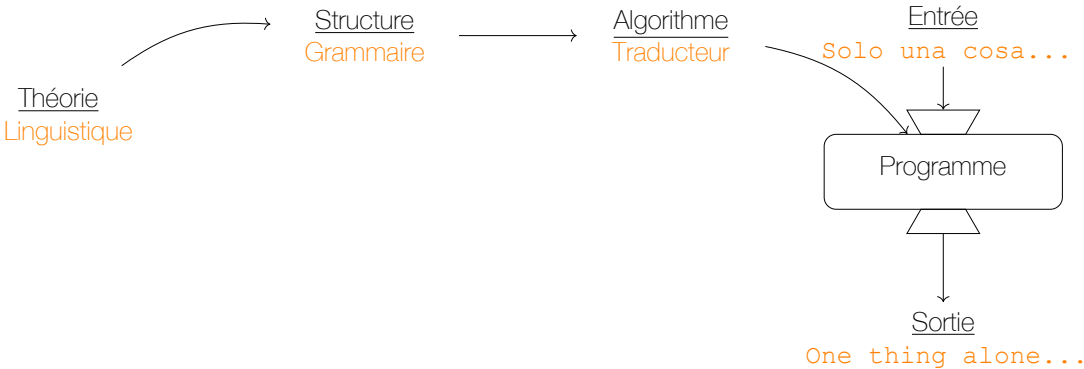
Other Approaches

- Activation Steering
- Training for Interpretability
- Causal Abstraction

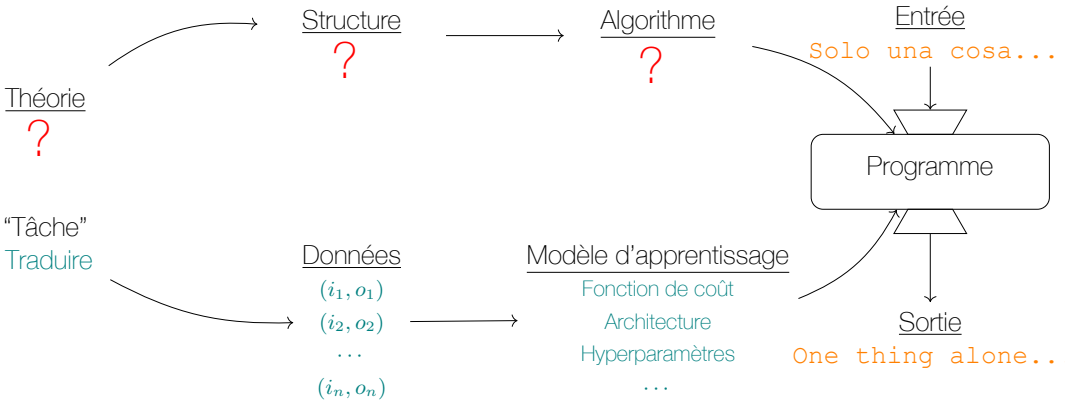
La structure implicite des données



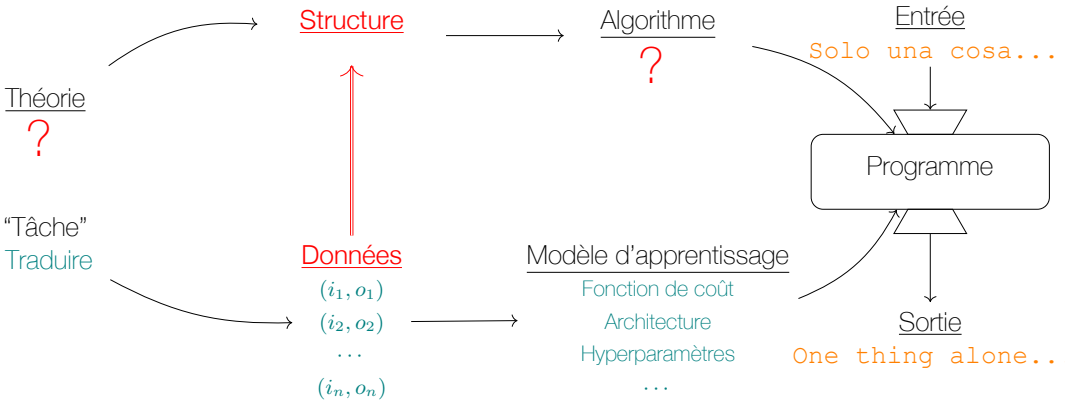
La structure implicite des données



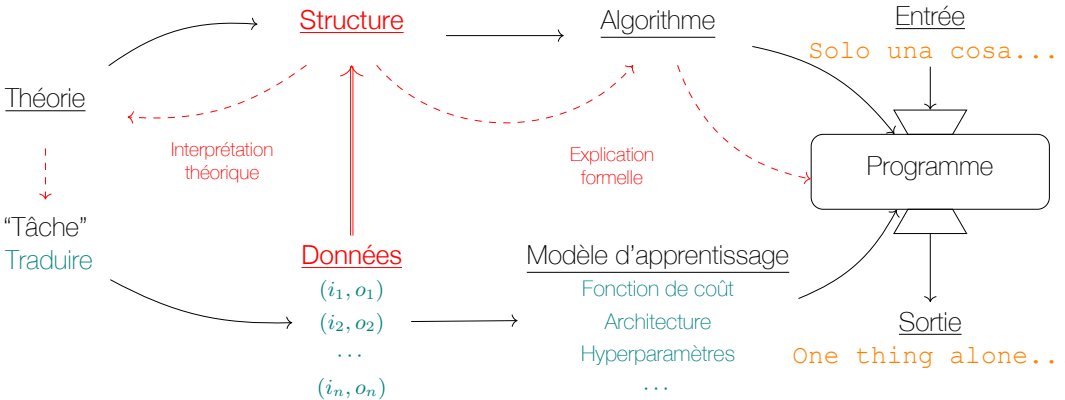
La structure implicite des données



La structure implicite des données

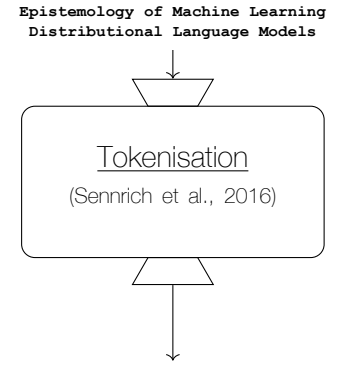


La structure implicite des données



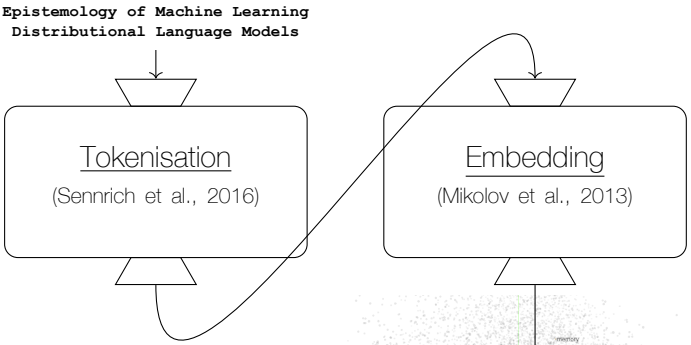
Outline

Explicabilité formelle



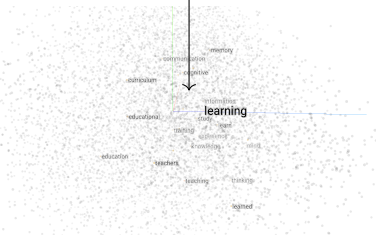
(https://tiktokenizer.vercel.app)

Explicabilité formelle



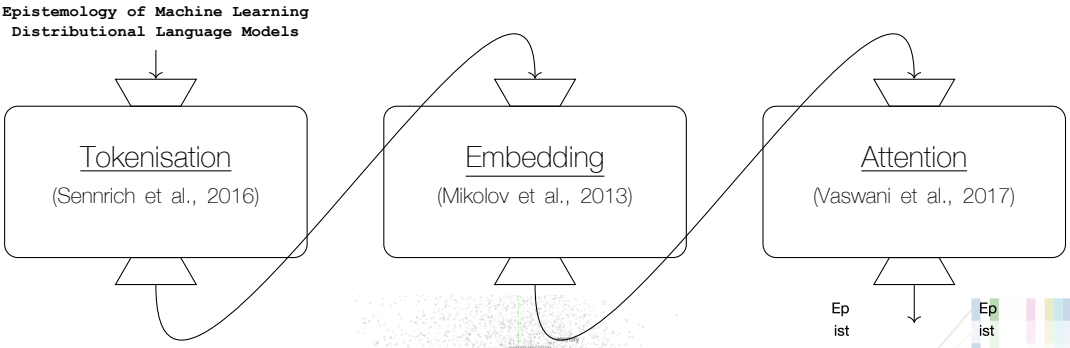
Epistemology of Machine Learning
Distributional Language Models

(https://tiktokenizer.vercel.app)

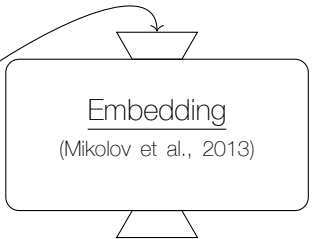


(https://projector.tensorflow.org)

Explicabilité formelle

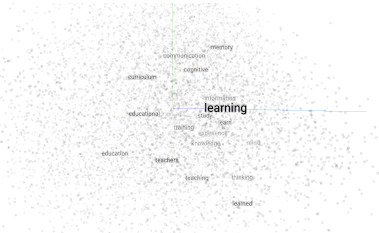


Explicabilité formelle




```
Epistemology of Machine Learning
Distributional Language Models
```

(https://tiktokenizer.vercel.app)



(https://projector.tensorflow.org)

La structure des embeddings

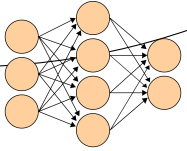
Structure

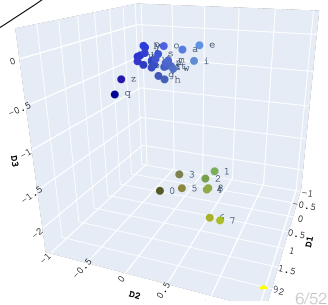
?

{-, /, 0, 1, 2, . . . , 8, 9, =,
a, b, c, . . . , w, x, y, z, é}

Embedding

Données





word2vec expliqué (Levy and Goldberg, 2014)

$$\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)]$$

Où:

$\vec{w}$	=	représentation vectorielle du mot w
$\vec{c}$	=	représentation vectorielle du contexte c
$\sigma(x)$	=	$\frac{1}{1+e^{-x}}$
k	=	nombre d'échantillons "négatifs" (arbitraires)
c_N	=	contexte arbitraire tiré de P_D
$P_D(c)$	=	distribution unigramme empirique de c dans les données D, c-à-d $\frac{\#(c)}{\|D\|}$

word2vec expliqué (Levy and Goldberg, 2014)

$$\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)]$$

word2vec expliqué (Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w, c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

word2vec expliqué (Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w, c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial(\vec{w} \cdot \vec{c})} = 0$$

word2vec expliqué (Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w,c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial(\vec{w} \cdot \vec{c})} = 0 \quad \text{quand} \quad \vec{w} \cdot \vec{c} \ = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

word2vec expliqué (Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w,c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial(\vec{w} \cdot \vec{c})} = 0 \quad \text{quand} \quad \vec{w} \cdot \vec{c} = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

$$= PMI(w,c) - \log k$$

word2vec expliqué (Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w,c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial(\vec{w} \cdot \vec{c})} = 0 \quad \text{quand} \quad \vec{w} \cdot \vec{c} = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

$$= PMI(w,c) - \log k$$

Contrainte supplémentaire :
$\vec{w}$ et $\vec{c}$ doivent être de faible dimension

word2vec expliqué (Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w,c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial(\vec{w} \cdot \vec{c})} = 0 \quad \text{quand} \quad \vec{w} \cdot \vec{c} = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

$$= PMI(w,c) - \log k$$

Contrainte supplémentaire :
$\vec{w}$ et $\vec{c}$ doivent être de faible dimension

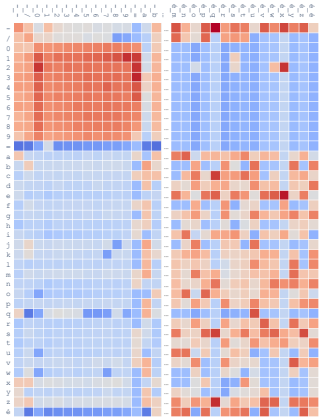
La décomposition en valeurs singulières (SVD) fournit une solution exacte à ce problème d'optimisation.

Exemple: Caractères dans Wikipédia

$$W = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

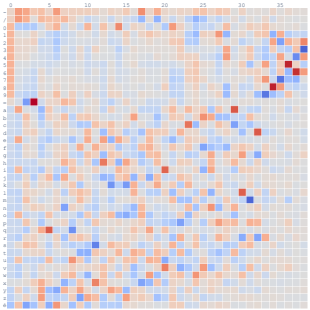
$$C = X \times X = \{\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\}$$

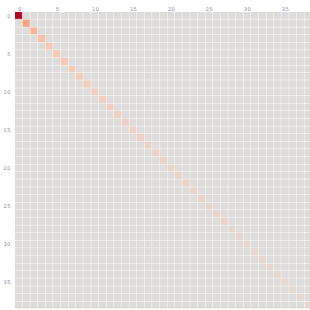
$$\begin{aligned} M_{wc} &= \operatorname{pmi}(w, c) \\ &= \log \frac{p(w, c)}{p(w)p(c)} \end{aligned}$$

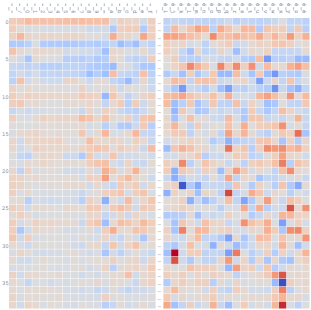


SVD d'une matrice pmi des caractères dans Wikipédia

U Σ V^T

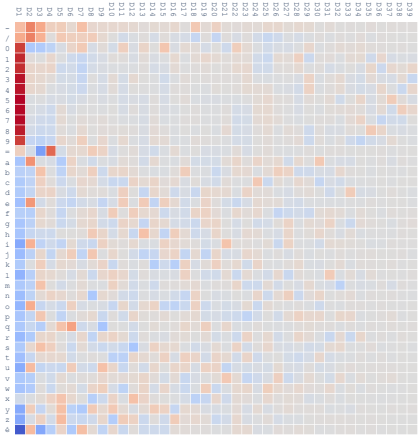






Tronquer

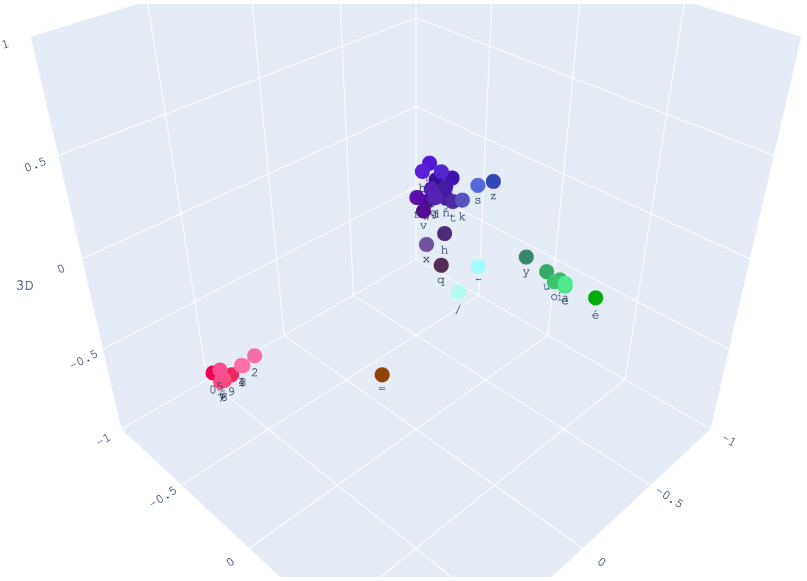
$$U \times \Sigma$$



$$\hat{U} \times \hat{\Sigma}$$

Plonger

$$\hat{U} \times \hat{\Sigma}$$



La structure des embeddings

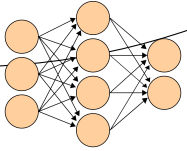
Structure

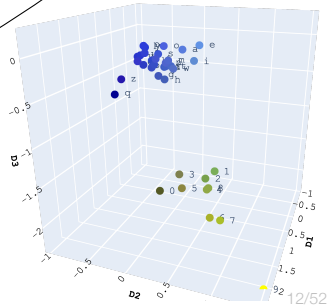
?

{-, /, 0, 1, 2, . . . , 8, 9, =,
a, b, c, . . . , w, x, y, z, é}

Embedding

Données





La structure des embeddings

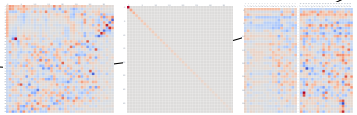
Structure

?

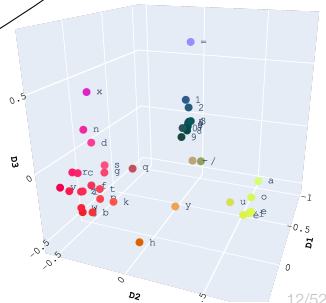
{-, /, 0, 1, 2, . . . , 8, 9, =,
a, b, c, . . . , w, x, y, z, é}

Embedding

Données



SVD



4 Why does this produce good word representations?

Good question. We don't really know.

The distributional hypothesis states that words in similar contexts have similar meanings. The objective above clearly tries to increase the quantity $v_w \cdot v_c$ for good word-context pairs, and decrease it for bad ones. Intuitively, this means that words that share many contexts will be similar to each other (note also that contexts sharing many words will also be similar to each other). This is, however, very hand-wavy.

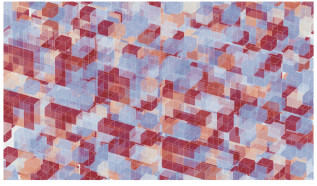
Can we make this intuition more precise? We'd really like to see something more formal.

(Goldberg and Levy, 2014)

The structure of meaning in languge (Bradley et al., 2024)

The Structure of Meaning in Language: Parallel Narratives in Linear Algebra and Category Theory

Tai-Danae Bradley, Juan Luis Gastaldi, and John Terilla



Introduction

Categories for AI, an online program about category theory intelligence in particular. While this article is by no means a comprehensive report on that event, the popularity of "Cats for AI" — the five introductory lectures have been viewed thousands of times — signals the growing prevalence of category theoretic tools in AI.

One way that category theory is gaining traction in machine learning is by providing a formal way to discuss how learning systems can be put together. This article has a different and somewhat narrow focus. It's about how a fundamental piece of AI technology used in language modeling can be understood, with the aid of categorical thinking, as a process that extracts structural features of language from purely syntactical input. The idea that structure arises from form may not be a surprise for many readers — category theoretic ideas have been a major influence in pure

Vecteurs de mots comme fonctions sur des ensembles

$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

$$Y = X \times X = \{\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\}$$

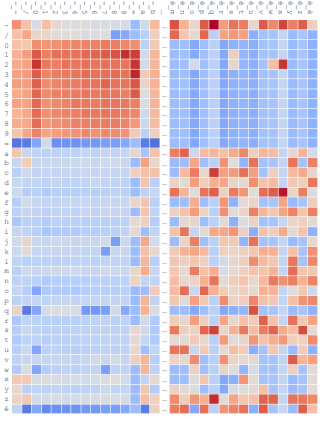
Vecteurs de mots comme fonctions sur des ensembles

$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

$$Y = X \times X = \{\,\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\,\}$$

$$M \colon X \times Y \to \mathbb{R}$$

$$(x, y) \mapsto \mathrm{pmi}(x, y)$$



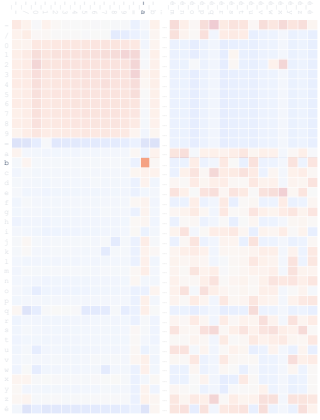
Vecteurs de mots comme fonctions sur des ensembles

$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

$$Y = X \times X = \{\,\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\,\}$$

$$M \colon X \times Y \to \mathbb{R}$$

$$(x, y) \mapsto \mathrm{pmi}(x, y)$$



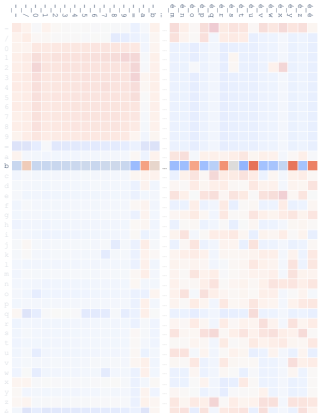
Vecteurs de mots comme fonctions sur des ensembles

$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

$$Y = X \times X = \{\,\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\,\}$$

$$M \colon X \times Y \to \mathbb{R}$$
$$(x, y) \mapsto \mathrm{pmi}(x, y)$$

$$M_x \colon X \to \mathbb{R}^Y$$
$$x \mapsto M(x, -)$$



Vecteurs de mots comme fonctions sur des ensembles

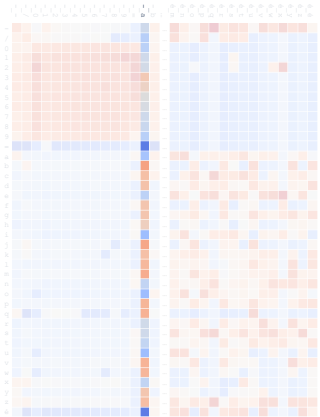
$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

$$Y = X \times X = \{\,\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\,\}$$

$$M \colon X \times Y \to \mathbb{R}$$
$$(x, y) \mapsto \mathrm{pmi}(x, y)$$

$$M_x \colon X \to \mathbb{R}^Y$$
$$x \mapsto M(x, -)$$

$$M_y \colon Y \to \mathbb{R}^X$$
$$y \mapsto M(-, y)$$



Vecteurs de mots comme fonctions sur des ensembles

$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

$$Y = X \times X = \{\,\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\,\}$$

$$M \colon X \times Y \to \mathbb{R}$$
$$(x, y) \mapsto \mathrm{pmi}(x, y)$$

$$M_x \colon X \to \mathbb{R}^Y$$
$$x \mapsto M(x, -)$$

$$M_y \colon Y \to \mathbb{R}^X$$
$$y \mapsto M(-, y)$$

$$X \xrightarrow{\quad M_x \quad} \mathbb{R}^Y$$

$$\mathbb{R}^X \xleftarrow[M_y]{} Y$$

Vecteurs de mots comme fonctions sur des ensembles

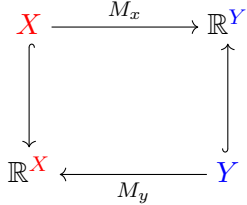
$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

$$Y = X \times X = \{\,\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\,\}$$

$$M \colon X \times Y \to \mathbb{R}$$
$$(x, y) \mapsto \mathrm{pmi}(x, y)$$

$$M_x \colon X \to \mathbb{R}^Y$$
$$x \mapsto M(x, -)$$

$$M_y \colon Y \to \mathbb{R}^X$$
$$y \mapsto M(-, y)$$



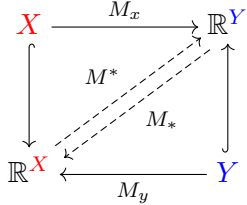
Vecteurs de mots comme fonctions sur des ensembles

$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$
$$Y = X \times X = \{\,\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\,\}$$

$$M \colon X \times Y \to \mathbb{R}$$
$$(x, y) \mapsto \mathrm{pmi}(x, y)$$

$$M_x \colon X \to \mathbb{R}^Y$$
$$x \mapsto M(x, -)$$

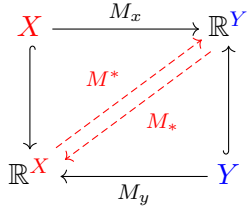
$$M_y \colon Y \to \mathbb{R}^X$$
$$y \mapsto M(-, y)$$



$$M^* \colon \mathbb{R}^X \to \mathbb{R}^Y$$
$$M_* \colon \mathbb{R}^Y \to \mathbb{R}^X$$

Vecteurs de mots comme fonctions sur des ensembles

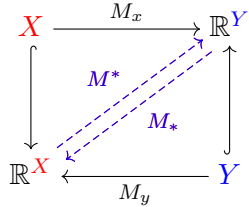
$M_* M^* : \mathbb{R}^X \to \mathbb{R}^X$



Vecteurs de mots comme fonctions sur des ensembles

$M_* M^* : \mathbb{R}^X \to \mathbb{R}^X$

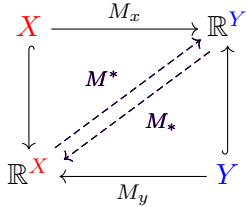
$M^* M_* : \mathbb{R}^Y \to \mathbb{R}^Y$



Vecteurs de mots comme fonctions sur des ensembles

$$M_* M^* \colon \mathbb{R}^X \to \mathbb{R}^X$$
$$M^* M_* \colon \mathbb{R}^Y \to \mathbb{R}^Y$$

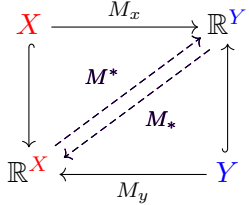
$$\{u_1, \dots, u_m\} \subset \mathbb{R}^X$$
$$\{v_1, \dots, v_n\} \subset \mathbb{R}^Y$$
$$\{\lambda_1, \dots, \lambda_{\min(m,n)}, 0, \dots, 0\}$$



Vecteurs de mots comme fonctions sur des ensembles

$$M_* M^* : \mathbb{R}^X \to \mathbb{R}^X$$
$$M^* M_* : \mathbb{R}^Y \to \mathbb{R}^Y$$

$$\{u_1, \ldots, u_m\} \subset \mathbb{R}^X$$
$$\{v_1, \ldots, v_n\} \subset \mathbb{R}^Y$$
$$\{\lambda_1, \ldots, \lambda_{\min(m,n)}, 0, \ldots, 0\}$$



$$M = U \Sigma V^T$$

$$U := [u_1, \ldots, u_m]$$
$$V := [v_1, \ldots, v_n]$$
$$\Sigma := \begin{bmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\lambda_r} \end{bmatrix}$$

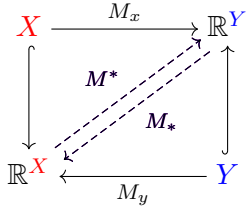
Vecteurs de mots comme fonctions sur des ensembles

$$M_* M^* : \mathbb{R}^X \to \mathbb{R}^X$$
$$M^* M_* : \mathbb{R}^Y \to \mathbb{R}^Y$$

$$\{u_1, \ldots, u_m\} \subset \mathbb{R}^X$$
$$\{v_1, \ldots, v_n\} \subset \mathbb{R}^Y$$
$$\{\lambda_1, \ldots, \lambda_{\min(m,n)}, 0, \ldots, 0\}$$

$$M_* M^* u_i = \lambda_i u_i$$
$$M^* M_* v_i = \lambda_i v_i$$

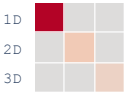
Les u_i and v_i sont des points fixes (linéaires)!



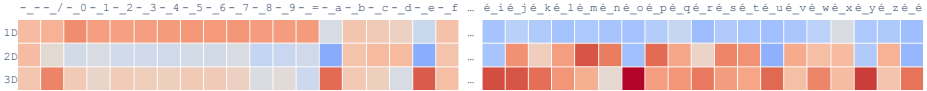
$$M = U \Sigma V^T \qquad U := [u_1, \ldots, u_m] \qquad V := [v_1, \ldots, v_n] \qquad \Sigma := \begin{bmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\lambda_r} \end{bmatrix}$$

Vecteurs propres de $M_* M^*$:

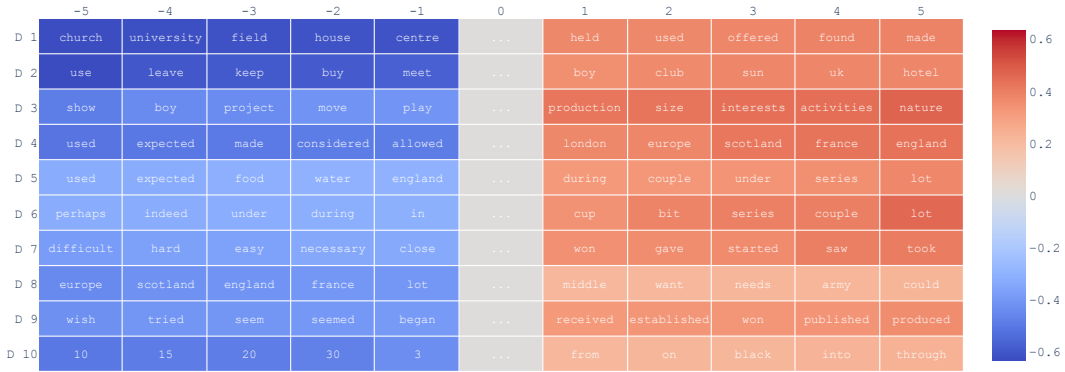
Valeurs propres de $M_* M^*$ et $M^* M_*$:



Vecteurs propres de $M^* M_*$:



Mots


-5 -4 -3 -2 -1 0 1 2 3 4 5
D 1 church university field house centre ... held used offered found made
D 2 use leave keep buy meet ... boy club sun uk hotel
D 3 show boy project move play ... production size interests activities nature
D 4 used expected made considered allowed ... london europe scotland france england
D 5 used expected food water england ... during couple under series lot
D 6 perhaps indeed under during in ... cup bit series couple lot
D 7 difficult hard easy necessary close ... won gave started saw took
D 8 europe scotland england france lot ... middle want needs army could
D 9 wish tried seem seemed began ... received established won published produced
D 10 10 15 20 30 3 ... from on black into through
0.6
0.4
0.2
0
−0.2
−0.4
−0.6

La structure des embeddings

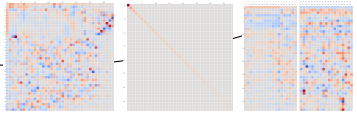
Structure

?

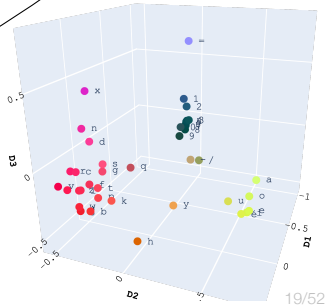
{-, /, 0, 1, 2, . . . , 8, 9, =,
a, b, c, . . . , w, x, y, z, é}

Embedding

Données



SVD



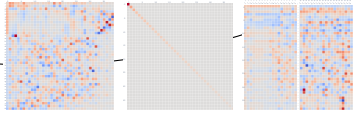
La structure des embeddings

Structure

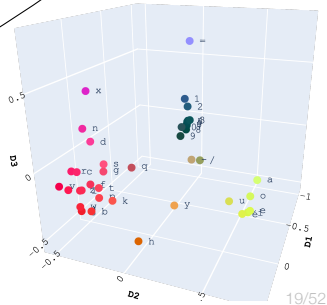
{-, /, 0, 1, 2, . . . , 8, 9, =,
a, b, c, . . . , w, x, y, z, é}

Embedding

Données



SVD



La structure des embeddings

Structure

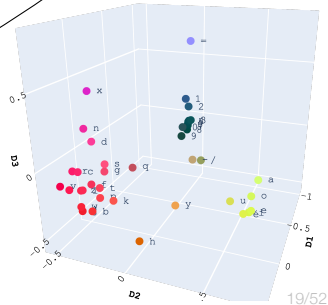
?

Données

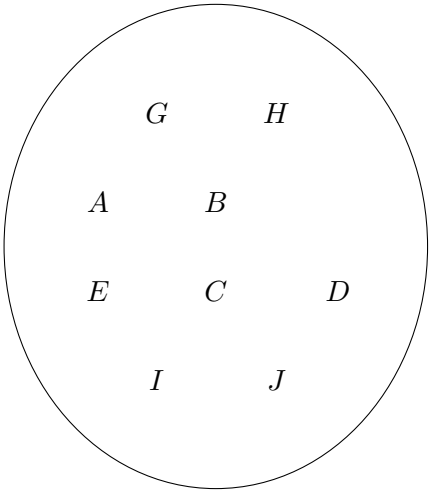
$$C^{op} \times D \to \bar{\mathbb{R}}$$

$\{-, /, 0, 1, 2, \ldots, 8, 9, =,$
$a, b, c, \ldots, w, x, y, z, é\}$

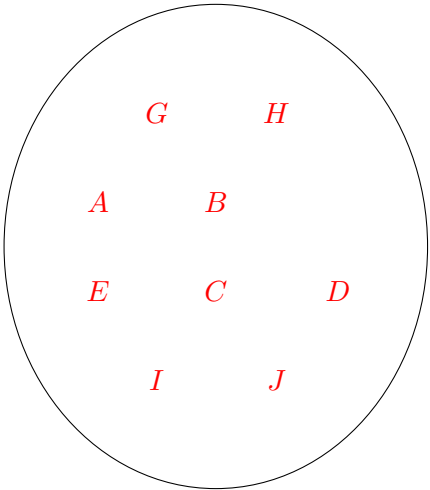
Embedding



Une catégorie est comme un ensemble muni d'une structure



Une catégorie est comme un ensemble muni d'une structure

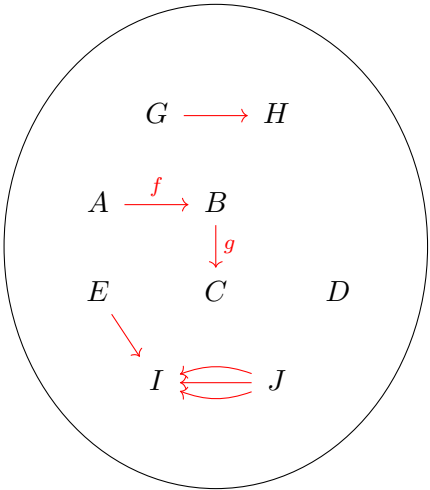


Definition (Category – Awodey, 2010)

Data:

- Objects: $A, B, C, \ldots$

Une catégorie est comme un ensemble muni d'une structure

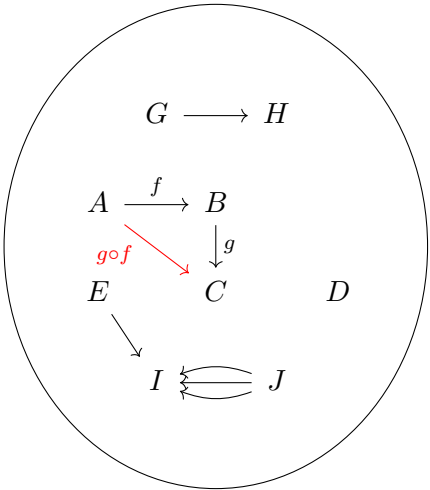


Definition (Category – Awodey, 2010)

Data:

- Objects: $A, B, C, \ldots$
- Arrows: $f, g, \ldots$

Une catégorie est comme un ensemble muni d'une structure



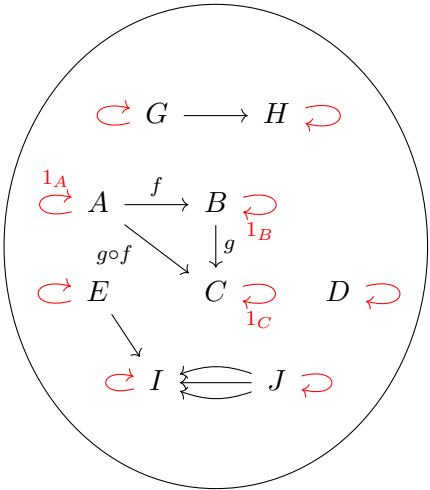
Definition (Category – Awodey, 2010)

Data:

- Objects: $A, B, C, \ldots$
- Arrows: $f, g, \ldots$
- Composition: Given $f : A \to B$ and $g : B \to C$, there is given an arrow

$$g \circ f : A \to C$$

Une catégorie est comme un ensemble muni d'une structure



Definition (Category – Awodey, 2010)

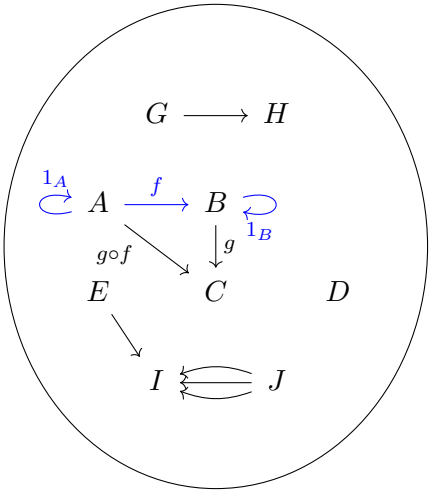
Data:

- Objects: $A, B, C, \ldots$
- Arrows: $f, g, \ldots$
- Composition: Given $f : A \to B$ and $g : B \to C$, there is given an arrow

$$g \circ f : A \to C$$

- Identity: For each A, there is $1_A : A \to A$

Une catégorie est comme un ensemble muni d'une structure



Definition (Category – Awodey, 2010)

Data:

- Objects: $A, B, C, \ldots$
- Arrows: $f, g, \ldots$
- Composition: Given $f : A \to B$ and $g : B \to C$, there is given an arrow

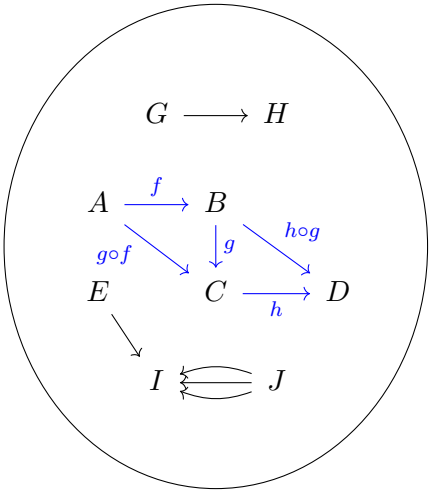
$$g \circ f : A \to C$$

- Identity: For each A, there is $1_A : A \to A$

Laws:

- Unit: $f \circ 1_A = f = 1_B \circ f$

Une catégorie est comme un ensemble muni d'une structure



Definition (Category – Awodey, 2010)

Data:

- Objects: $A, B, C, \ldots$
- Arrows: $f, g, \ldots$
- Composition: Given $f : A \to B$ and $g : B \to C$, there is given an arrow

$$g \circ f : A \to C$$

- Identity: For each A, there is $1_A : A \to A$

Laws:

- Unit: $f \circ 1_A = f = 1_B \circ f$
- Associativity: $f \circ (g \circ h) = (f \circ g) \circ h$

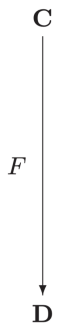
Un foncteur est une application entre catégories

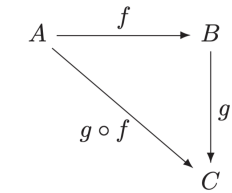
Definition (Functor – Awodey, 2010)

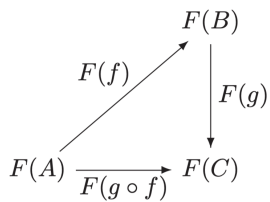
A *functor*

$$F\colon \mathsf{C} \to \mathsf{D}$$

between categories C and D is a mapping of objects to objects and arrows to arrows, in such a way that







Un foncteur est une application entre catégories

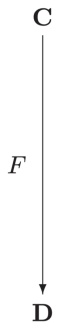
Definition (Functor – Awodey, 2010)

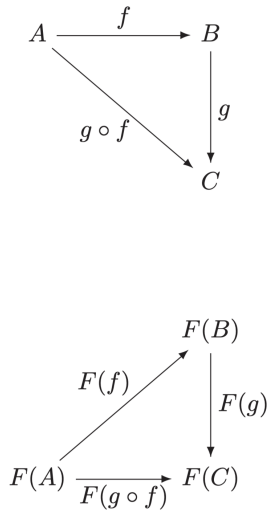
A *functor*

$$F\colon \mathsf{C} \to \mathsf{D}$$

between categories C and D is a mapping of objects to objects and arrows to arrows, in such a way that

(a) $F(f : A \to B) = F(f) : F(A) \to F(B)$





Un foncteur est une application entre catégories

Definition (Functor – Awodey, 2010)

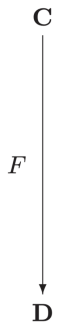
A *functor*

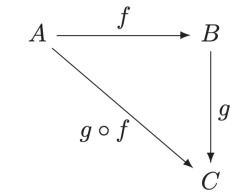
$$F\colon \mathsf{C} \to \mathsf{D}$$

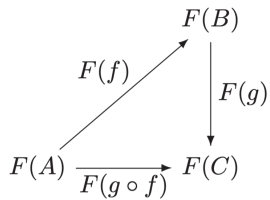
between categories C and D is a mapping of objects to objects and arrows to arrows, in such a way that

(a) $F(f : A \to B) = F(f) : F(A) \to F(B)$

(b) $F(1_A) = 1_{F(A)}$







Un foncteur est une application entre catégories

Definition (Functor – Awodey, 2010)

A *functor*

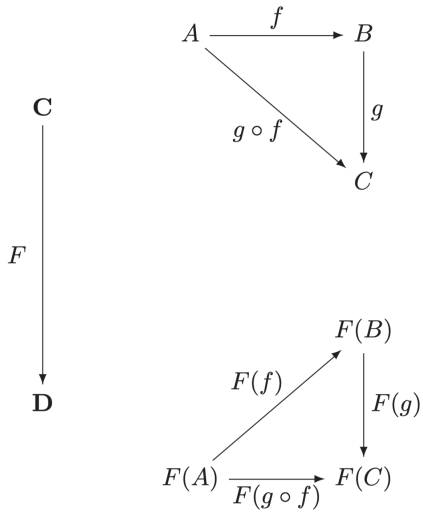
$$F\colon \mathsf{C} \to \mathsf{D}$$

between categories C and D is a mapping of objects to objects and arrows to arrows, in such a way that

(a) $F(f : A \to B) = F(f) : F(A) \to F(B)$

(b) $F(1_A) = 1_{F(A)}$

(c) $F(g \circ f) = F(g) \circ F(f)$



Produit de categories

Definition 2.15. In any category $\mathbf{C}$, a *product diagram* for the objects A and B consists of an object P and arrows

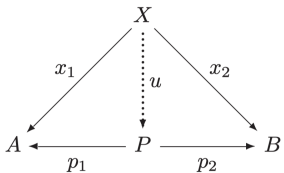
$$A \xleftarrow{p_1} P \xrightarrow{p_2} B$$

satisfying the following UMP:

Given any diagram of the form

$$A \xleftarrow{x_1} X \xrightarrow{x_2} B$$

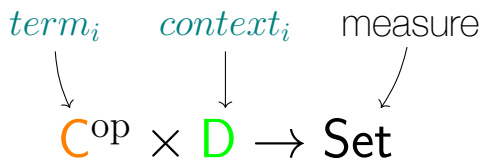
there exists a unique $u : X \to P$, making the diagram



commute, that is, such that $x_1 = p_1 u$ and $x_2 = p_2 u$.

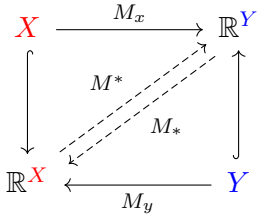
(Awodey, 2010)

Un profoncteur est un foncteur du produit de deux catégories quelconques vers la catégorie Set



Opérateur “distributionnel” et points fixes

$$M : X \times Y \to \mathbb{R}$$

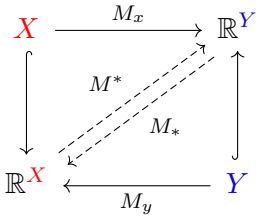


$$\left.\begin{array}{l} M_* M^* : \mathbb{R}^X \to \mathbb{R}^X \\ M^* M_* : \mathbb{R}^Y \to \mathbb{R}^Y \end{array}\right\} \text{ opérateur distributionnel}$$

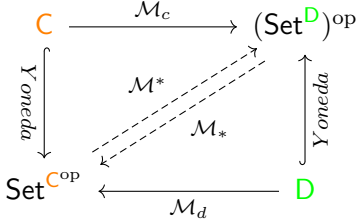
$$\left.\begin{array}{l} \{u_i \in \mathbb{R}^X \,|\, M_* M^* u_i = \lambda_i u_i\} \\ \{v_i \in \mathbb{R}^Y \,|\, M^* M_* v_i = \lambda_i v_i\} \end{array}\right\} \text{ points fixes}$$

Opérateur “distributionnel” et points fixes

$M : X \times Y \to \mathbb{R}$



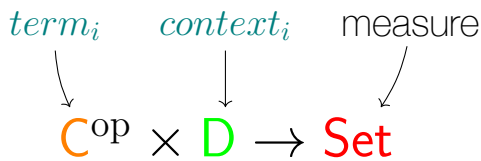
$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \mathsf{Set}$



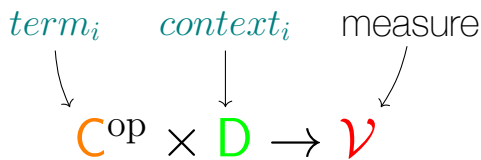
$$\left.\begin{array}{l} M_*M^* \colon \mathbb{R}^X \to \mathbb{R}^X \\ M^*M_* \colon \mathbb{R}^Y \to \mathbb{R}^Y \end{array}\right\} \text{ opérateur distributionnel } \left\{\begin{array}{l} \mathcal{M}_*\mathcal{M}^* \colon \mathsf{Set}^{\mathsf{C}^{\mathrm{op}}} \to \mathsf{Set}^{\mathsf{C}^{\mathrm{op}}} \\ \mathcal{M}^*\mathcal{M}_* \colon (\mathsf{Set}^{\mathsf{D}})^{\mathrm{op}} \to (\mathsf{Set}^{\mathsf{D}})^{\mathrm{op}} \end{array}\right.$$

$$\left.\begin{array}{l} \{u_i \in \mathbb{R}^X \mid M_*M^*u_i = \lambda_i u_i\} \\ \{v_i \in \mathbb{R}^Y \mid M^*M_*v_i = \lambda_i v_i\} \end{array}\right\} \text{ points fixes } \left\{\begin{array}{l} \{f \in \mathsf{Set}^{\mathsf{C}^{\mathrm{op}}} \mid \mathcal{M}_*\mathcal{M}^*(f) \cong f\} \\ \{g \in (\mathsf{Set}^{\mathsf{D}})^{\mathrm{op}} \mid \mathcal{M}^*\mathcal{M}_*(g) \cong g\} \end{array}\right.$$

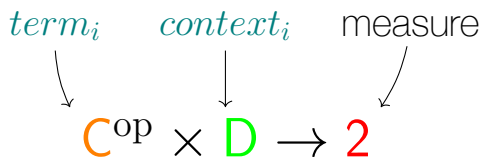
Profoncteur de catégories enrichies



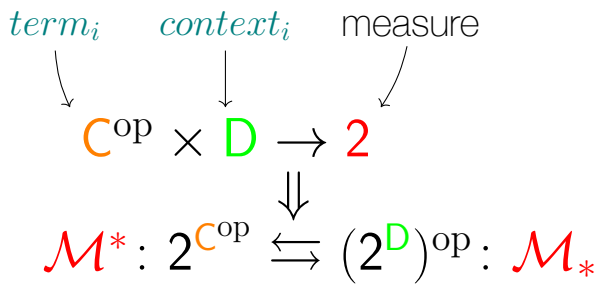
Profoncteur de catégories enrichies



Profoncteur de catégories enrichies

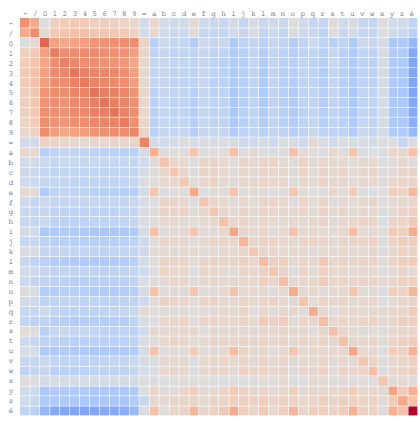


Profoncteur de catégories enrichies

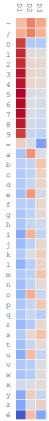


Points fixes booléans

$$M_* M^* u = \lambda u$$



$\times$

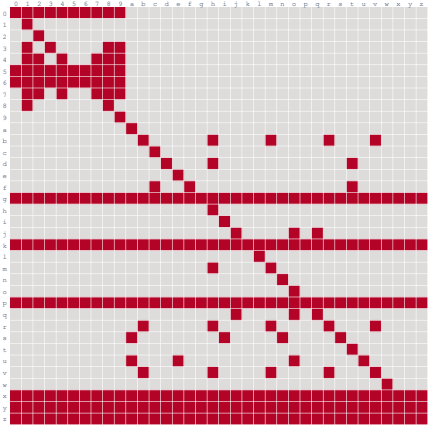


$=$

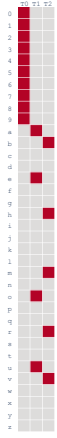
$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix}$$

Points fixes booléans

$$\mathcal{M}_* \mathcal{M}^* f = f$$

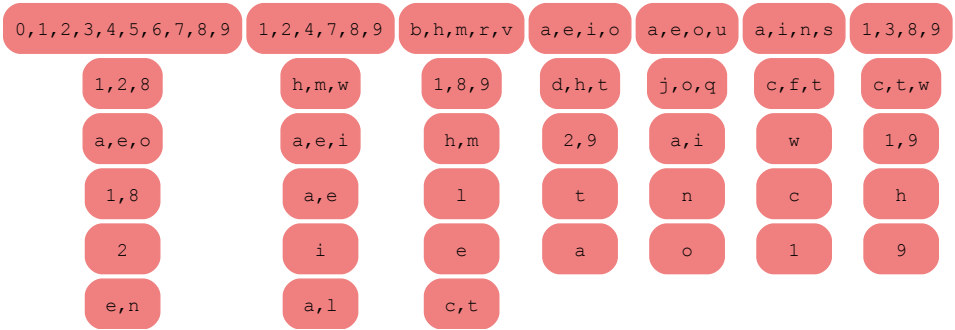


$\star$

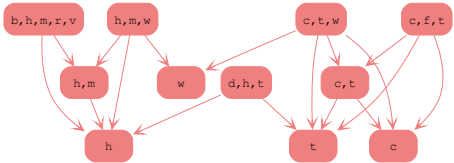


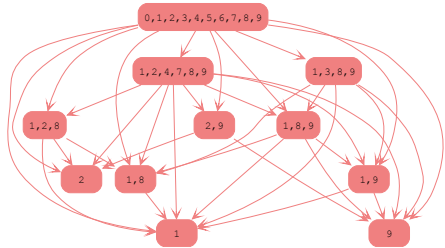
“Eigensets”

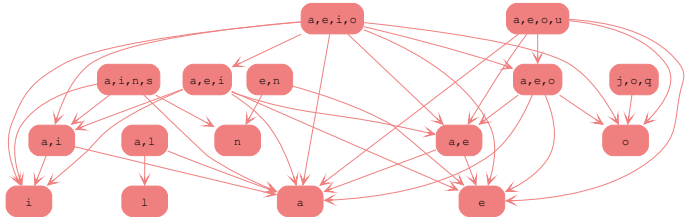
$$\mathcal{M}_* \mathcal{M}^* f = f$$



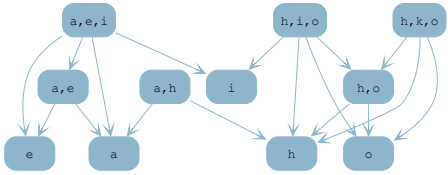
Structure d'ordre partiel

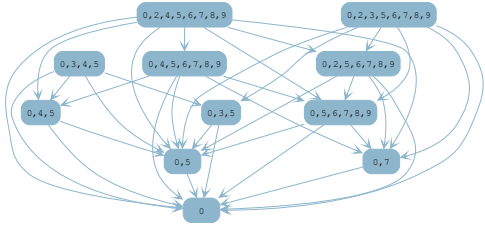


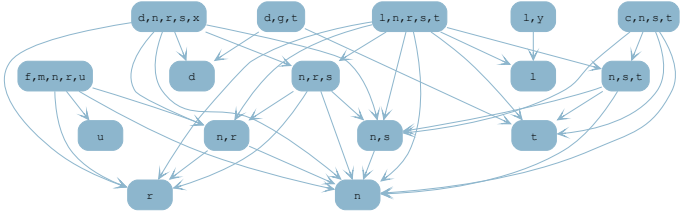




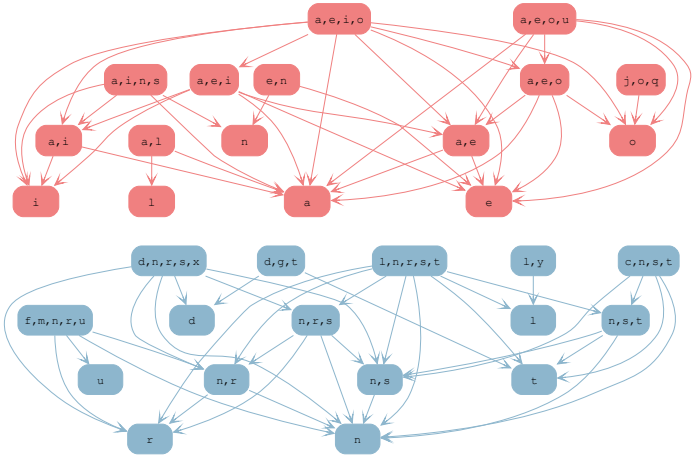
Ordre partiel dual



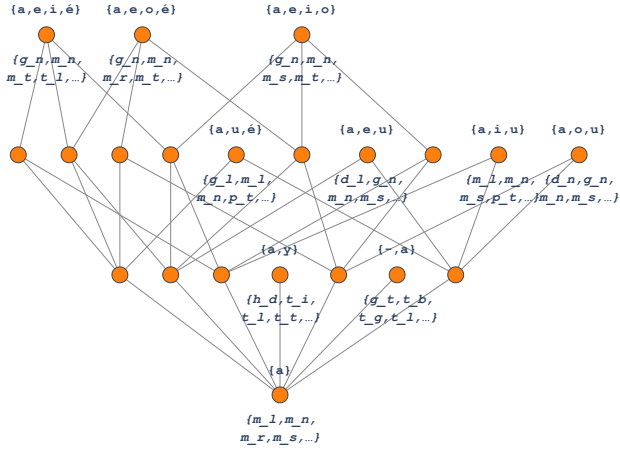


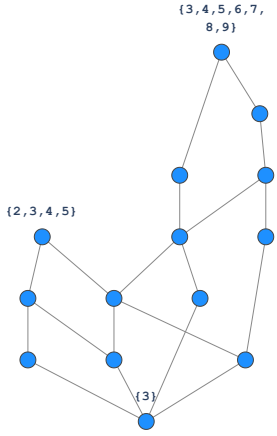


Couplage des points fixes

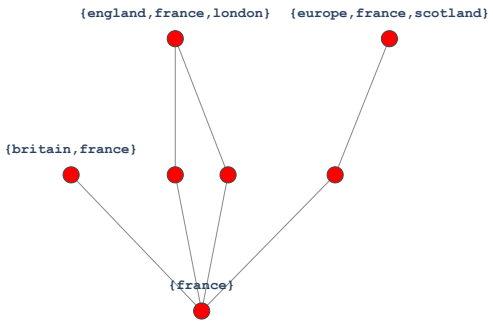


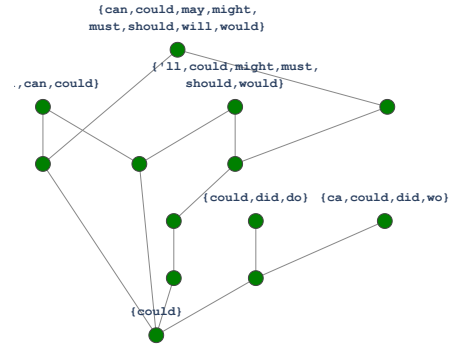
Concepts formels



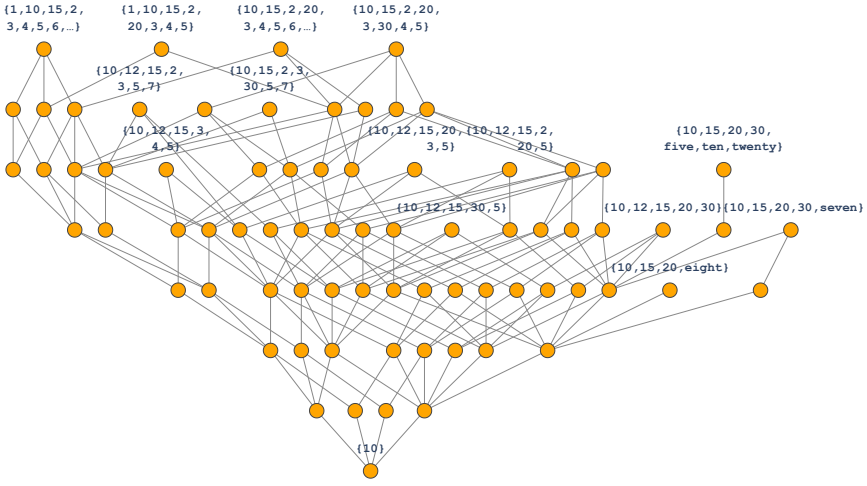


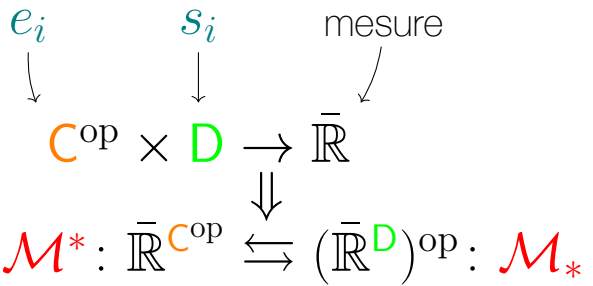
Concepts formels (mots)



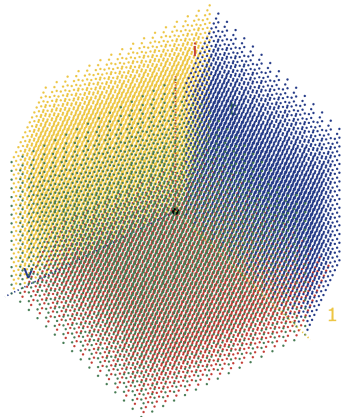


Concepts formels (mots)



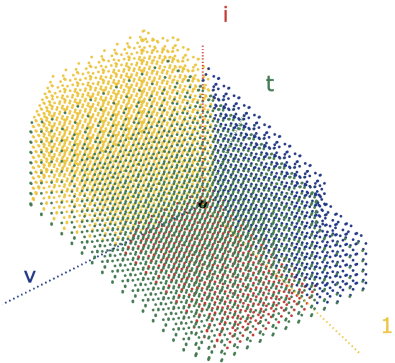

e_i
s_i
mesure
$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \bar{\mathbb{R}}$
$\mathcal{M}^* : \bar{\mathbb{R}}^{\mathsf{C}^{\mathrm{op}}} \leftrightarrows (\bar{\mathbb{R}}^{\mathsf{D}})^{\mathrm{op}} : \mathcal{M}_*$

Profoncteur et structures du noyau

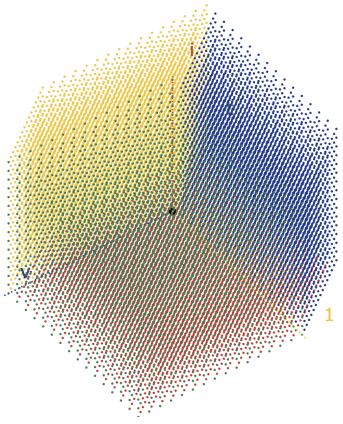


$$\begin{array}{ccc} \mathsf{C}^{\mathrm{op}} \times \mathsf{D} & \longrightarrow & \bar{\mathbb{R}} \\ & \Downarrow & \\ \mathcal{M}^* \colon \bar{\mathbb{R}}^{\mathsf{C}^{\mathrm{op}}} & \leftrightarrows & \left(\bar{\mathbb{R}}^{\mathsf{D}}\right)^{\mathrm{op}} \colon \mathcal{M}_* \end{array}$$

Profoncteur et structures du noyau



$\mathcal{M}_* \mathcal{M}^*$



$$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \longrightarrow \bar{\mathbb{R}}$$

$$\Downarrow$$

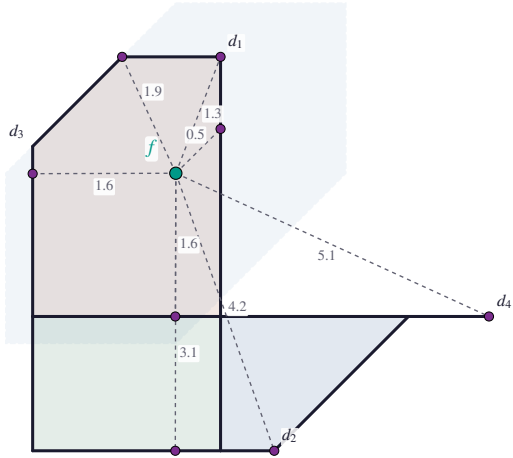
$$\mathcal{M}^* : \bar{\mathbb{R}}^{\mathsf{C}^{\mathrm{op}}} \leftrightarrows \left(\bar{\mathbb{R}}^{\mathsf{D}}\right)^{\mathrm{op}} : \mathcal{M}_*$$

Géométrie du noyau (Gastaldi et al., 2026c)

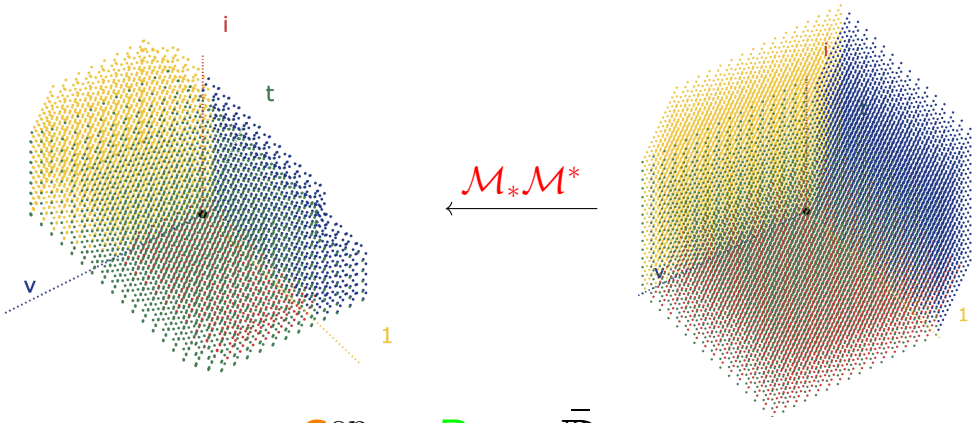
Projective metric geometry of tropical nuclei: gap matrices, event loci, and order chambers

Juan Luis Gastaldi[1†], Samantha Jarvis[2†], Thomas Seiller[3†], John Terilla[2,4†]

[1]ETH Zurich, Zurich, Switzerland.
[2]Queens College, City University of New York, New York, NY, USA.
[3]CNRS, Paris, France.
[4]The Graduate Center, City University of New York, New York, NY, USA.



Géométrie du noyau

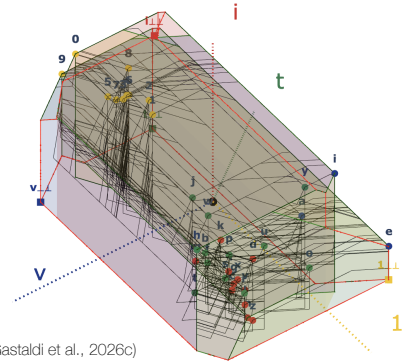


$$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \longrightarrow \bar{\mathbb{R}}$$

$$\Downarrow$$

$$\mathcal{M}^* : \bar{\mathbb{R}}^{\mathsf{C}^{\mathrm{op}}} \leftrightarrows \left(\bar{\mathbb{R}}^{\mathsf{D}}\right)^{\mathrm{op}} : \mathcal{M}_*$$

Géométrie du noyau



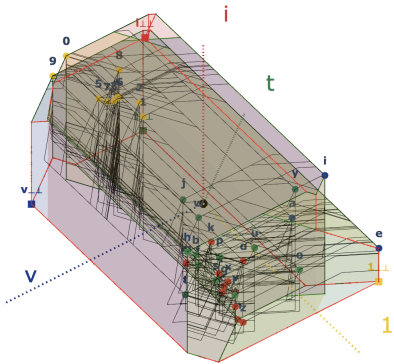
(Gastaldi et al., 2026c)

$$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \longrightarrow \bar{\mathbb{R}}$$

$$\Downarrow$$

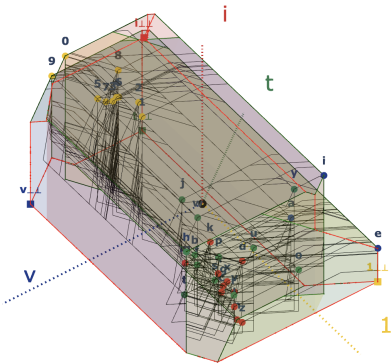
$$\mathcal{M}^* : \bar{\mathbb{R}}^{\mathsf{C}^{\mathrm{op}}} \leftrightarrows \left(\bar{\mathbb{R}}^{\mathsf{D}}\right)^{\mathrm{op}} : \mathcal{M}_*$$

Géométrie du noyau



- Action de jauge
- Métrique projective tropicale interne
- Décomposition en cellules polhyedriques
- Invariant par jauge externe

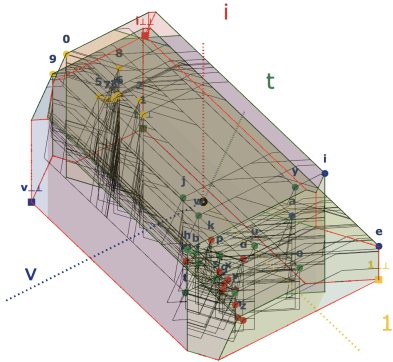
Géométrie du noyau



$$\mathcal{M} = \mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \bar{\mathbb{R}}$$
$$\bar{\mathbb{R}} = ([-\inf, \inf], \leq, +)$$

- Action de jauge
- Métrique projective tropicale interne
- Décomposition en cellules polhyedriques
- Invariant par jauge externe

Géométrie du noyau

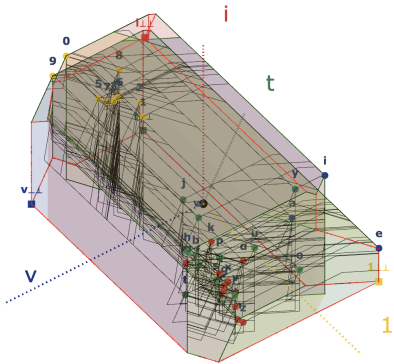


$$\mathcal{M} = \mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \bar{\mathbb{R}}$$
$$\bar{\mathbb{R}} = ([-\inf, \inf], \leq, +)$$

$$\mathcal{M}^* f(d) := \min_{c \in \mathsf{C}} (\mathcal{M}(c, d) - f(c))$$
$$\mathcal{M}^* g(c) := \min_{d \in \mathsf{D}} (\mathcal{M}(c, d) - g(d))$$

- Action de jauge
- Métrique projective tropicale interne
- Décomposition en cellules polhyedriques
- Invariant par jauge externe

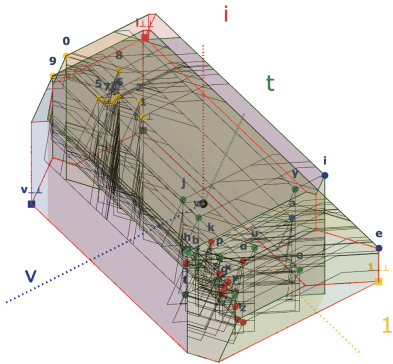
Géométrie du noyau



$$(f, g) \in \mathrm{Nuc}(\mathcal{M}) \implies (f + \lambda, g - \lambda) \in \mathrm{Nuc}(\mathcal{M})$$

- Action de jauge
- Métrique projective tropicale interne
- Décomposition en cellules polhyedriques
- Invariant par jauge externe

Géométrie du noyau

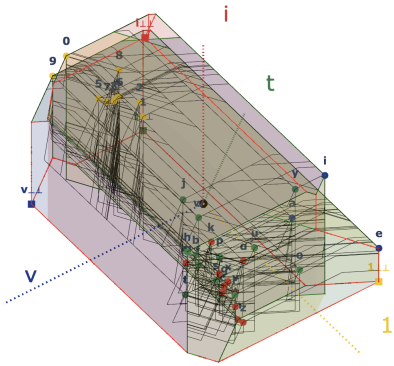


$$(f, g) \in \mathrm{Nuc}(\mathcal{M}) \implies (f + \lambda, g - \lambda) \in \mathrm{Nuc}(\mathcal{M})$$

La projectivisation $\mathbb{P}\mathrm{Nuc}(\mathcal{M})$ est un
espace compact polyédrique

- Action de jauge
- Métrique projective tropicale interne
- Décomposition en cellules polhyedriques
- Invariant par jauge externe

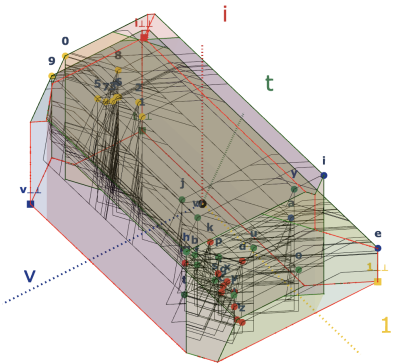
Géométrie du noyau



- Action de jauge
- Métrique projective tropicale interne
- Décomposition en cellules polhyedriques
- Invariant par jauge externe

$$d_{\mathsf{C}}(f, f') = \max_c(f(c) - f'(c)) - \min_c(f(c) - f'(c))$$

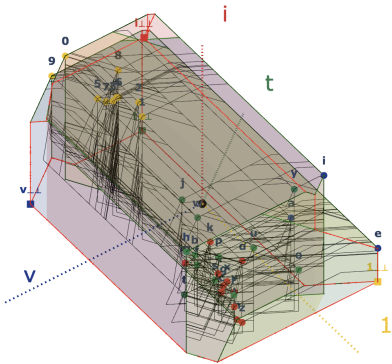
Géométrie du noyau



Cellules définies par les par les couples (c, d) pour lesquelles $f(c) + g(d) = M(c, d)$

- Action de jauge
- Métrique projective tropicale interne
- Décomposition en cellules polhyedriques
- Invariant par jauge externe

Géométrie du noyau



$$\delta^{f,g}(c,d) = \mathcal{M}(c,d) - f(c) - g(d)$$

- Action de jauge
- Métrique projective tropicale interne
- Décomposition en cellules polhyedriques
- Invariant par jauge externe

Logique du noyau (Gastaldi et al., 2026b)

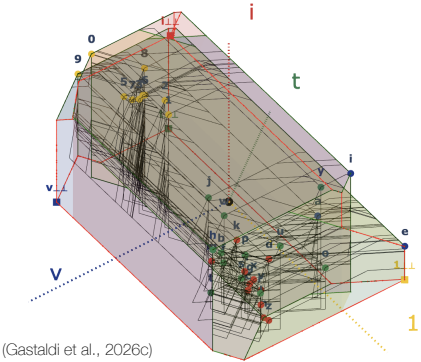
A CALCULUS OF TYPES IN ISBELL NUCLEI

JUAN LUIS GASTALDI, SAMANTHA JARVIS, THOMAS SEILLER, AND JOHN TERILLA

Abstract. We identify two constructions from different mathematical traditions. In linear logic and realisability, logical types are generated rather than fixed in advance: one begins with a universe of realisers equipped with execution, uses orthogonality to test their interactions, and takes types to be the biorthogonally closed subsets. In enriched Isbell duality, a quantitative relation induces an adjunction whose fixed points form a category, its nucleus. These constructions proceed by different means; we show that, in the present setting, they produce the same objects.

The shared datum is minimal: an associative product, called execution, and a real-valued measurement, with no compatibility assumed between them. The failure of the measurement to be additive is at once the relation defining orthogonality and the quantitative relation whose Isbell nucleus we form, and the types cut out by orthogonality are exactly the fixed points of the associated adjunction. The identification pays off in both directions. The most natural product of types fails to be associative; repairing this failure forces a different notion of type, sensitive to both sides of a composite, on which the induced product is associative and, when execution has units, carries two residuals. What emerges is a noncommutative Lambek calculus, derived directly from execution and orthogonality rather than imposed. In the reverse direction, each such type, read on the categorical side, generates a quantitative relation of its own, and with it a derived adjunction and a further generation of types; these derived types are again types of the original situation, computed by the residuals of the Lambek calculus. We also prove a coherence theorem for the threefold arrangements of this construction and, in the finite-dimensional case, give explicit formulas for the product.

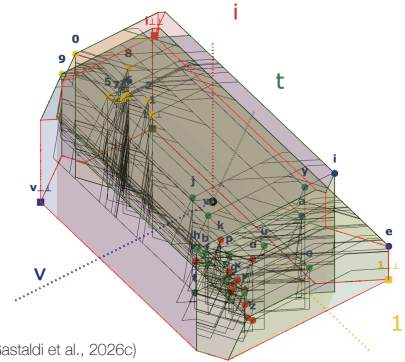
Logique du noyau



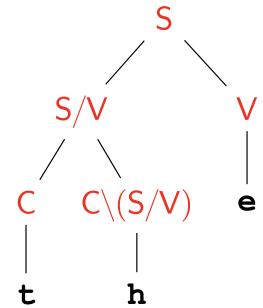
(Gastaldi et al., 2026c)

$$\begin{array}{ccc} \mathsf{C}^{\mathrm{op}} \times \mathsf{D} & \longrightarrow & \bar{\mathbb{R}} \\ & \Downarrow & \\ \mathcal{M}^* : \bar{\mathbb{R}}^{\mathsf{C}^{\mathrm{op}}} & \leftrightarrows & \left(\bar{\mathbb{R}}^{\mathsf{D}}\right)^{\mathrm{op}} : \mathcal{M}_* \end{array}$$

Logique du noyau



(Gastaldi et al., 2026c)



(Gastaldi & Pellissier, 2021)
(Gastaldi et al., 2026a)

$$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \bar{\mathbb{R}}$$

$$\Downarrow$$

$$\mathcal{M}^* : \bar{\mathbb{R}}^{\mathsf{C}^{\mathrm{op}}} \leftrightarrows \left(\bar{\mathbb{R}}^{\mathsf{D}}\right)^{\mathrm{op}} : \mathcal{M}_*$$

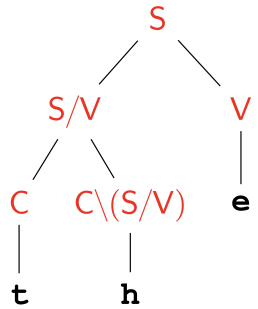
Logique du Noyau

Définition: Polaire/Orthogonal - Girard, 2006

[É]tant donnée une fonction binaire $a, b \rightsquigarrow \langle a|b\rangle : A \times B \to C$ et un sous-ensemble $P \subset C$ (le « pôle »), on peut définir le *polaire* $X^\perp \subset B$ d'un sous-ensemble $X \subset A$ (resp. $Y^\perp \subset A$ d'un sous-ensemble $Y \subset B$) par :

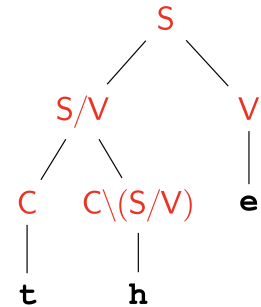
$$X^\perp := \{y \in B : \forall x \in X, \langle a|b\rangle \in P\}$$
$$Y^\perp := \{x \in A : \forall y \in Y, \langle a|b\rangle \in P\}$$

- L'application « polaire » est décroissante: $X \subset X' \Rightarrow X'^\perp \subset X^\perp$.
- L'ensemble $\mathrm{Pol}(A) \subset \mathcal{P}(A)$ des ensembles *polaires*, i.e., de la forme $Y^\perp$, est stable par intersections arbitraires. En particulier, A est polaire et $X^{\perp\perp}$ est le plus petit ensemble polaire contenant X.
- En conséquence, $X^{\perp\perp\perp} = X^\perp$.



Logique du Noyau

- Réalisabilité linéaire (Seiller, 2024)
- Produit tensoriel sur les types
- Implications à droite et à gauche

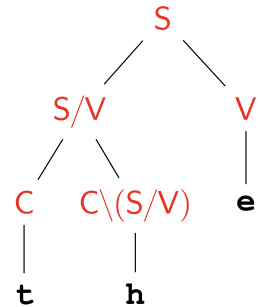


- Ensemble P
- $\bullet : P \times P \to P$ (opération associative)
- $[\![\cdot,\cdot]\!]_m : P \times P \to \mathbf{R}$ (“measurement”)

$$[\![p_1 \bullet p_2, p_3]\!]_m + [\![p_1, p_2]\!]_m = [\![p_1, p_2 \bullet p_3]\!]_m + [\![p_2, p_3]\!]_m$$

Logique du Noyau

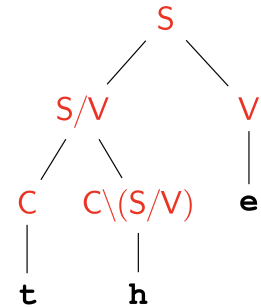
- Réalisabilité linéaire (Seiller, 2024)
- Produit tensoriel sur les types
- Implications à droite et à gauche



$$A \cdot B = {}^{\perp}(\{a \cdot b \mid a \in A, b \in B\}^{\perp})$$

Logique du Noyau

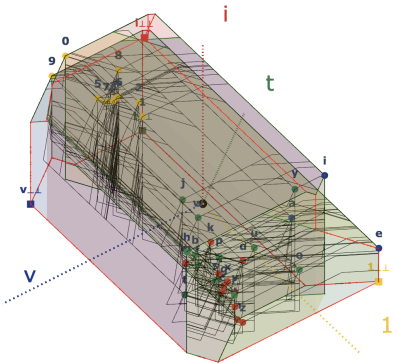
- Réalisabilité linéaire (Seiller, 2024)
- Produit tensoriel sur les types
- Implications à droite et à gauche



$$A \cdot B = {}^{\perp}(\{a \cdot b \mid a \in A, b \in B\}^{\perp})$$

$$A \multimap_r B = \{f \mid \forall a \in A, fa \in B\}$$
$$A \multimap_r B = \{f \mid \forall a \in A, af \in B\}$$

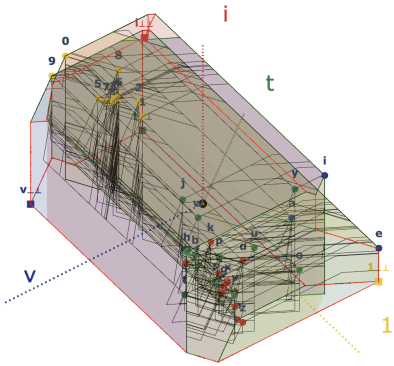
Explicabilité formelle



Embedding

Géométrie et types

Explicabilité formelle



Tokenisation
Renormalisation

Embedding
Géométrie et types

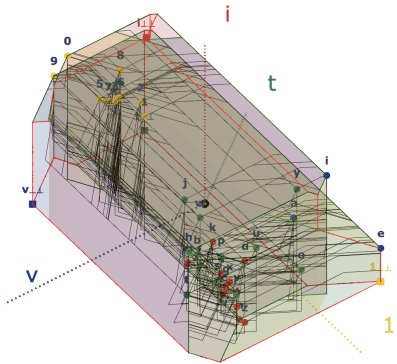
Attention
Produit monoïdal

Outline

Interprétabilité théorique

Théorie

?



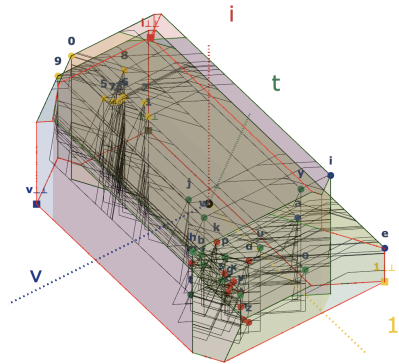
Interprétabilité théorique

Hypothèse distributionnelle

Le contenu des unités linguistiques est déterminé par leur *distribution* dans un corpus.

$$\mathbf{C}^{\mathrm{op}} \times \mathbf{D} \rightarrow \bar{\mathbb{R}}$$

Théorie



Interprétabilité théorique

Hypothèse distributionnelle

Le contenu des unités linguistiques est déterminé par leur *distribution* dans un corpus.

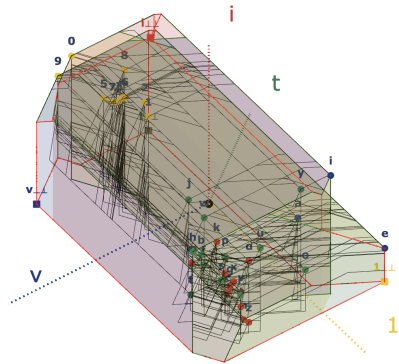
$$\mathbf{C}^{\mathrm{op}} \times \mathbf{D} \rightarrow \bar{\mathbb{R}}$$

Théorie

$$\bar{\mathbb{R}}^{\mathbf{C}^{\mathrm{op}}} \leftrightarrows (\bar{\mathbb{R}}^{\mathbf{D}})^{\mathrm{op}}$$

Hypothèse structurale

Le contenu linguistique est l'effet d'une *structure* virtuelle dérivée des pratiques linguistiques dans une communauté.

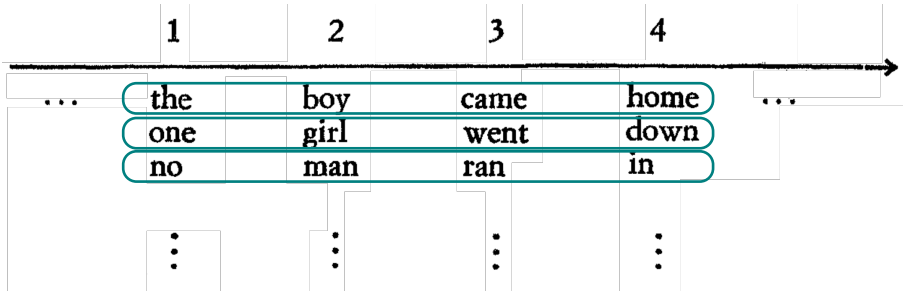


L'Hypothèse Structurale

"Il semble légitime en tous cas de poser a priori l'hypothèse qu'à tout *processus* répond un *système* qui permette de l'analyser et de le décrire au moyen d'un nombre restreint de prémisses. Il doit être possible de considérer tout processus comme composé d'un nombre limité d'éléments qui réapparaissent constamment dans de nouvelles combinaisons. On devrait pouvoir, en se fondant sur l'analyse du processus, regrouper ces éléments en classes, chaque classe étant définie par l'homogénéité de ses possibilités combinatoires, et pouvoir, à partir de ce classement préalable, établir un calcul général exhaustif des combinaisons possibles."

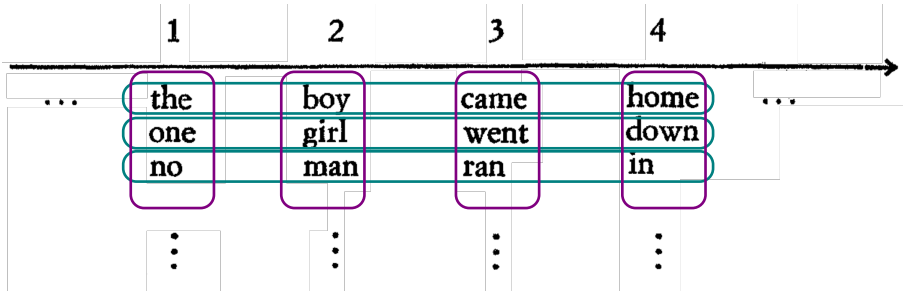
(Hjelmslev, *Prolégomènes à une théorie du langage*, p. 16)

Points fixes comme paradigmes



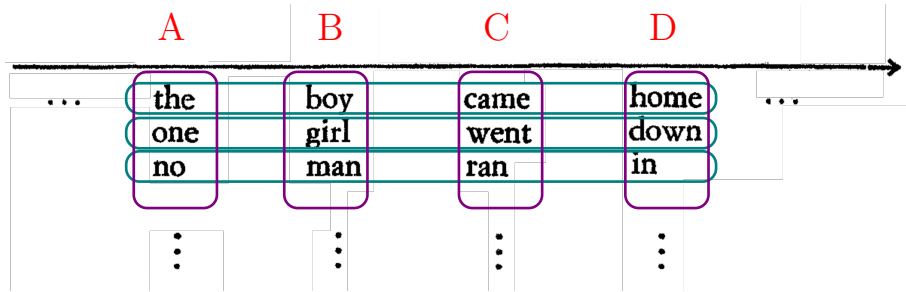
(Hjelmslev, 1971a)

Points fixes comme paradigmes



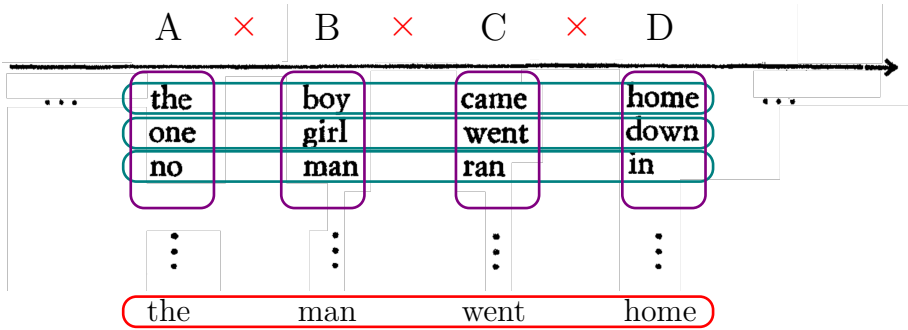
(Hjelmslev, 1971a)

Points fixes comme paradigmes



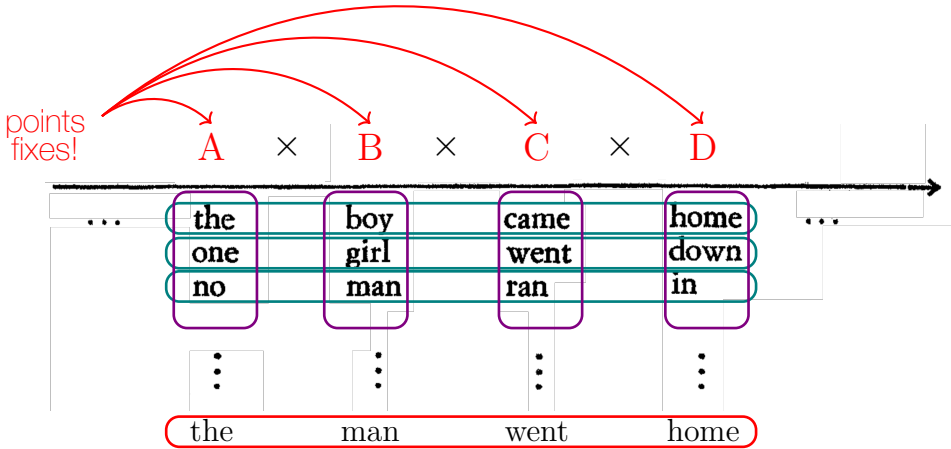
(Hjelmslev, 1971a)

Points fixes comme paradigmes



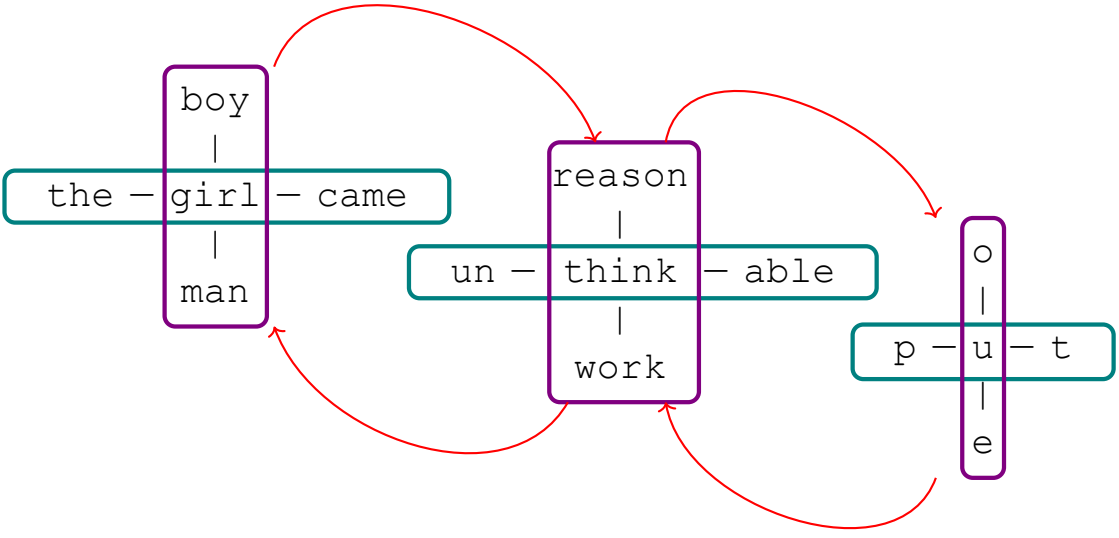
(Hjelmslev, 1971a)

Points fixes comme paradigmes

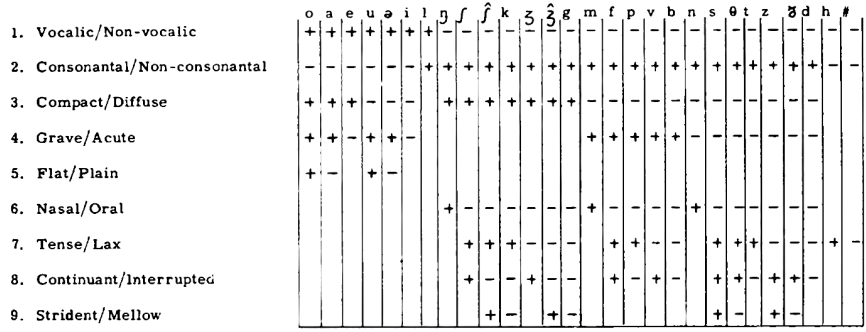


(Hjelmslev, 1971a)

Structure stratifiée



Points fixes comme traits distinctifs



(Jakobson et al., 1952)

Points fixes comme traits distinctifs

points fixes!!!

points fixes!!!

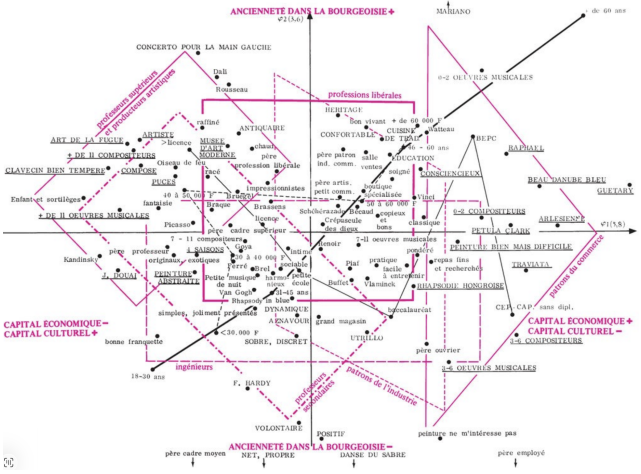
	o	a	e	u	ə	i	l	ŋ	ʃ	ʃ̂	k	z	ʒ̂	g	m	f	p	v	b	n	s	θ	t	z	ʒ	d	h	#
1. Vocalic/Non-vocalic	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
2. Consonantal/Non-consonantal	−	−	−	−	−	−	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−
3. Compact/Diffuse	+	+	+	−	−	−		+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−			
4. Grave/Acute	+	+	−	+	+	−									+	+	+	+	+	−	−	−	−	−	−	−			
5. Flat/Plain	+	−		+	−																								
6. Nasal/Oral								+	−	−	−	−	−	−	+	−	−	−	−	+	−	−	−	−	−	−			
7. Tense/Lax									+	+	+	−	−			+	+	−	−		+	+	+	−	−	−	+	−
8. Continuant/Interrupted									+	−	−	+	−	−		+	−	+	−		+	+	−	+	+	−		
9. Strident/Mellow										+	−		+	−							+	−		+	−			

(Jakobson et al., 1952)

Noyau et espace social

"S'il est vrai que [...] la classe dominante constitue un espace relativement autonome dont la structure est définie par la distribution entre ses membres des différentes espèces de capital, [...] on doit retrouver ces structures dans l'espace des styles de vie [...]. C'est ce que l'on a essayé d'établir en soumettant à l'analyse des correspondances l'ensemble des données recueillies."

(Bourdieu, 1979)

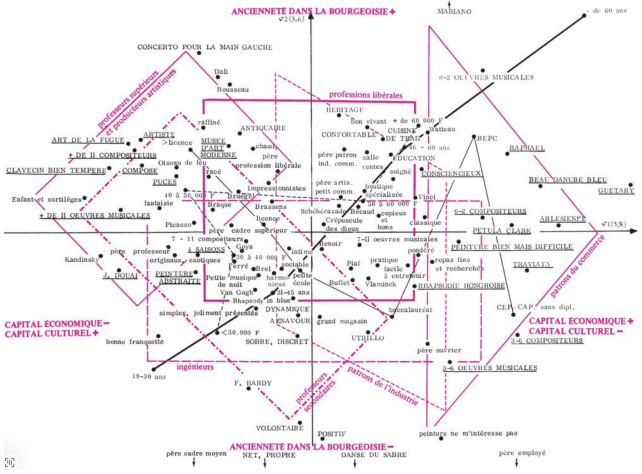


(Bourdieu, 1979)

Noyau et espace social

"Doing a data analysis, in good mathematics, is simply searching eigenvectors, all the science of it (the art) is just to find the right matrix to diagonalize"

(J.-P. Benzécri, 1973)

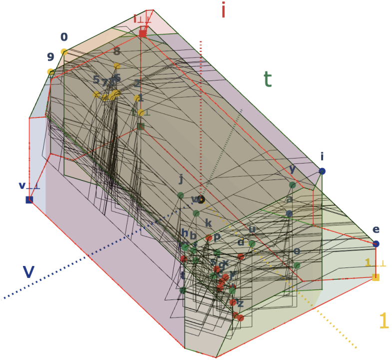


(Bourdieu, 1979)

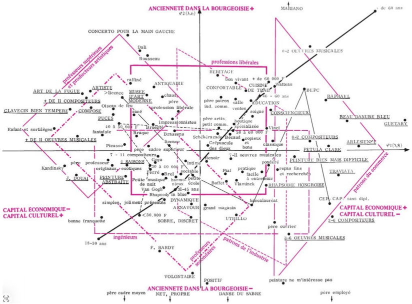
Interprétabilité théorique

Traits distinctifs
Unités
Classes
Relations
Structures

<u>Théorie</u>

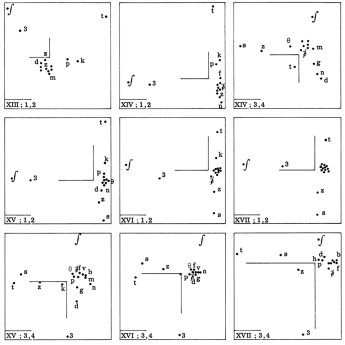


Linguistique
Sociologie
Anthropologie
Histoire
Géographie
Phil. des Sc.
...



(Bourdieu, 1979)

Sciences humaines et sciences formelles



v. p.	1°	2°	3°	4°	5°
XII	,99	,98	,97	,92	,86
XV	,71	,41	,37	,37	,19

v. p.	1°	2°	3°	4°	5°
XIII	,94	,92	,86	,83	,75
XVI	,55	,29	,25	,19	,04

v. p.	1°	2°	3°	4°	5°
XIV	,87	,62	,49	,47	,45
XVII	,44	,23	,17	,09	,02

(J. P. Benzécri, 1976)

1) une voyelle neutre (amorphe), caractérisée par l'absence de chacune des propriétés β, φ, χ, λ : [ə] ;

2) quatre types élémentaires de voyelles, chacun caractérisé par une seule propriété :

$$[\varepsilon] = \varphi\,,\ [\mathrm{a}] = \beta\,,\ [\mathrm{y}] = \chi\,,\ [\mathrm{ɯ}] = \lambda$$

3) six voyelles distinctes, chacune caractérisée par deux propriétés :

$$[\mathrm{ø}] = \varphi\beta\,,\ [\mathrm{i}] = \chi\varphi\,,\ [\mathrm{u}] = \chi\beta\,,\ [\mathrm{e}] = \lambda\varphi\,,\ [\mathrm{o}] = \lambda\beta\,,\ [\mathrm{ɔ}] = \chi\lambda\,;$$

4) quatre voyelles combinées, chacune caractérisée par trois propriétés :

$$[\mathrm{ɛ}] = \chi\lambda\varphi\,,\ [\mathrm{œ}] = \chi\lambda\beta\,,\ [\mathrm{ɯ}] = \varphi\beta\chi\,,\ [\mathrm{ɔ}] = \varphi\beta\lambda\,;$$

5) une voyelle polymorphe, caractérisée par les quatre propriétés considérées : la voyelle russe [ɯ] = $\varphi\beta\chi\lambda$.

On obtient alors le diagramme de la figure 2.

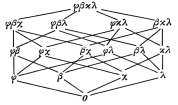
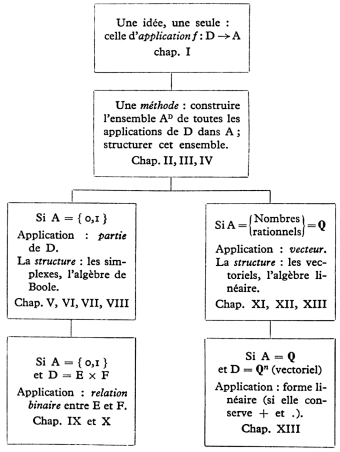


FIG. 2.

(Marcus, 1967)

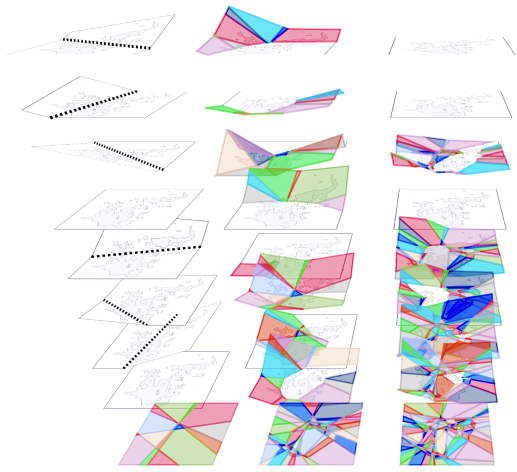
Organigramme



(Barbut, 1967)

Vers un structuralisme génératif

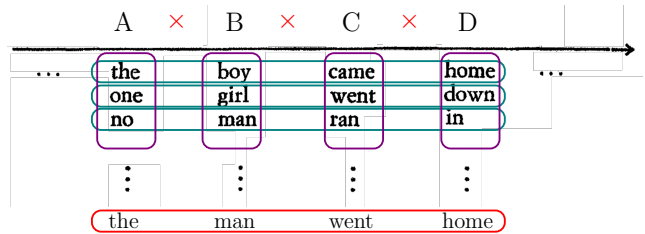
- Géométrie tropicale polyhédrique
- Réalisabilité et logique linéaire



(Welch Labs, 2025)

Vers un structuralisme génératif

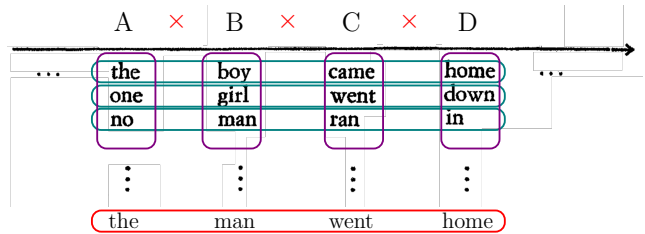
- Géométrie tropicale polyhédrique
- Réalisabilité et logique linéaire



(Hjelmslev, 1971a)

Vers un structuralisme génératif

- Géométrie tropicale polyhédrique
- Réalisabilité et logique linéaire

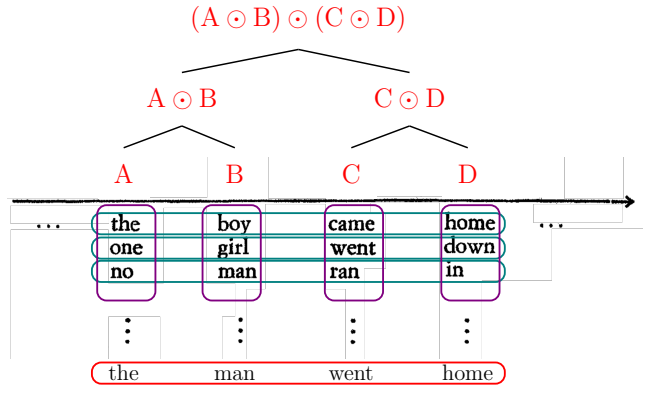


(Hjelmslev, 1971a)

$$(f \odot g)(x) := \sup_{cd=x} (f(c) + g(d) - M(c,d))$$

Vers un structuralisme génératif

- Géométrie tropicale polyhédrique
- Réalisabilité et logique linéaire



(Hjelmslev, 1971a)

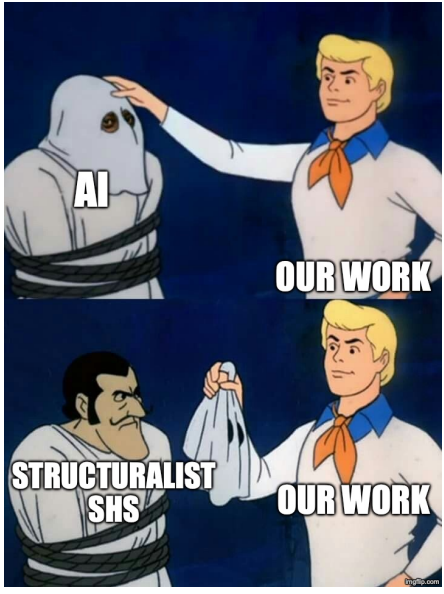
$$(f \odot g)(x) := \sup_{cd=x} (f(c) + g(d) - M(c,d))$$

Outline

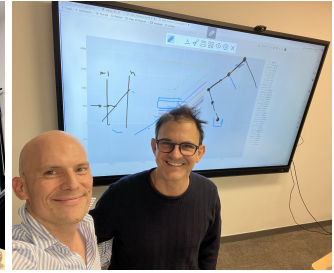
Conclusion


WORD EMBEDDINGS
LEVY& GOLDBERG
SVD
LEVY& GOLDBERG
imgflip.com

Conclusion



Collaborations



J. Terilla (CUNY), T.-D. Bradley (SandboxAQ), L. Pellissier (Paris-Est Créteil), Th. Seiller (CNRS), S. Jarvis (CUNY)

Reference Papers

- Gastaldi, J. L. (2021).Why Can Computers Understand Natural Language? *Philosophy & Technology*, *34*(1), 149–214
- Gastaldi, J. L., & Pellissier, L. (2021).The Calculus of Language: Explicit Representation of Emergent Linguistic Structure through Type-Theoretical Paradigms. *Interdisciplinary Science Reviews*, *46*(4), 569–590
- Bradley, T.-D., Gastaldi, J. L., & Terilla, J. (2024).The Structure of Meaning in Language: Parallel Narratives in Linear Algebra and Category Theory. *Notices of the American Mathematical Society*
- Gastaldi, J. L., Jarvis, S., Seiller, T., & Terilla, J. (2026c). Projective metric geometry of tropical nuclei: Gap matrices, event loci, and order chambers.
- Gastaldi, J. L., Jarvis, S., Seiller, T., & Terilla, J. (2026a). A calculus of types in isbell nuclei.

Références I

Awodey, S. (2010). *Category theory* (2nd). Oxford University Press, Inc.

Barbut, M. (1967). *Mathématiques et sciences humaines. tome i: Combinatoire et algèbre*. Presses Universitaires de France.

Benzécri, J. P. (1976).Sur le codage réduit d'un vecteur de description en analyse des correspondances. *Les cahiers de l'analyse des données*, *1*(2), 127–136.

Benzécri, J.-P. (1973). *L'analyse des données. 2 L'analyse des correspondances* [Et coll.]. Bordas.

Bourdieu, P. (1979). *La distinction: Critique sociale du jugement*. Éditions de Minuit.

Bradley, T.-D., Gastaldi, J. L., & Terilla, J. (2024).The Structure of Meaning in Language: Parallel Narratives in Linear Algebra and Category Theory. *Notices of the American Mathematical Society*.

Gastaldi, J. L. (2021).Why Can Computers Understand Natural Language? *Philosophy & Technology*, *34*(1), 149–214.

Gastaldi, J. L. (2024).Content from Expressions. The Place of Textuality in Deep Learning Approaches to Mathematics. *Synthese (under review)*.

Gastaldi, J. L., & Pellissier, L. (2021).The Calculus of Language: Explicit Representation of Emergent Linguistic Structure through Type-Theoretical Paradigms. *Interdisciplinary Science Reviews*, *46*(4), 569–590.

Gastaldi, J. L., Jarvis, S., Seiller, T., & Terilla, J. (2026a). A calculus of types in isbell nuclei.

Gastaldi, J. L., Jarvis, S., Seiller, T., & Terilla, J. (2026b). *Logical structures in $\mathbb{R}$-enriched adjunctions* [Working paper. Accessible at https://www.giannigastaldi.com/assets/pdf/pubs/GastaldiJarvisEtAl2025.pdf].

Gastaldi, J. L., Jarvis, S., Seiller, T., & Terilla, J. (2026c). Projective metric geometry of tropical nuclei: Gap matrices, event loci, and order chambers.

Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., Arora, A., Wu, Z., Goodman, N., Potts, C., & Icard, T. (2025). Causal abstraction: A theoretical foundation for mechanistic interpretability.

Girard, J.-Y. (2006). *Le point aveugle: Cours de logique. vers la perfection*. Editions Hermann.

Références II

Goldberg, Y., & Levy, O. (2014).Word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *CoRR*, *abs/1402.3722*.

Hjelmslev, L. (1971a). La structure fondamentale du langage. In *Prolégomènes à une théorie du langage* [Prolégomènes à une theorie du langage] (pp. 177–231). Éditions de Minuit.

Hjelmslev, L. (1971b). *Prolégomènes à une théorie du langage*. Éditions de Minuit.

Jakobson, R., Fant, G. M., & Halle, M. (1952). *Preliminaries to speech analysis: The distinctive features and their correlates*. MIT Press.

Levy, O., & Goldberg, Y. (2014).Neural word embedding as implicit matrix factorization. *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, 2177–2185.

Marcus, S. (1967). *Introduction mathématique à la linguistique structurale*. Dunod.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013).Distributed representations of words and phrases and their compositionality. *CoRR*, *abs/1310.4546*.

Seiller, T. (2024). *Mathematical informatics* [Habilitation thesis].

Sennrich, R., Haddow, B., & Birch, A. (2016).Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the ACL*, 1715–1725.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.

CAMS - EHESS
Séminaire Systèmes complexes en sciences sociales
2 Juin 2026 – Paris, France

Explicabilité formelle et interprétabilité théorique des modèles neuronaux

Juan Luis (Gianni) Gastaldi

ETH zürich

www.giannigastaldi.com