Concours chercheurs 2026
CR Section 53 - Concours n° 53/02

Formal Explainability and Theoretical Interpretability of Machine Learning Distributional Language Models

Juan Luis Gastaldi

ETH zürich

www.giannigastaldi.com

Parcours

1997-2007 Recherche pré-doctorale
Sciences Po, Philosophie, Mathématiques
Argentine, France
(UNR, ENS, Paris 1, UPMC)

2008-2014 Recherche doctorale
Philo et Hist des Sciences
France
(Bordeaux Montaigne)

2015-2022 Professeur d'Enseignement Artistique
Philo, Hist des Idées, Esthétique
France
(MO.CO.ESBA)

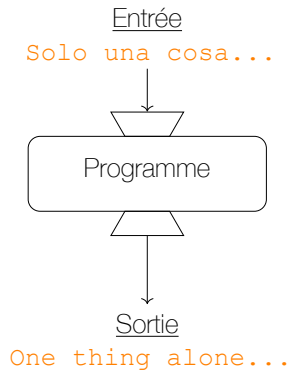
2015-2022 Recherche post-doctorale
Philo et Informatique
France, Suisse, Tchéquie, USA, Israel
(Paris 7, ETH, MSCA, CUNY, CMU, Cohn)

2023-Présent Chargé de recherche et de cours
Nouvelle recherche doctorale
Informatique (ML, TAL, IA)
Suisse, USA (ETH Zurich, CUNY)

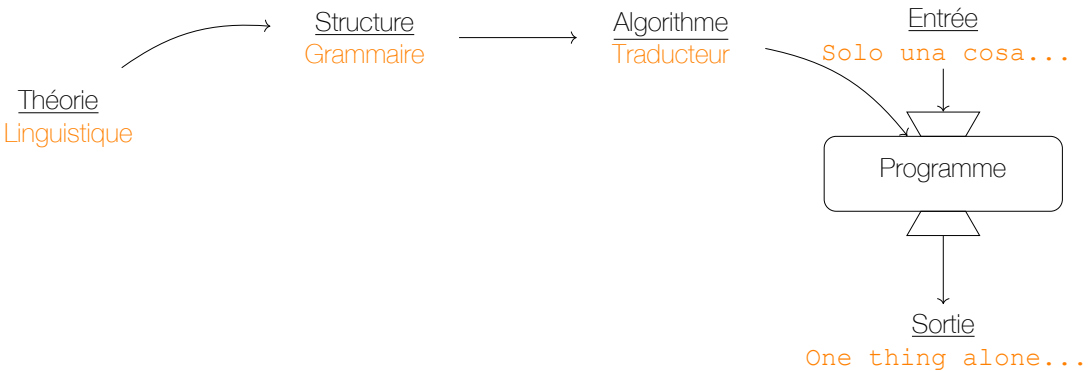
Recherche

- **Formalisme critique**: Articulation entre les humanités et les sciences formelles
- Approche **philosophique**, **historique**, **théorique** et **technique**
- Thèse en Philosophie: Philosophie et histoire de la **mathématisation de la logique** (Gastaldi, 2014)
- Thèse en Informatique: **Segmentation et structure** dans des **modèles distributionnels** de langage (Bradley et al., 2024, Gastaldi et al., 2025, Gastaldi et al., 2026b, Gastaldi et al., 2026a)

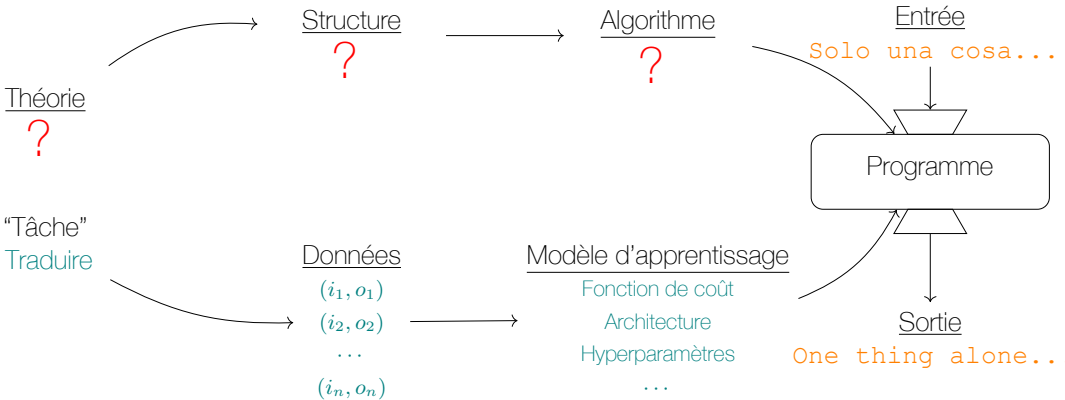
Motivation: La structure implicite des données



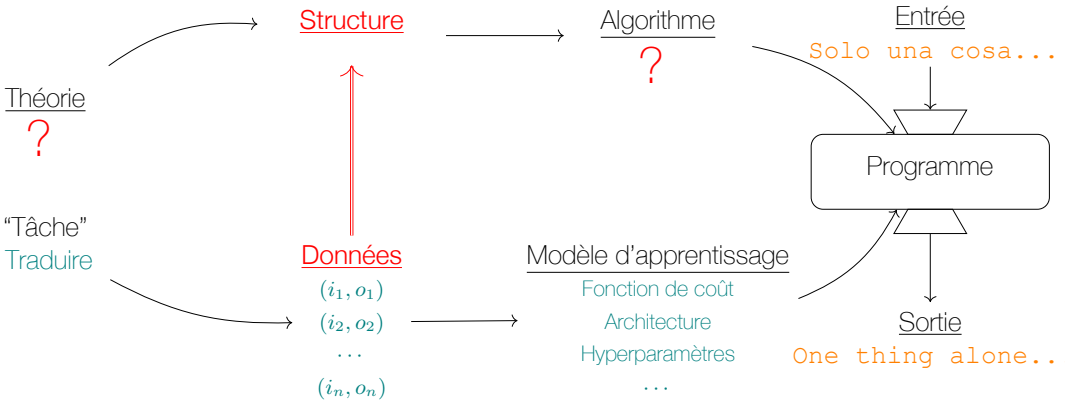
Motivation: La structure implicite des données



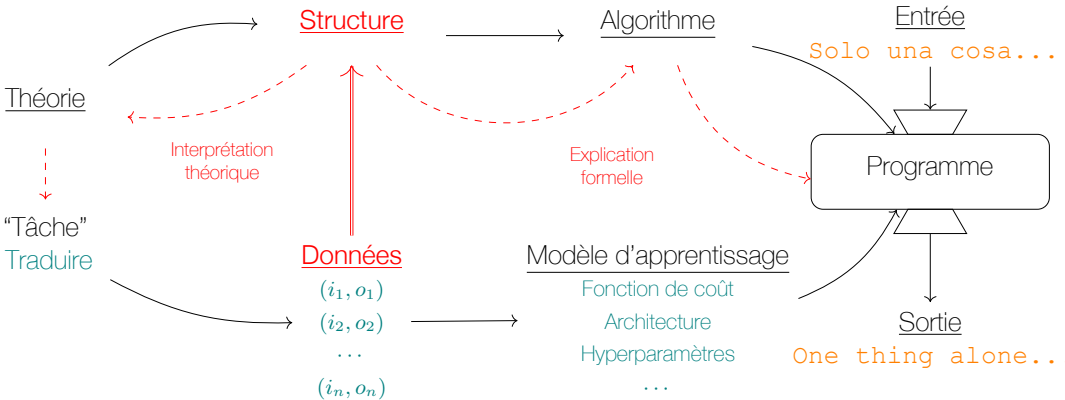
Motivation: La structure implicite des données



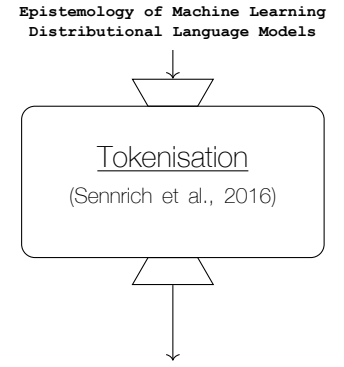
Motivation: La structure implicite des données



Motivation: La structure implicite des données

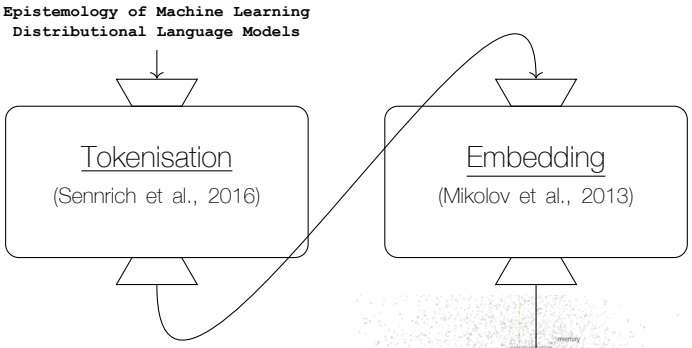


Axe 1: Explicabilité formelle



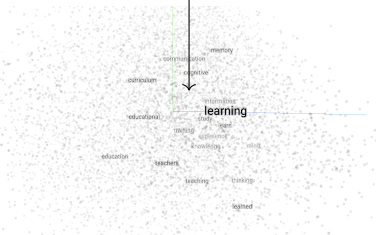
(https://tiktokenizer.vercel.app)

Axe 1: Explicabilité formelle



Epistemology of Machine Learning
Distributional Language Models

(https://tiktokenizer.vercel.app)



(https://projector.tensorflow.org)

Axe 1: Explicabilité formelle

Epistemology of Machine Learning
Distributional Language Models

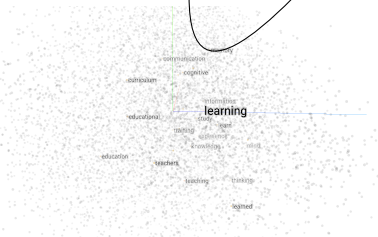
Tokenisation
(Sennrich et al., 2016)

Embedding
(Mikolov et al., 2013)

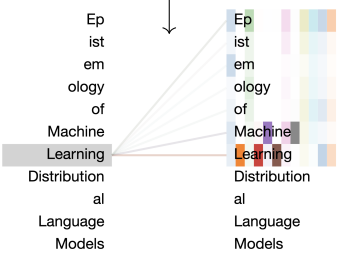
Attention
(Vaswani et al., 2017)

Epistemology of Machine Learning
Distributional Language Models

(https://tiktokenizer.vercel.app)

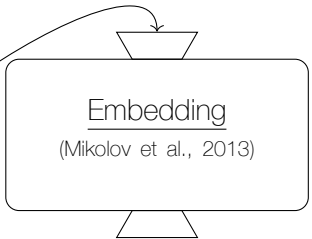


(https://projector.tensorflow.org)



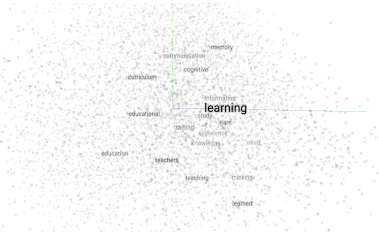
(https://github.com/jessevig/bertviz)

Axe 1: Explicabilité formelle



Epistemology of Machine Learning
Distributional Language Models

(https://tiktokenizer.vercel.app)



(https://projector.tensorflow.org)

Axe 1: La structure des embeddings

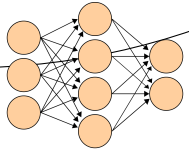
Structure

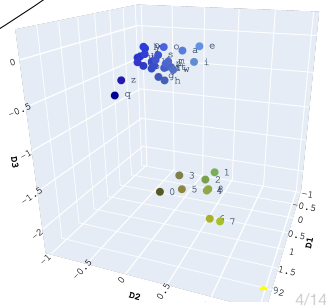
?

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

Embedding

Données





Axe 1: La structure des embeddings

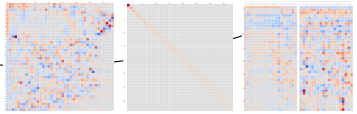
Structure

?

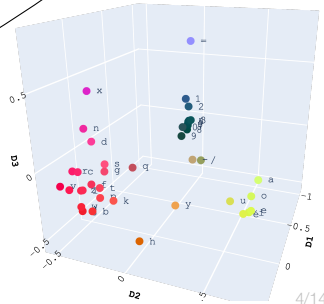
{-, /, 0, 1, 2, . . . , 8, 9, =,
a, b, c, . . . , w, x, y, z, é}

Embedding

Données



SVD



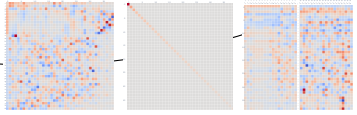
Axe 1: La structure des embeddings

Structure

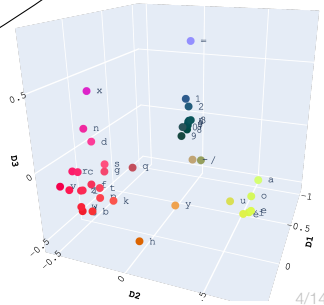
{-, /, 0, 1, 2, . . . , 8, 9, =,
a, b, c, . . . , w, x, y, z, é}

Embedding

Données



SVD



Axe 1: La structure des embeddings

Structure

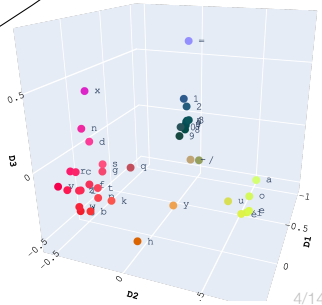
?

Données

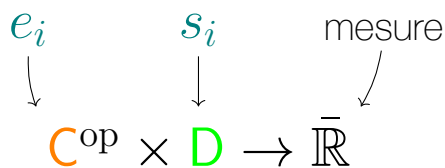
$$C^{op} \times D \to \bar{\mathbb{R}}$$

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

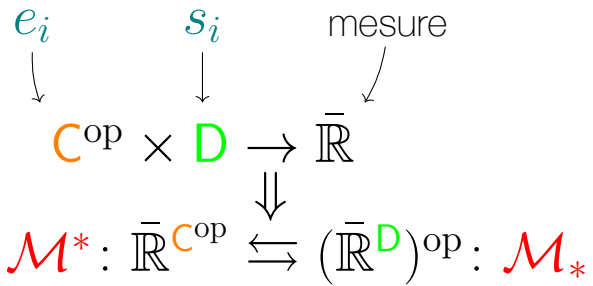
Embedding



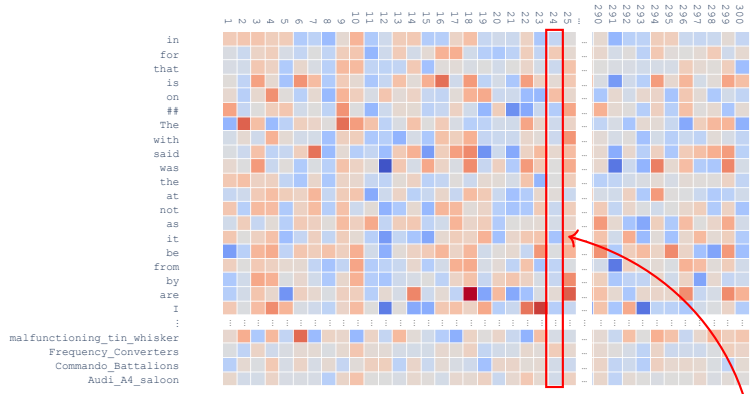
Axe 1: Profoncteur et structures du noyau



Axe 1: Profoncteur et structures du noyau



Axe 1: Profoncteur et structures du noyau



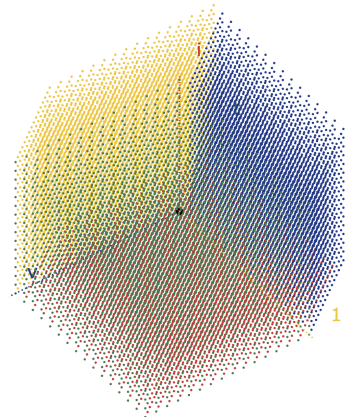
$$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \bar{\mathbb{R}}$$

$$\Downarrow$$

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

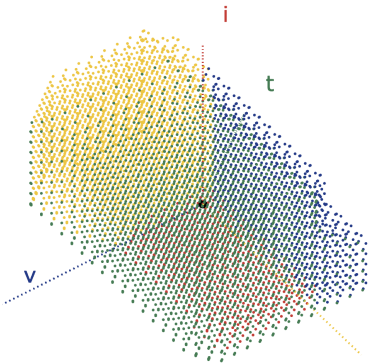
point fixe

Axe 1: Profoncteur et structures du noyau

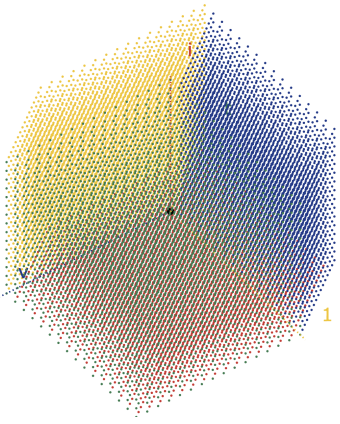


$$\begin{array}{ccc} \mathsf{C}^{\mathrm{op}} \times \mathsf{D} & \to & \bar{\mathbb{R}} \\ & \Downarrow & \\ \mathcal{M}^* \colon \bar{\mathbb{R}}^{\mathsf{C}^{\mathrm{op}}} & \leftrightarrows & \left(\bar{\mathbb{R}}^{\mathsf{D}}\right)^{\mathrm{op}} \colon \mathcal{M}_* \end{array}$$

Axe 1: Profoncteur et structures du noyau

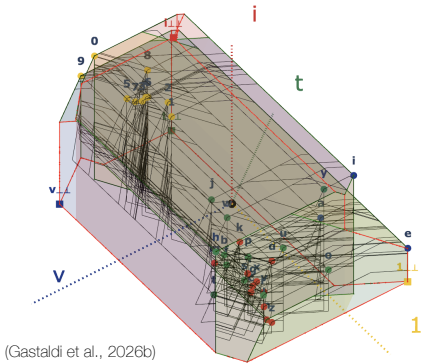


$\mathcal{M}_* \mathcal{M}^*$



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

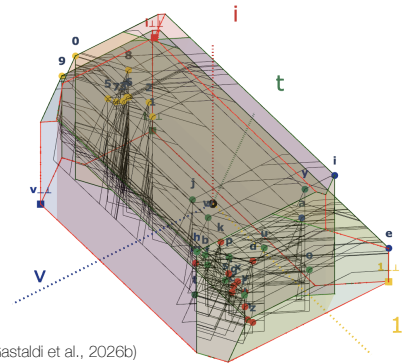
Axe 1: Profoncteur et structures du noyau



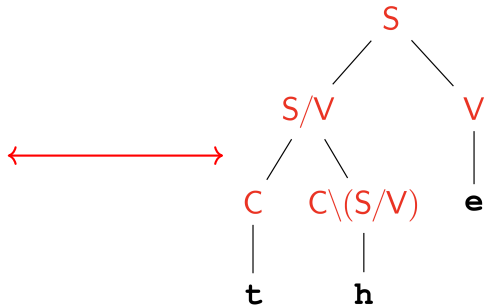
(Gastaldi et al., 2026b)

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

Axe 1: Profoncteur et structures du noyau



(Gastaldi et al., 2026b)



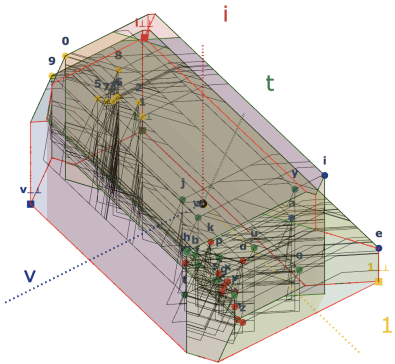
(Gastaldi & Pellissier, 2021)
(Gastaldi et al., 2026a)

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

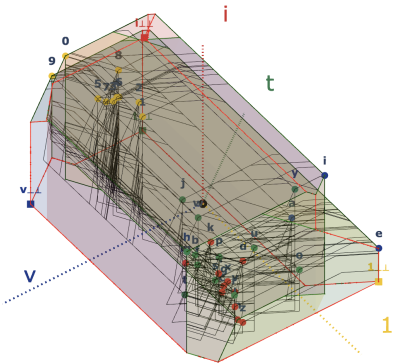
$$\Downarrow$$

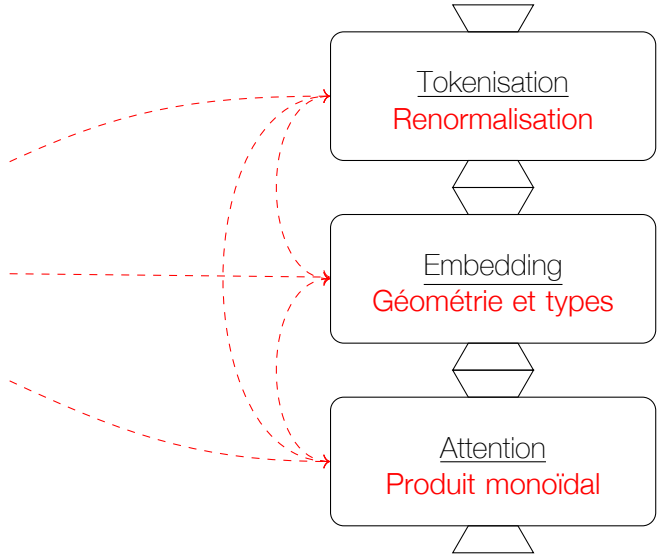
$$\mathcal{M}^* : \bar{\mathbb{R}}^{\mathsf{C}^{\mathrm{op}}} \leftrightarrows \left(\bar{\mathbb{R}}^{\mathsf{D}}\right)^{\mathrm{op}} : \mathcal{M}_*$$

Axe 1: Explicabilité formelle



Axe 1: Explicabilité formelle

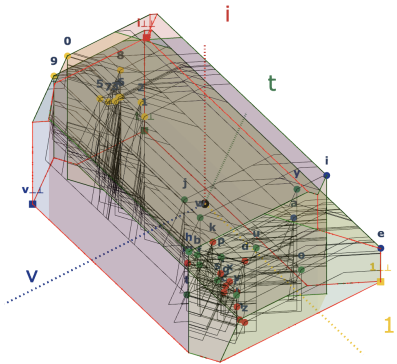




Axe 2: Interprétabilité théorique

Théorie

?



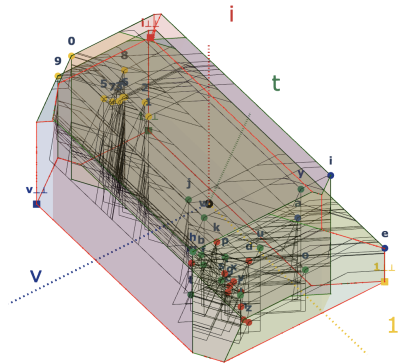
Axe 2: Interprétabilité théorique

Hypothèse distributionnelle

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

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

Théorie



Axe 2: Interprétabilité théorique

Hypothèse distributionnelle

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

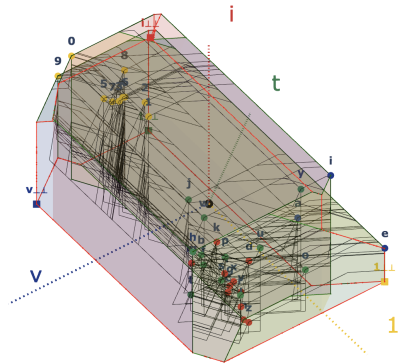
$$\mathbf{C}^{\mathrm{op}} \times \mathbf{D} \rightarrow \bar{\mathbb{R}}$$

Théorie

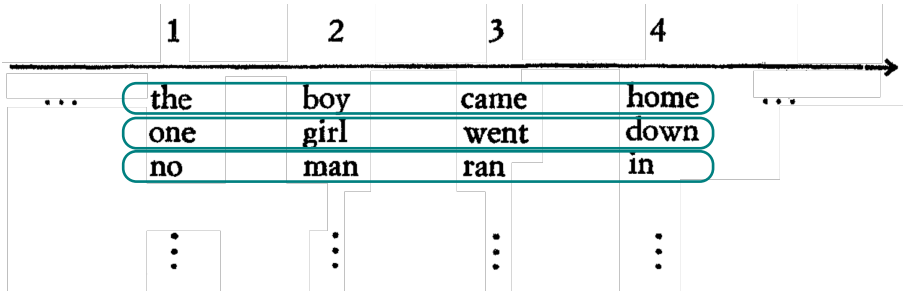
$$\bar{\mathbb{R}}^{\mathbf{C}^{\mathrm{op}}} \leftrightarrows (\bar{\mathbb{R}}^{\mathbf{D}})^{\mathrm{op}}$$

Hypothèse structurale

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

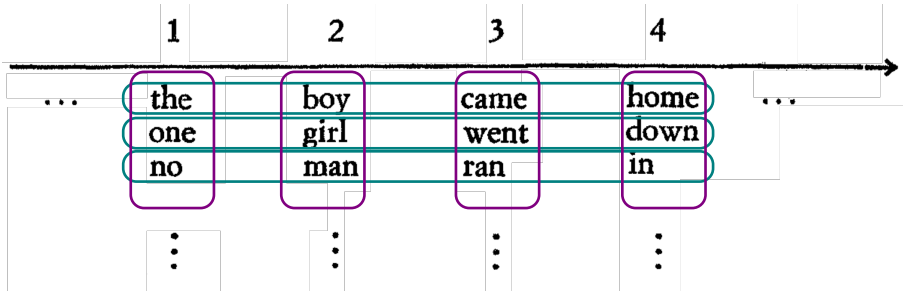


Axe 2: Interprétabilité théorique



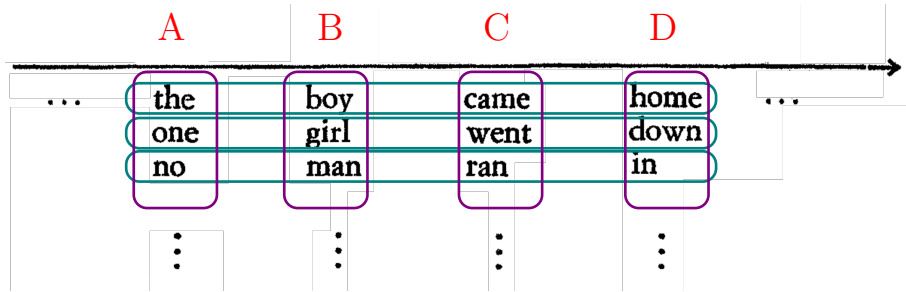
(Hjelmslev, 1971)

Axe 2: Interprétabilité théorique



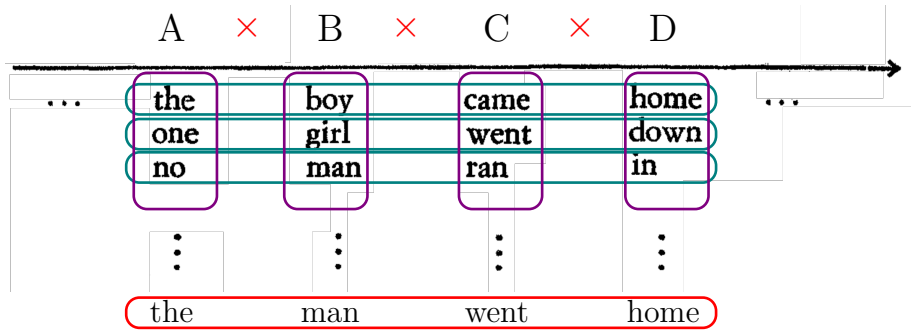
(Hjelmslev, 1971)

Axe 2: Interprétabilité théorique



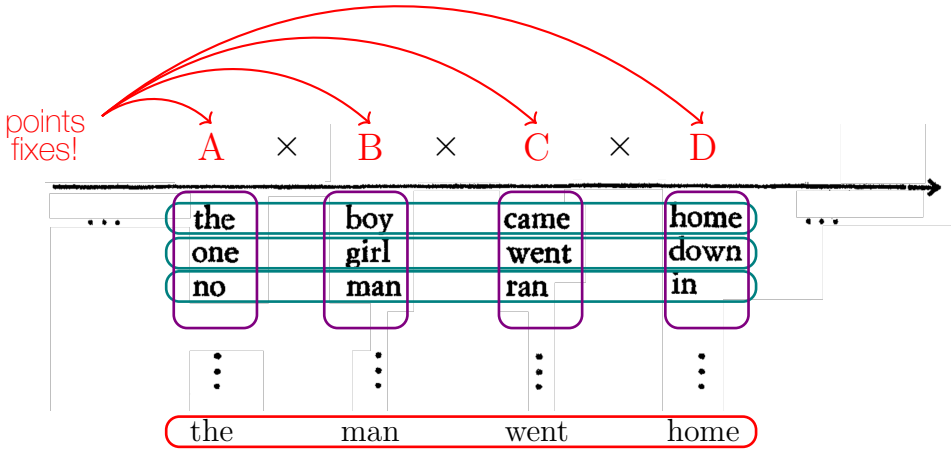
(Hjelmslev, 1971)

Axe 2: Interprétabilité théorique



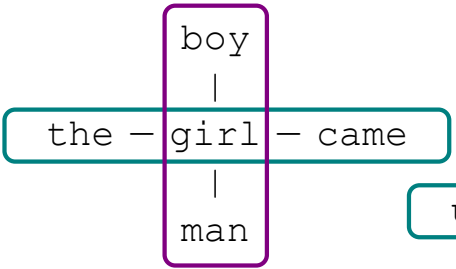
(Hjelmslev, 1971)

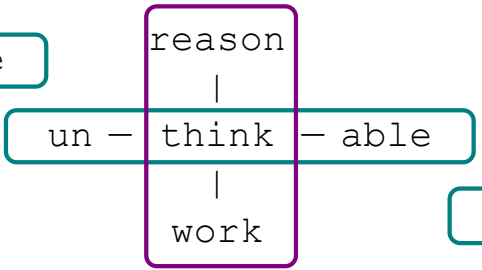
Axe 2: Interprétabilité théorique

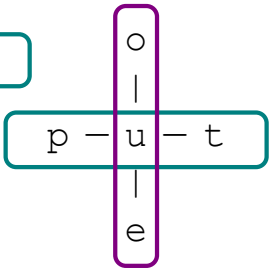


(Hjelmslev, 1971)

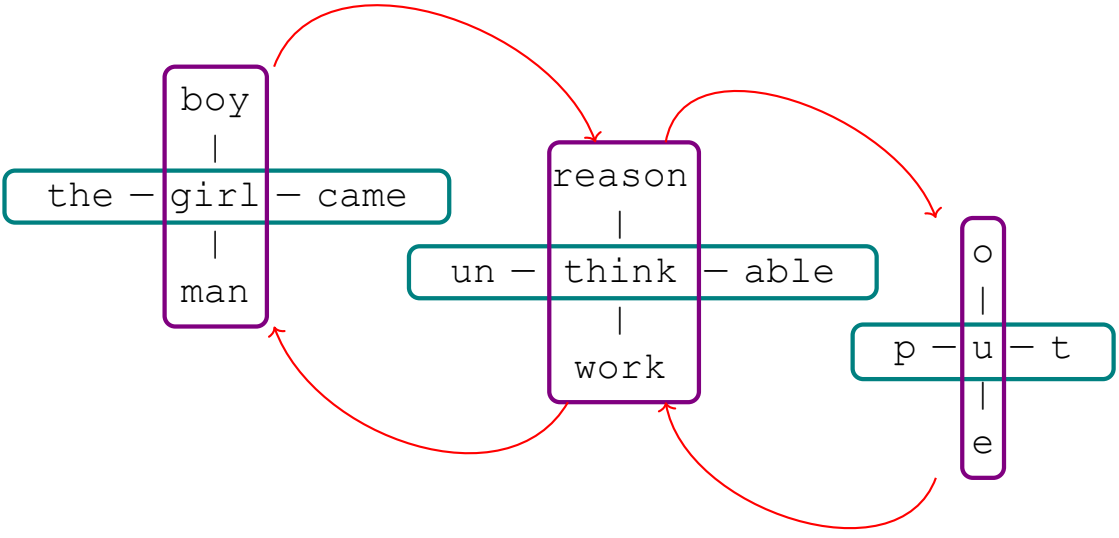
Axe 2: Interprétabilité théorique







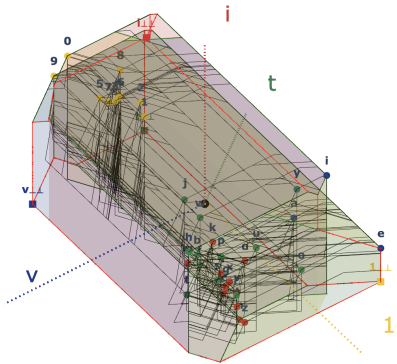
Axe 2: Interprétabilité théorique



Axe 2: Interprétabilité théorique

Traits distinctifs
Unités
Classes
Relations
Structures

Théorie

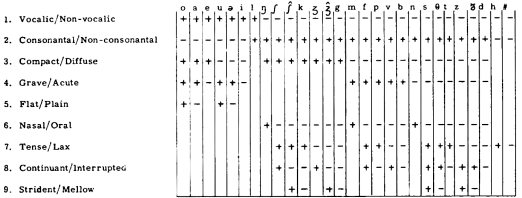


Axe 2: Interprétabilité théorique

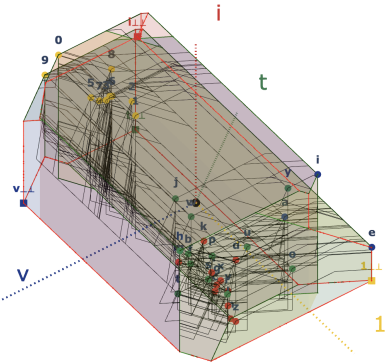
Traits distinctifs
Unités
Classes
Relations
Structures

Théorie

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



(Jakobson et al., 1952)

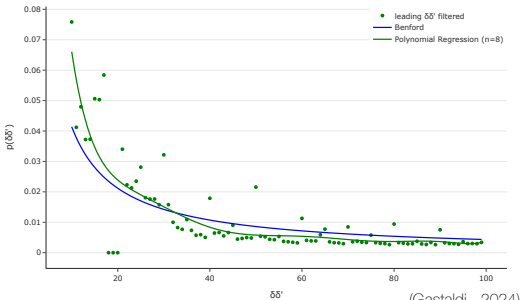


Axe 2: Interprétabilité théorique

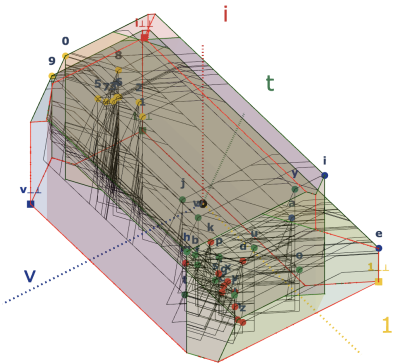
Traits distinctifs
Unités
Classes
Relations
Structures

Théorie

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



(Gastaldi, 2024)

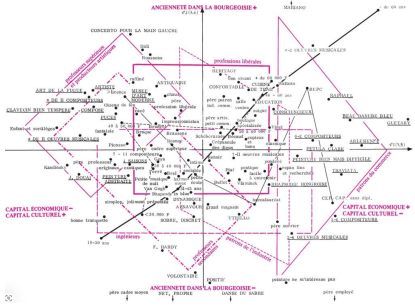


Axe 2: Interprétabilité théorique

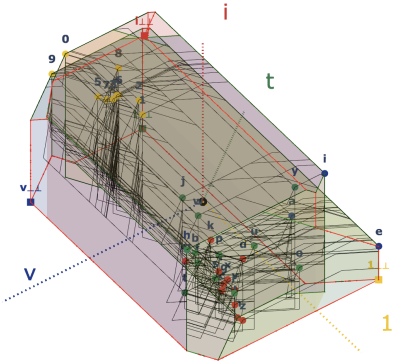
Traits distinctifs
Unités
Classes
Relations
Structures

Théorie

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



(Bourdieu, 1979)



CAMS · IHPST · SPHERE · STL

Intégration scientifique

- Inscription dans des **axes** combinant **philosophie**, **sc. formelles** et **sc. humaines**
- Environnement favorable à l'**interdisciplinarité**
- Apport d'une **expertise en informatique** rare dans des environnements SHS

Intégration effective

- Contact établi avec les **directions**
- Accueil accepté par les **conseils de laboratoire**
- **Lettres de soutien** jointes au dossier
- **Rencontres avec des collègues** dans chaque laboratoire
- **Présentation de mes travaux** à plusieurs occasions

Intégration - Détails

CAMS, UMR 8557 (EHESS)

- Unité **pluridisciplinaire** CNRS-EHESS (SHS & Mathématiques)
- Axe **cognitions** individuelles et collectives (IA connexionniste, modélisation en SHS)
- Axe **outils fondamentaux** (sciences formelles pour les SHS)
- Axe **morphogenèse** (méthodes symboliques et connexionnistes)

SPHERE, UMR 7219 (Paris Cité)

- Axe **Histoire & philosophie des mathématiques**
- Axe **Interdisciplinarité** en HPS
- Articulation entre approches **épistémologiques, historiques, linguistiques et sociologiques**
- Liens avec labos d'informatique et linguistique (**IRIF**, **LLF**)

IHPST, UMR 8590 (Panthéon-Sorbonne)

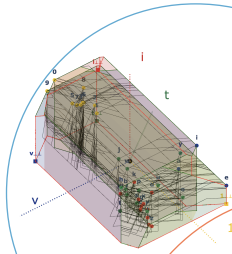
- Axe **Logique, mathématiques, informatique** (algorithmes, méthodes formelles, philosophie de l'IA)
- Axe transversal **Sciences en société** (impact social de l'IA)
- Écosystème **AISorb**
- Profil **interdisciplinaire** valorisé (philosophie & histoire des sc. formelles, informatique)

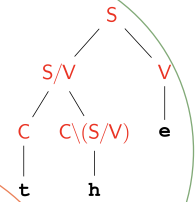
STL, UMR 8163 (Lille)

- Croisement **Philosophie et Linguistique**
- Thématique **Construction du sens en contexte** (Langues et langage)
- Thématique **Frontières de la science** (Normes, œuvres, discours)
- Groupe de travail sur l'**IA** actif

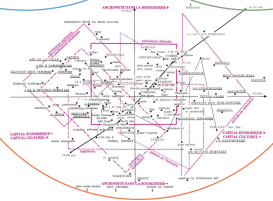
Programme

Mathématiques
Catégories enrichies
Géométrie tropicale
…





SGT



Logique informatique
Théorie des types
Réalisabilité linéaire
…

Humanités
Épistémologie
Linguistique
Sociologie
…

Profil

Compétences interdisciplinaires

- **Doctorats** en *Philo* et en *Info* (en cours)
- **Publications** en *Humanités* (Phil&Tech, Minds and Machines, Synthese, ...) et *Info et Maths* (ACL, EMNLP, ICLR, ICML, AMS, ...)
- **Conférences invitées**

2025: 9 conférences (Montréal, NYU, CUNY, Cambridge, Ca' Foscari, Côte d'Azur, Dagstuhl, ...)
2026: 7 conférences (ITU, Caltech, Côte d'Azur, Paris 8, EHESS, Monash Malaisie, Bâle, ...)

Participation à la communauté scientifique

- **Collaborations internationales** actives et interdisciplinaires (ETH, CUNY, CNRS, ITU Copenhague, U. Milan, U. Bâle, ...)
- **Évaluateur** (Horizon Europe, FNRS, ANR)
- **Reviewer** (Nature SR, Phil&Tech, Minds and Machines, HSSC, ACL, ICLR, NeurIPS, Compositionality, ...)
- **(Vice)-Président** de HaPoC (jusqu'en 2025)

Enseignement et encadrement

- **Enseignement** interdisciplinaire international (Argentine, France, Tchéquie, Suisse)
- **Encadrement** d'étudiants en Philosophie, Informatique, Art (L, M, D)

Gestion de la recherche

- **Directeur** du Dép. de Recherche (MO.CO.ESBA)
- **Directeur exécutif** du Turing Center (ETH)
- **Marie Skłodowska-Curie** Fellow

Références I

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.

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. (2014, September). *Une archéologie de la logique du sens : arithmétique et contenu dans le processus de mathématisation de la logique au XIXe siècle* (Publication No. 2014BOR30035) [Theses]. Université Michel de Montaigne - Bordeaux III.

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., Terilla, J., Malagutti, L., DuSell, B., Vieira, T., & Cotterell, R. (2025).The Foundations of Tokenization: Statistical and Computational Concerns. *The Thirteenth International Conference on Learning Representations (ICLR)*.

Gastaldi, J. L., Jarvis, S., Seiller, T., & Terilla, J. (2026a). *A Calculus of Types in Isbell Nuclei* [Under review. Accessible at https://www.giannigastaldi.com/assets/pdf/pubs/GastaldiJarvisEtAl2025.pdf].

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

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

Références II

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

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.

Concours chercheurs 2026
CR Section 53 - Concours n° 53/02

Formal Explainability and Theoretical Interpretability of Machine Learning Distributional Language Models

Juan Luis Gastaldi

ETH zürich

www.giannigastaldi.com