CNRS - Concours chercheurs 2025
CR Section 53 - Concours n° 53/03

Épistémologie des modèles distributionnels de langage par apprentissage machine

Explicabilité formelle et interprétabilité théorique

Juan Luis Gastaldi

`http://www.jlgastaldi.com/assets/gastaldi_cnrs_cr.pdf`

ETH zürich

12 Mai 2025

Parcours

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

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

2015-2022 Professeur d'Ens. Artistique
Philo et Hist des Idées
France
(MO.CO.ESBA)

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

2023-Présent Nouvelle recherche doctorale
Informatique (ML, TAL)
Suisse, USA
(ETH Zurich)

Parcours

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

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

2015-2022 Professeur d'Ens. Artistique
Philo et Hist des Idées
France
(MO.CO.ESBA)

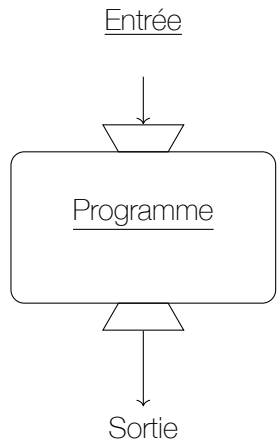
2015-2022 Recherche post-doctorale
Philo et Informatique
France, Suisse, Tchéquie, USA
(ETH, MSCA, CUNY, CMU)

2023-Présent Nouvelle recherche doctorale
Informatique (ML, TAL)
Suisse, USA
(ETH Zurich)

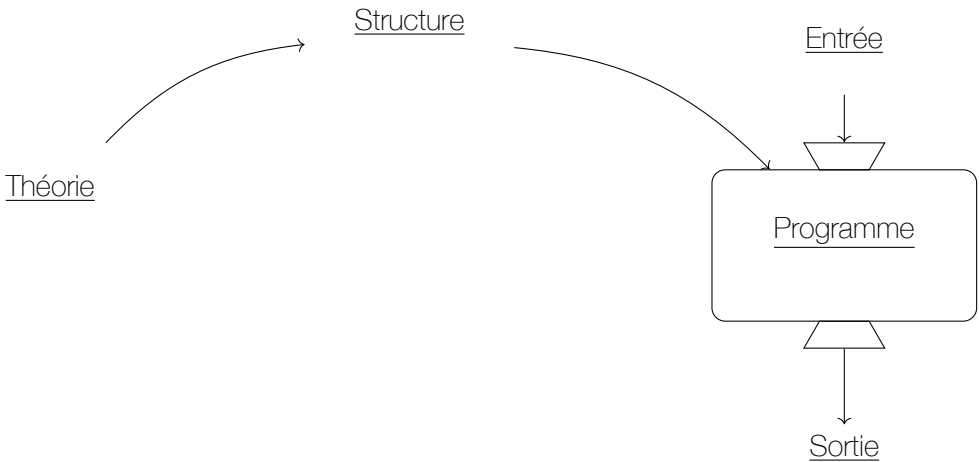
Recherche

- **Formalisme critique**: Alliance entre les humanités critiques 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: **Tokenisation en TAL** (Gastaldi et al., 2024; Giulianelli et al., 2024; Vieira et al., 2024; Zouhar et al., 2023a, 2023b)

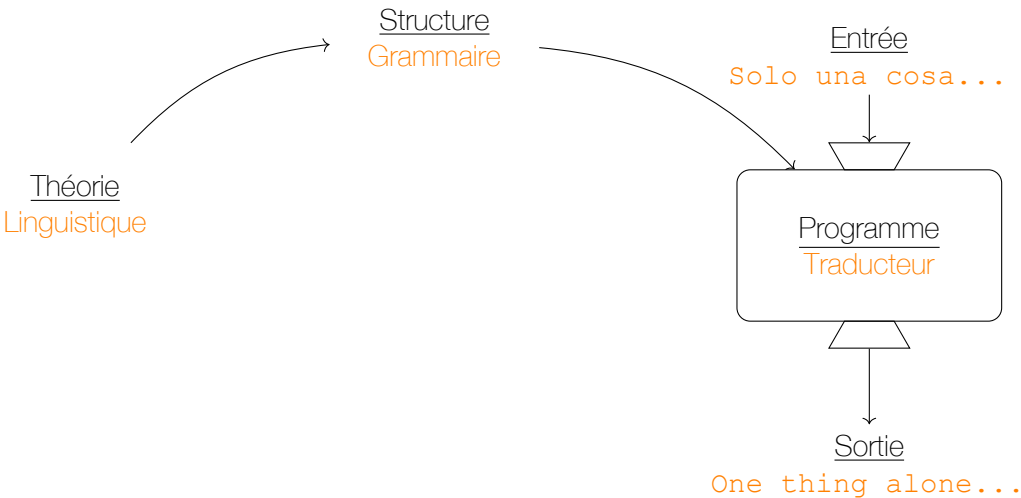
La structure implicite des données



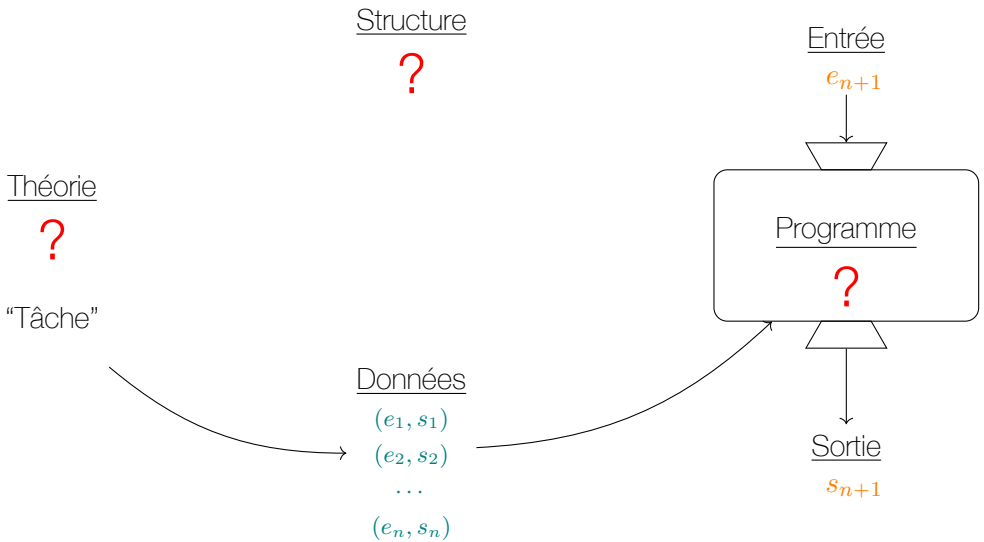
La structure implicite des données



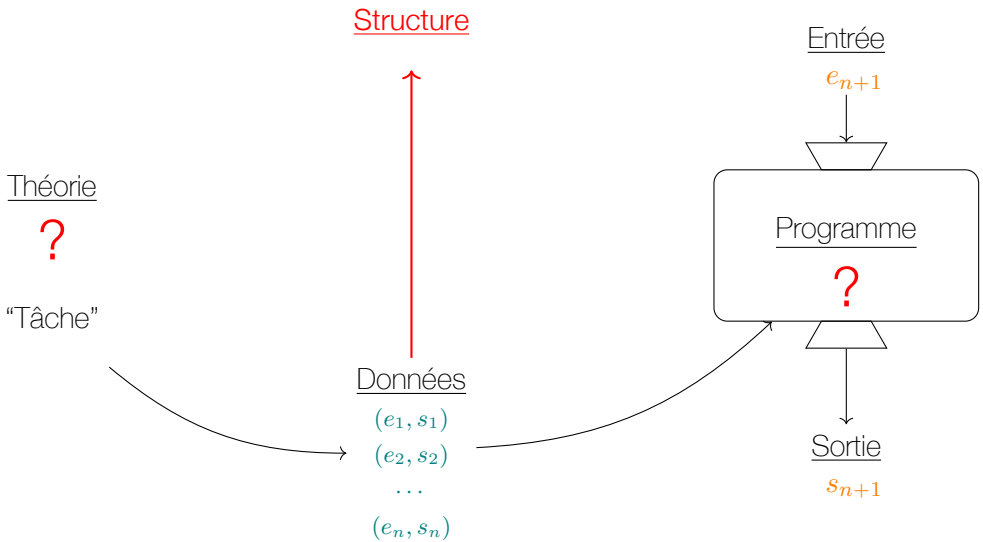
La structure implicite des données



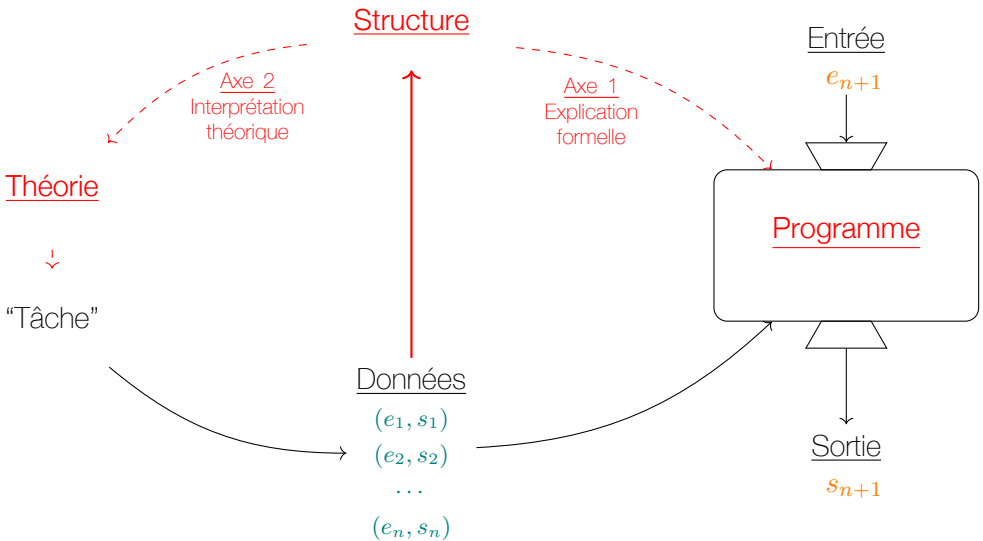
La structure implicite des données



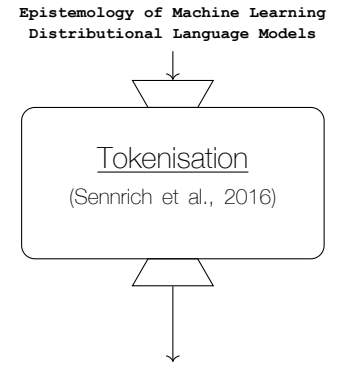
La structure implicite des données



La structure implicite des données

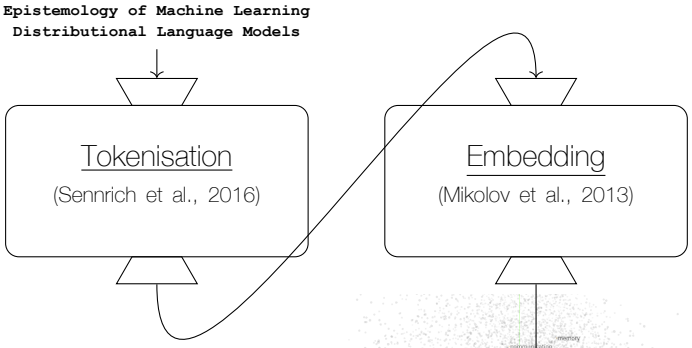


Axe 1: Explicabilité formelle



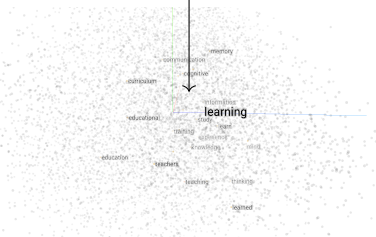
(https://tiktokenizer.vercel.app)

Axe 1: Explicabilité formelle



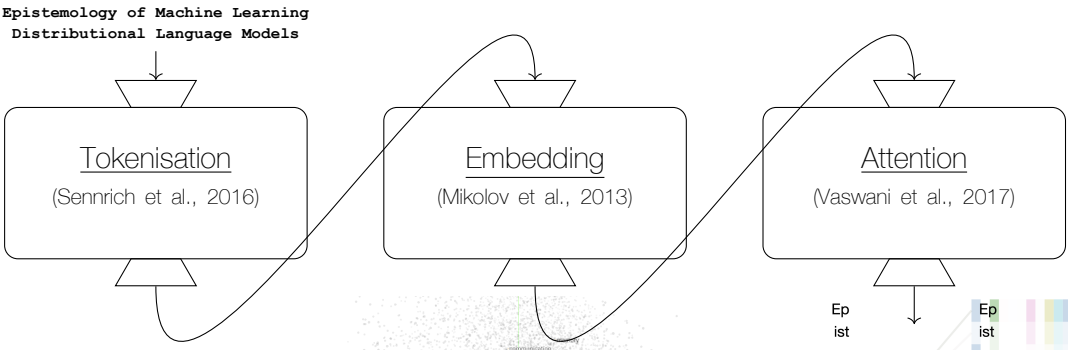
Epistemology of Machine Learning
Distributional Language Models

(https://tiktokenizer.vercel.app)

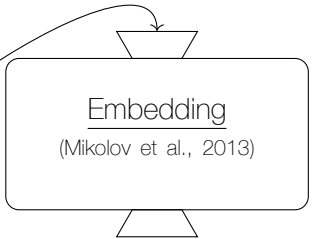


(https://projector.tensorflow.org)

Axe 1: Explicabilité formelle

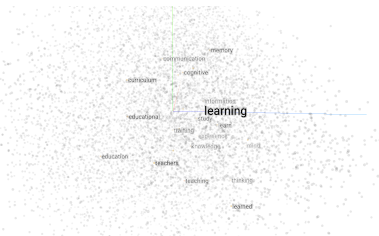


Axe 1: Explicabilité formelle



Epistemology of Machine Learning
Distributional Language Models

(https://tiktokenizer.vercel.app)



(https://projector.tensorflow.org)

La structure des embeddings

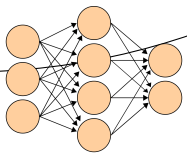
Structure

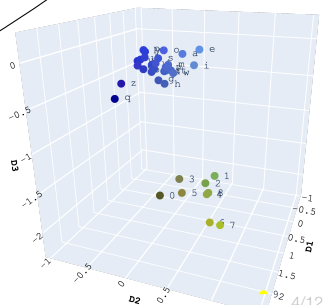
?

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

Embedding

Données





La structure des embeddings

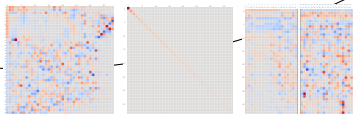
Structure

?

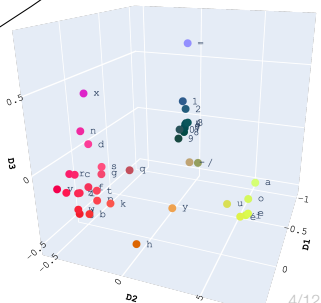
{-, /, 0, 1, 2, . . . , 8, 9, =,
a, b, c, . . . , w, x, y, z, é}

Embedding

Données

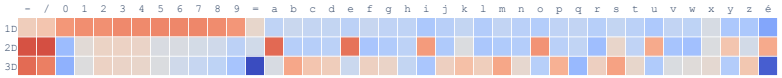


SVD



La structure des embeddings

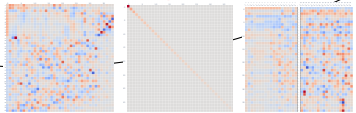
Structure



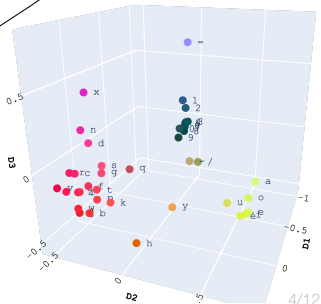
{-, /, 0, 1, 2, . . . , 8, 9, =,
a, b, c, . . . , w, x, y, z, é}

Embedding

Données



SVD



La structure des embeddings

Structure

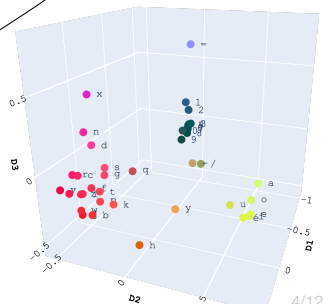
?

Données

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

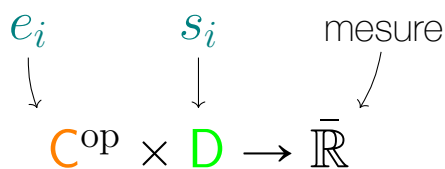
{-, /, 0, 1, 2, . . . , 8, 9, =,
a, b, c, . . . , w, x, y, z, é}

Embedding



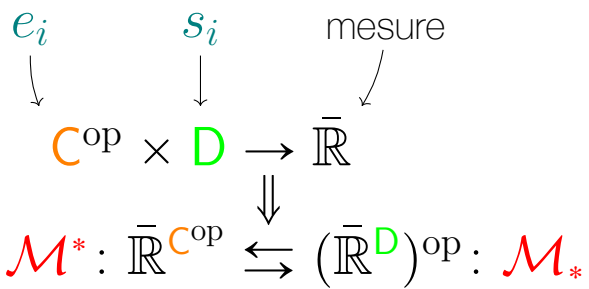
Structure

?

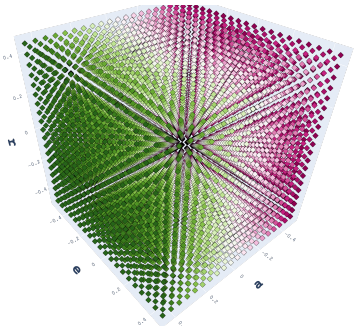


Structure

?


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

Structure

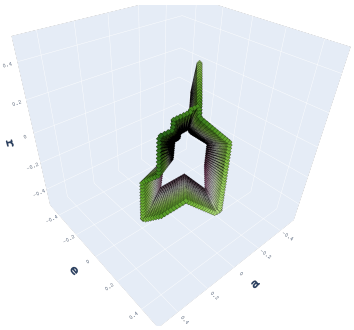


$$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \longrightarrow \bar{\mathbb{R}}$$

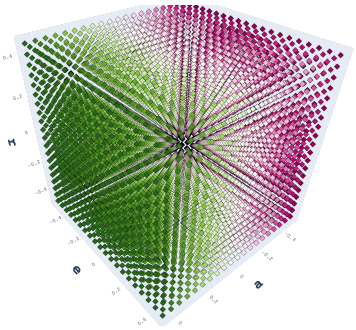
$$\Downarrow$$

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

Structure



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

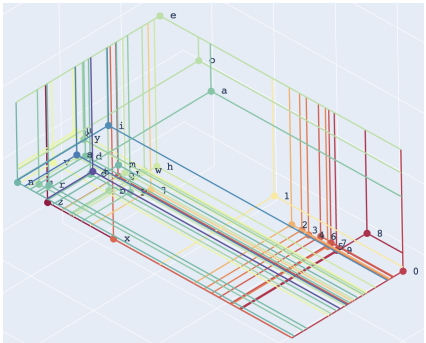


$$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \longrightarrow \bar{\mathbb{R}}$$

$$\Downarrow$$

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

Structure

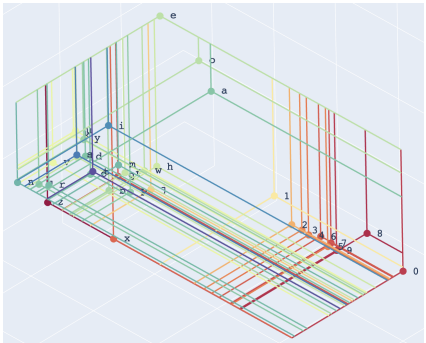


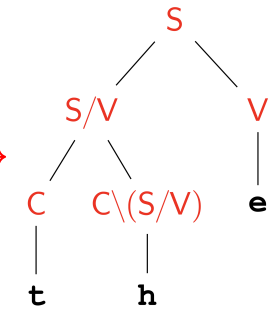
$$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \longrightarrow \bar{\mathbb{R}}$$

$$\Downarrow$$

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

Structure



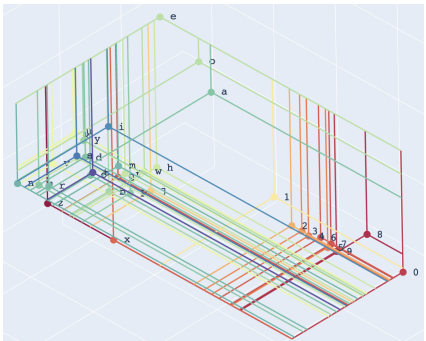


$$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \longrightarrow \bar{\mathbb{R}}$$

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

Axe 1: Objectifs

Structure

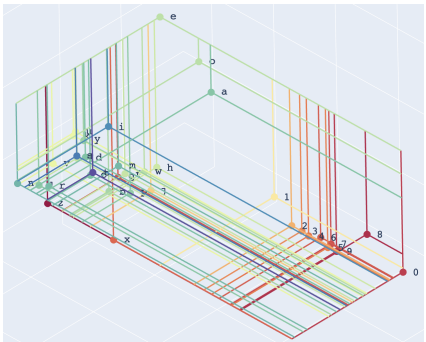


Embedding

Noyau et Types

Axe 1: Objectifs

Structure



Embedding

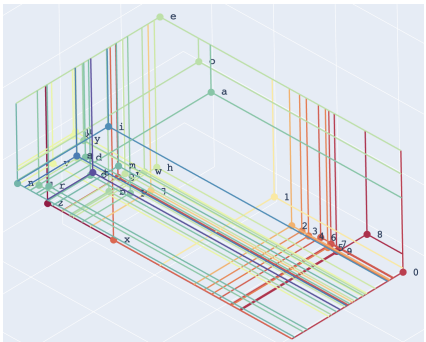
Noyau et Types

Attention

Compositionnalité

Axe 1: Objectifs

Structure



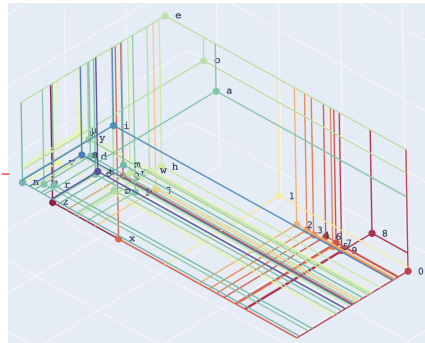
Tokenisation
Mesure et entropie

Embedding
Noyau et Types

Attention
Compositionnalité

Axe 2: Interprétabilité théorique

Structure



Théorie
"Tâche"

?

Axe 2: Interprétabilité théorique

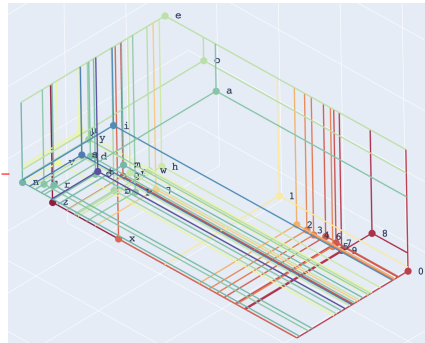
$$C^{\mathrm{op}} \times D \longrightarrow \bar{\mathbb{R}}$$

Hypothèse distributionnelle

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

Théorie
"Tâche"

Structure



Axe 2: Interprétabilité théorique

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

Hypothèse distributionnelle

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

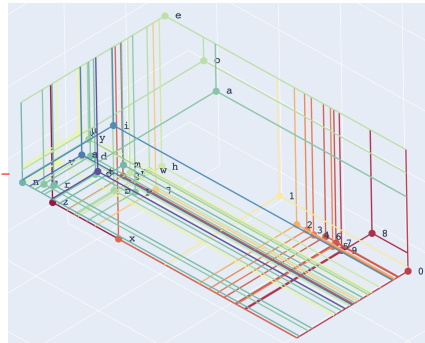
Théorie
"Tâche"

Hypothèse structurale

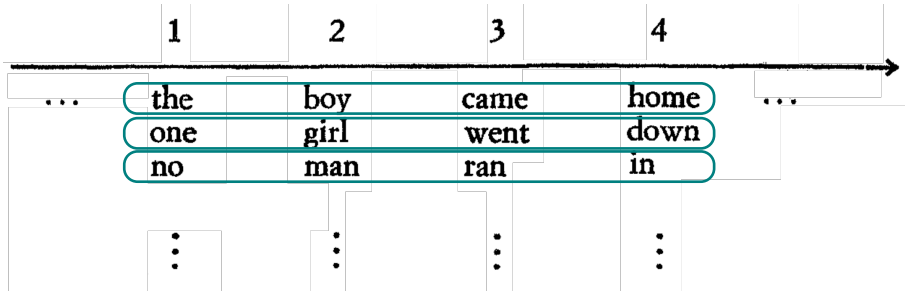
Le contenu linguistique est l'effet d'une **structure** virtuelle dérivée des pratiques linguistiques dans une communauté.

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

Structure

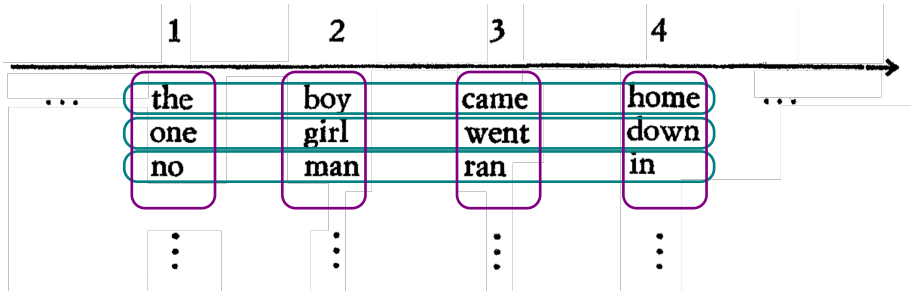


Syntagmes et Paradigmes



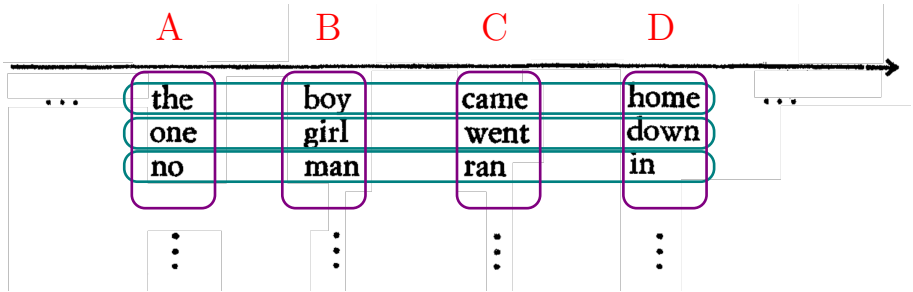
(Hjelmslev, 1971)

Syntagmes et Paradigmes



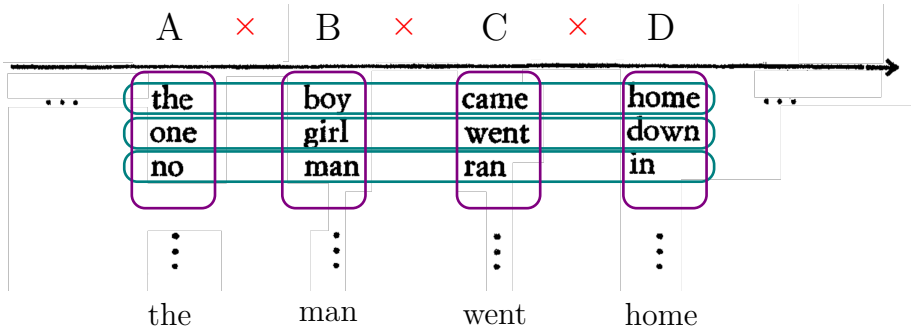
(Hjelmslev, 1971)

Syntagmes et Paradigmes



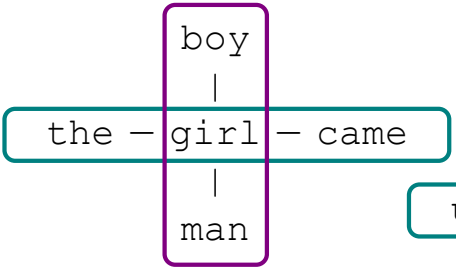
(Hjelmslev, 1971)

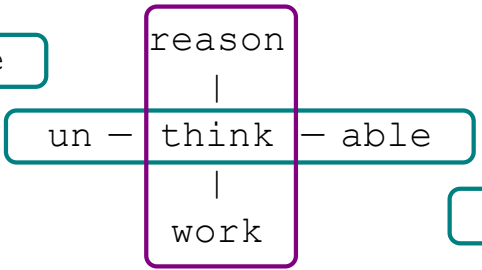
Syntagmes et Paradigmes

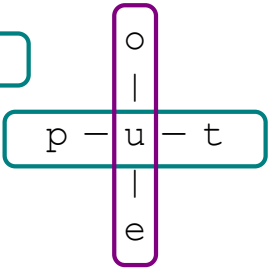


(Hjelmslev, 1971)

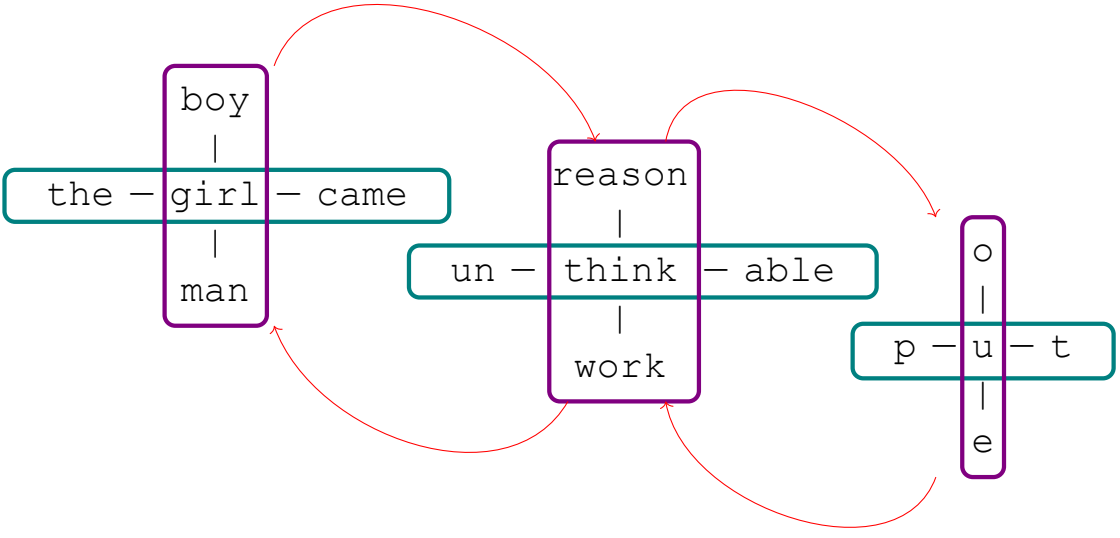
Stratification







Stratification

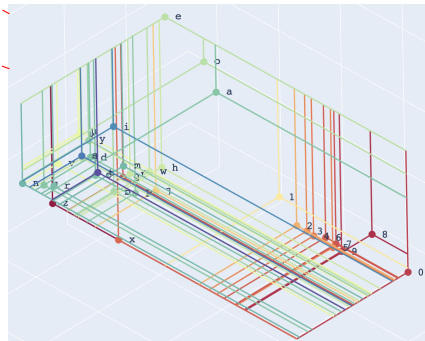


Axe 2: Objectifs

Unités
Classes
Relations

Sémantique
Syntaxe
Morphologie
Phonologie

Structure

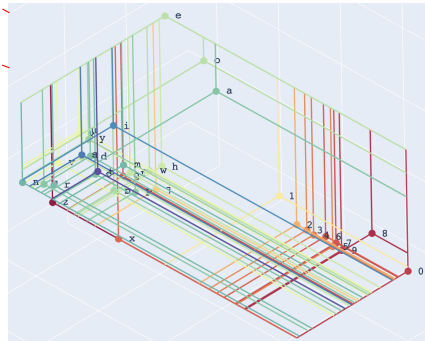


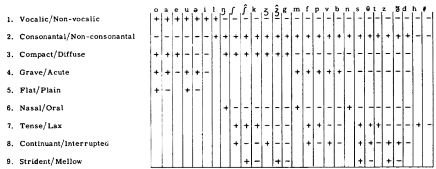
Axe 2: Objectifs

Unités
Classes
Relations

Sémantique
Syntaxe
Morphologie
Phonologie

Structure





(Jakobson et al., 1952)

Profil

Compétences interdisciplinaires

- **Doctorats** en Philo et en Info (en cours)
- **Publications** en Philo, Info et Maths
  (Phil&Tech, Minds and Machines, Synthese)
  (ACL, ICLR, ICML, AMS)
- **Conférences invitées**
  Montréal, NYC, Cambridge, Montpellier,
  Singapour (6 derniers mois)
  Venice (keynote), Dagstuhl (2 mois prochains)

Participation dans la communauté scientifique

- **(Vice)-Président** de HaPoC
- **Évaluateur** pour Horizon Europe (MSCA)
- **Reviewer** (Nature SR, Phil&Tech, HSSC, ACL)

Gestion de la recherche

- **Directeur** du Dép. de Rech. (MO.CO.ESBA)
- **Directeur executif** du Turing Center (ETH)
- **Marie Skłodowska-Curie** Fellow
- **Projet** "Human Forms"
  (soumis à E. Schmidt Foundation)
- **Cluster** "Foundations of AI"
  (CUNY, Simons Foundation)

Enseignement et encadrement

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

Intégration

LIPN, UMR 7030 (Paris)

- Équipe **LoCal** (Logique et Calcul)
- Accent sur les **fondements**
  (théorie des types, théorie de catégories, TAL)
- **Rapprochement** de différentes équipes
  (eg. axe "Sc. des données")
- Forte **interdisciplinarité**
  (santé, linguistique, physique, philosophie)

LIRMM, UMR 5506 (Montpellier)

- Équipe **TEXTE** (Exploration et exploitation de données textuelles)
- Accent sur les **applications**
  (grammaires catégorielles, TAL, th. des types)
- Activités **transversales**
  (eg. axe "IA et Sc. des données")
- Forte **interdisciplinarité**
  (projet Muse: "Nourrir, Soigner, Protéger")

Dans les deux cas

- **Collaboration** et **contact** avec des membres et la direction
- Intégration des aspects **épistémologiques** et **sociétaux** dans la recherche
- Présentation de mon travail aux équipes

Références I

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

Bourdieu, P. (1994). *Raisons pratiques: Sur la théorie de l'action*. Éditions du Seuil.

Foucault, M. (1966). *Les mots et les choses : Une archéologie des sciences humaines*. Gallimard.

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. https://tel.archives-ouvertes.fr/tel-01174485

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. https://doi.org/10.1080/03080188.2021.1890484

Gastaldi, J. L., Terilla, J., Malagutti, L., DuSell, B., Vieira, T., & Cotterell, R. (2024).The Foundations of Tokenization: Statistical and Computational Concerns. https://arxiv.org/abs/2407.11606

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

Giulianelli, M., Malagutti, L., Gastaldi, J. L., DuSell, B., Vieira, T., & Cotterell, R. (2024).On the Proper Treatment of Tokenization in Psycholinguistics [To appear in the Proceedings of EMNLP 2024]. https://arxiv.org/abs/2410.02691

Harris, Z. (1960). *Structural linguistics*. University of Chicago Press.

Hjelmslev, L. (1935). *La catégorie des cas*. Wilhelm Fink Verlag.

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.

Hjelmslev, L. (1975). *Résumé of a Theory of Language*. Nordisk Sprog-og Kulturforlag.

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

Références II

Latour, B., Jensen, P., Venturini, T., Grauwin, S., & Boullier, D. (2012).'The whole is always smaller than its parts' - a digital test of Gabriel Tardes' monads. *The British Journal of Sociology*, *63*(4), 590–615.

Lévi-Strauss, C. (1949). *Les structures élémentaires de la parenté*. Presses Universitaires de France.

Lévi-Strauss, C. (1962). *La pensée sauvage*. Plon.

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.

Spang-Hanssen, H. (1959). *Probability and structural classification in language description*. Rosenkilde; Bagger.

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. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Vieira, T., LeBrun, B., Giulianelli, M., Gastaldi, J. L., DuSell, B., Terilla, J., O'Donnell, T. J., & Cotterell, R. (2024). From language models over tokens to language models over characters. https://arxiv.org/abs/2412.03719

Zouhar, V., Meister, C., Gastaldi, J. L., Du, L., Sachan, M., & Cotterell, R. (2023b).Tokenization and the Noiseless Channel. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5184–5207. https://doi.org/10.18653/v1/2023.acl-long.284

Zouhar, V., Meister, C., Gastaldi, J. L., Du, L., Vieira, T., Sachan, M., & Cotterell, R. (2023a).A Formal Perspective on Byte-Pair Encoding. *Findings of the Association for Computational Linguistics: ACL 2023*, 598–614. https://doi.org/10.18653/v1/2023.findings-acl.38

CNRS - Concours chercheurs 2025
CR Section 53 - Concours n° 53/03

Épistémologie des modèles distributionnels de langage par apprentissage machine

Explicabilité formelle et interprétabilité théorique

Juan Luis Gastaldi

`http://www.jlgastaldi.com/assets/gastaldi_cnrs_cr.pdf`

ETH zürich

12 Mai 2025