Chat Token Vector
Università Ca' Foscari
Venice, Italy

Toward a Critical Formalism

Philosophical and Theoretical Effects of a Mathematical Critique of LLMs

Juan Luis Gastaldi

www.giannigastaldi.com

ETH zürich

June 12, 2025

Outline

Intro: Critique and Formalism

Epistemological Critique: LLMs as Formal Objects

Theoretical Critique: Formal Explainability
- The Algebra Behind the Embeddings
- The Structure Behind the Algebra
- The Categories Behind the Structure

Conclusion

Outline

Intro: Critique and Formalism

Epistemological Critique: LLMs as Formal Objects

Theoretical Critique: Formal Explainability
- The Algebra Behind the Embeddings
- The Structure Behind the Algebra
- The Categories Behind the Structure

Conclusion

Where Art Thou, Critique?

Where Art Thou, Critique?

- Good “externalist” critique

Where Art Thou, Critique?

- Good “externalist” critique
- Poor “internalist” critique

Where Art Thou, Critique?

- Good “externalist” critique
- Poor “internalist” critique
  - The main “critical” reference remains the “Stochastic Parrots” approach (Bender & Koller, 2020; Bender et al., 2021)

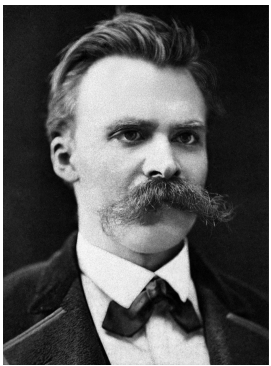
Where Art Thou, Critique?

- Good “externalist” critique
- Poor “internalist” critique
  - The main “critical” reference remains the “Stochastic Parrots” approach (Bender & Koller, 2020; Bender et al., 2021)
  - Kirschenbaum (2023):
    Bender et al.’s (2021) paper “offers a disarmingly linear account of how language, communication, intention, and meaning work, one that would seem to sidestep decades of scholarship around these same issues in literary theory [...] the passage would be red meat for a graduate critical-theory seminar.”

Where Art Thou, Critique?

- Good “externalist” critique
- Poor “internalist” critique
  - The main “critical” reference remains the “Stochastic Parrots” approach (Bender & Koller, 2020; Bender et al., 2021)
  - Kirschenbaum (2023):
    Bender et al.’s (2021) paper “offers a disarmingly linear account of how language, communication, intention, and meaning work, one that would seem to sidestep decades of scholarship around these same issues in literary theory [...] the passage would be red meat for a graduate critical-theory seminar.”
  - Underwood (2023):
    “The beautiful irony of this situation [...] is that a generation of humanists trained on Foucault have now rallied around “On the Dangers of Stochastic Parrots” to oppose a theory of language that their own disciplines invented, just at the moment when computer scientists are reluctantly beginning to accept it.”

The Birth of Contemporary Critique



"In some remote corner of the universe, flickering in the light of the countless solar systems into which it had been poured, there was once a planet on which clever animals invented cognition. It was the most arrogant and most mendacious minute in the 'history of the world'…"

"On Truth and Lying in a Non-Moral Sense"
(Nietzsche, 1873)

The Critical Argumentative Matrix

Knowledge depends on language

The Critical Argumentative Matrix

Knowledge depends on language

The relation between language and the world is essentially arbitrary

The Critical Argumentative Matrix

Knowledge depends on language

The relation between language and the world is essentially arbitrary

Any regularity in language/knowledge is not natural but cultural/social/political

The Critical Argumentative Matrix

Knowledge depends on language

The relation between language and the world is essentially arbitrary

Any regularity in language/knowledge is not natural but cultural/social/political

We should resist existing regularities and create new ones

The Critical Argumentative Matrix

Knowledge depends on language
(Epistemological)

The relation between language and the world is essentially arbitrary

Any regularity in language/knowledge is not natural but cultural/social/political
(Political)

We should resist existing regularities and create new ones
(Aesthetic)

The Critical Argumentative Matrix

Knowledge depends on language
(Epistemological)

[The relation between language and the world is essentially arbitrary?]

Any regularity in language/knowledge is not natural but cultural/social/political
(Political)

We should resist existing regularities and create new ones
(Aesthetic)

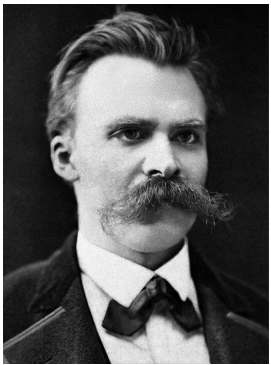
Critique and Formalism

- At the source of this situation is the new foundational role played by formal sciences in the 20th century
  - For a theory of language: Carnap, Gödel, Turing, Shannon, Harris, Chomsky…

Critique and Formalism

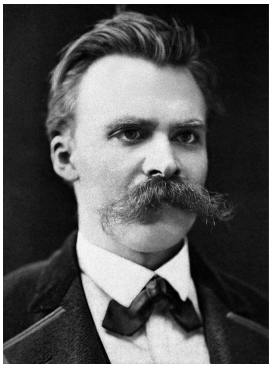
- At the source of this situation is the new foundational role played by formal sciences in the 20th century
  - For a theory of language: Carnap, Gödel, Turing, Shannon, Harris, Chomsky…
- The critical tradition has either withdrawn from the areas conquered by formal approaches, or made formal approaches the target of criticism

Critique vs Formalism



- “their eyes merely glide across the surface of things and see ‘forms’; nowhere does their perception lead into truth; instead it is content to receive stimuli and, as it were, to play with its fingers on the back of things.”

Critique vs Formalism



- "their eyes merely glide across the surface of things and see 'forms'; nowhere does their perception lead into truth; instead it is content to receive stimuli and, as it were, to play with its fingers on the back of things."
- "...the great edifice of concepts exhibits the rigid regularity of a *Roman columbarium*, while logic breathes out that air of severity and coolness which is peculiar to mathematics."
- "...if we are forced to comprehend all things under these forms alone, then it is no longer wonderful that what we comprehend in all these things is actually nothing other than these very forms; for all of them must exhibit the laws of number, and number is precisely that which is most astonishing about things."
- "...the only things we really know about them [laws of nature] are things which we bring to bear on them: time and space, in other words, relations of succession and number."

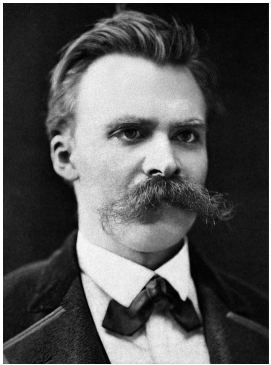
For a Critical Formalism

- At the source of this situation is the new foundational role played by formal sciences in the 20th century
  - For a theory of language: Carnap, Gödel, Turing, Shannon, Harris, Chomsky…
- The critical tradition has either withdrawn from the areas conquered by formal approaches, or made formal approaches the target of criticism

For a Critical Formalism

- At the source of this situation is the new foundational role played by formal sciences in the 20th century
  - For a theory of language: Carnap, Gödel, Turing, Shannon, Harris, Chomsky…
- The critical tradition has either withdrawn from the areas conquered by formal approaches, or made formal approaches the target of criticism
- We need a new strategy: Elaborate a critical formalism

For a Critical Formalism



"...all peoples have just such a mathematically divided firmament of concepts above them [...] Here, one can certainly admire humanity as a mighty architectural genius who succeeds in erecting the infinitely complicated cathedral of concepts on moving foundations, or even, one might say, on flowing water; admittedly, in order to rest on such foundations, it has to be like a thing constructed from cobwebs, so delicate that it can be carried off on the waves and yet so firm as not to be blown apart by the wind."

(Nietzsche, 1873)

For a Critical Formalism

- In the case of AI, a critical formalism can provide:
  - New epistemological tools countering dogmatic perspectives stemming from within the field
  - New theoretical tools contributing to the non-dogmatic positive production of knowledge

Outline

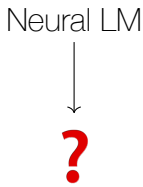
Intro: Critique and Formalism

Epistemological Critique: LLMs as Formal Objects

Theoretical Critique: Formal Explainability
- The Algebra Behind the Embeddings
- The Structure Behind the Algebra
- The Categories Behind the Structure

Conclusion

Neural LMs as Computable Functions

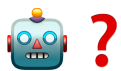


Neural LMs as Computable Functions

Neural LM

Neural LMs as Computable Functions

Neural LM



Neural LMs as Computable Functions

Neural LM

Neural LMs as Computable Functions

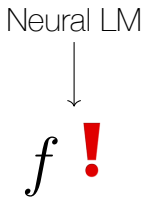
Neural LM

Neural LMs as Computable Functions

Neural LM

?

Neural LMs as Computable Functions



Neural LMs as Computable Functions

Neural LM

Function

$$f : \mathbb{R}^n \xrightarrow{f_1} \mathbb{R}^{n_1} \xrightarrow{f_2} \cdots \xrightarrow{f_K} \mathbb{R}^{n_K} \xrightarrow{g} \mathbb{R}^m$$

Neural LMs as Computable Functions

Neural LM

↓

Computable

Function

$$f : \mathbb{R}^n \xrightarrow{f_1} \mathbb{R}^{n_1} \xrightarrow{f_2} \cdots \xrightarrow{f_K} \mathbb{R}^{n_K} \xrightarrow{g} \mathbb{R}^m$$

Neural LMs as Computable Functions

Neural LM

↓

Computable

Function

$$f : \mathbb{R}^n \xrightarrow{f_1} \mathbb{R}^{n_1} \xrightarrow{f_2} \cdots \xrightarrow{f_K} \mathbb{R}^{n_K} \xrightarrow{g} \mathbb{R}^m$$

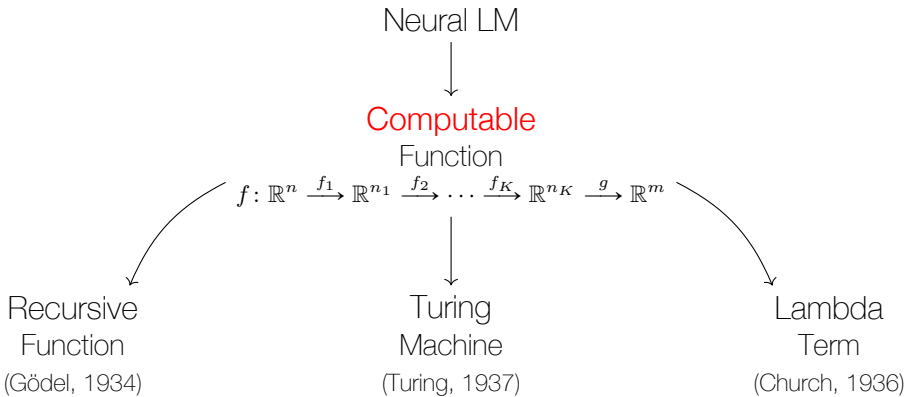
↓

Turing

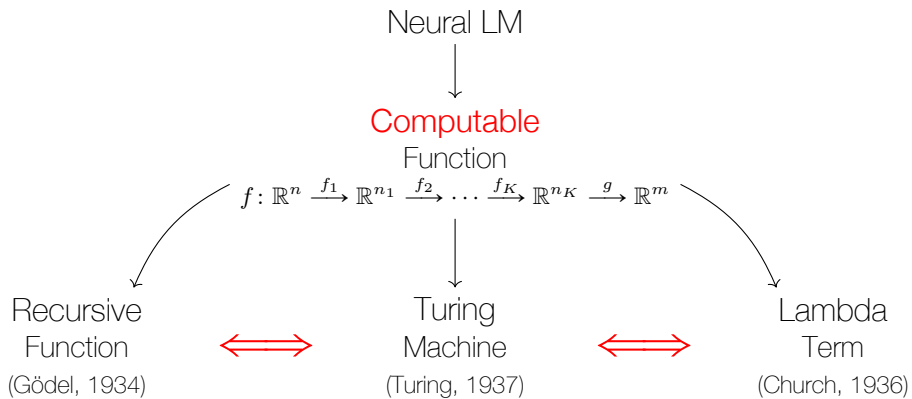
Machine

(Turing, 1937)

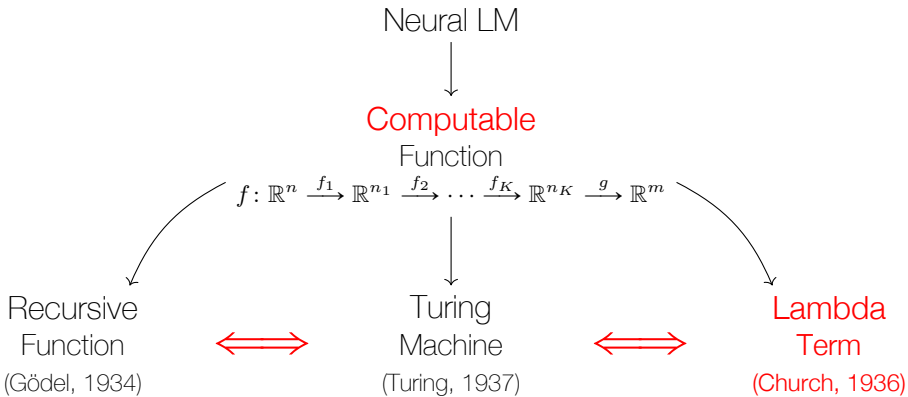
Neural LMs as Computable Functions



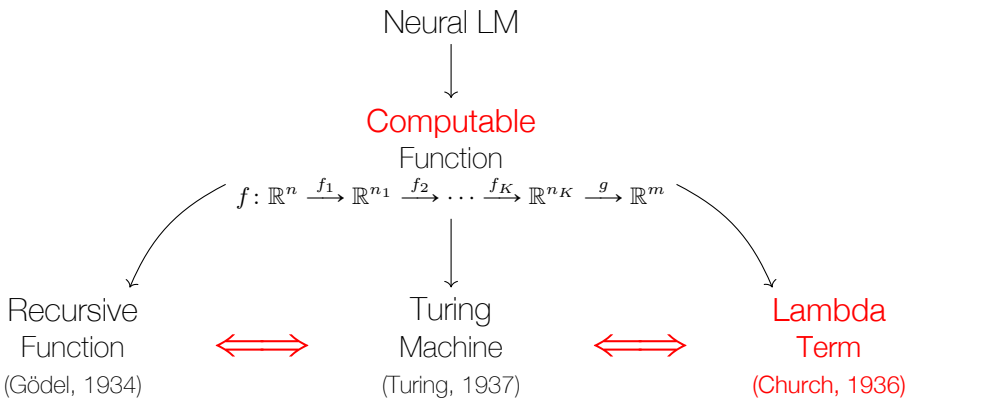
Neural LMs as Computable Functions



Neural LMs as Computable Functions



Neural LMs as Computable Functions



λ-abstraction and β-reduction in λ-calculus

$$yxz$$

λ-abstraction and β-reduction in λ-calculus

$$\lambda x.yxz$$

λ-abstraction and β-reduction in λ-calculus

$$(\lambda x.yxz)t$$

λ-abstraction and β-reduction in λ-calculus

$$(\lambda x.yxz)t$$

$$\downarrow$$

$$ytz$$

$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$

$$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$$

0: $\lambda f.\lambda x.x$

1: $\lambda f.\lambda x.fx$

2: $\lambda f.\lambda x.f(fx)$

3: $\lambda f.\lambda x.f(f(fx))$

4: $\lambda f.\lambda x.f(f(f(fx)))$

5: $\lambda f.\lambda x.f(f(f(f(fx))))$

$\ldots$

n: $\lambda f.\lambda x.\underbrace{f(\ldots(f}_{n \text{ times}}x)\ldots)$

Empirical Evaluation

$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$

0: $\lambda f.\lambda x.x$

1: $\lambda f.\lambda x.fx$

2: $\lambda f.\lambda x.f(fx)$

3: $\lambda f.\lambda x.f(f(fx))$

4: $\lambda f.\lambda x.f(f(f(fx)))$

5: $\lambda f.\lambda x.f(f(f(f(fx))))$

$\ldots$

n: $\lambda f.\lambda x.\underbrace{f(\ldots(f}_{n \text{ times}}x)\ldots)$

$\lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)(\lambda f.\lambda x.f(fx))(\lambda f.\lambda x.f(f(fx)))$

Empirical Evaluation

$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$

0: $\lambda f.\lambda x.x$

1: $\lambda f.\lambda x.fx$

2: $\lambda f.\lambda x.f(fx)$

3: $\lambda f.\lambda x.f(f(fx))$

4: $\lambda f.\lambda x.f(f(f(fx)))$

5: $\lambda f.\lambda x.f(f(f(f(fx))))$

$\ldots$

n: $\lambda f.\lambda x.\underbrace{f(\ldots(f}_{n \text{ times}}x)\ldots)$

$\lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)(\lambda f.\lambda x.f(fx))(\lambda f.\lambda x.f(f(fx)))$

$\lambda f.\lambda x.f(f(f(f(fx))))$

Empirical Evaluation

$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$
$P' := \lambda r.\lambda s.\lambda f.\lambda x.f(f(f(f(fx))))$

0: $\lambda f.\lambda x.x$
1: $\lambda f.\lambda x.fx$
2: $\lambda f.\lambda x.f(fx)$
3: $\lambda f.\lambda x.f(f(fx))$
4: $\lambda f.\lambda x.f(f(f(fx)))$
5: $\lambda f.\lambda x.f(f(f(f(fx))))$
$\ldots$
n: $\lambda f.\lambda x.\underbrace{f(\ldots(f}_{n \text{ times}}x)\ldots)$

$\lambda r.\lambda s.\lambda f.\lambda x.f(f(f(f(fx))))(\lambda f.\lambda x.f(fx))(\lambda f.\lambda x.f(f(fx)))$

$\lambda f.\lambda x.f(f(f(f(fx))))$

$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$

0: $\lambda f.\lambda x.x$

1: $\lambda f.\lambda x.fx$

2: $\lambda f.\lambda x.f(fx)$

3: $\lambda f.\lambda x.f(f(fx))$

4: $\lambda f.\lambda x.f(f(f(fx)))$

5: $\lambda f.\lambda x.f(f(f(f(fx))))$

$\ldots$

n: $\lambda f.\lambda x.\underbrace{f(\ldots(f}_{n \text{ times}}\, x)\ldots)$

$\lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)(\lambda f.\lambda x.f(fx))(\lambda f.\lambda x.f(f(fx)))$

$\vdots$

$\lambda f.\lambda x.f(f(f(f(fx))))$

$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$

0: $\lambda f.\lambda x.x$

1: $\lambda f.\lambda x.fx$

2: $\lambda f.\lambda x.f(fx)$

3: $\lambda f.\lambda x.f(f(fx))$

4: $\lambda f.\lambda x.f(f(f(fx)))$

5: $\lambda f.\lambda x.f(f(f(f(fx))))$

$\ldots$

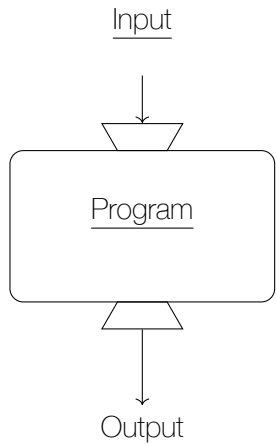
n: $\lambda f.\lambda x.\underbrace{f(\ldots(f}_{n \text{ times}}x)\ldots)$

$$\lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)(\lambda f.\lambda x.f(fx))(\lambda f.\lambda x.f(f(fx)))$$
$$\lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)(\lambda g.\lambda y.g(gy))(\lambda h.\lambda z.h(h(hz)))$$
$$\lambda n.\lambda f.\lambda x.(\lambda g.\lambda y.g(gy))f(nfx)(\lambda h.\lambda z.h(h(hz)))$$
$$\lambda n.\lambda f.\lambda x.(\lambda g.\lambda y.g(gy))f(nfx)(\lambda h.\lambda z.h(h(hz)))$$
$$\lambda f.\lambda x.(\lambda g.\lambda y.g(gy))f((\lambda h.\lambda z.h(h(hz)))fx)$$
$$\lambda f.\lambda x.(\lambda y.f(fy))((\lambda h.\lambda z.h(h(hz)))fx)$$
$$\lambda f.\lambda x.(\lambda y.f(fy))((\lambda z.f(f(fz)))x)$$
$$\lambda f.\lambda x.(\lambda y.f(fy))(f(f(fx)))$$
$$\lambda f.\lambda x.f(f(f(f(fx))))$$

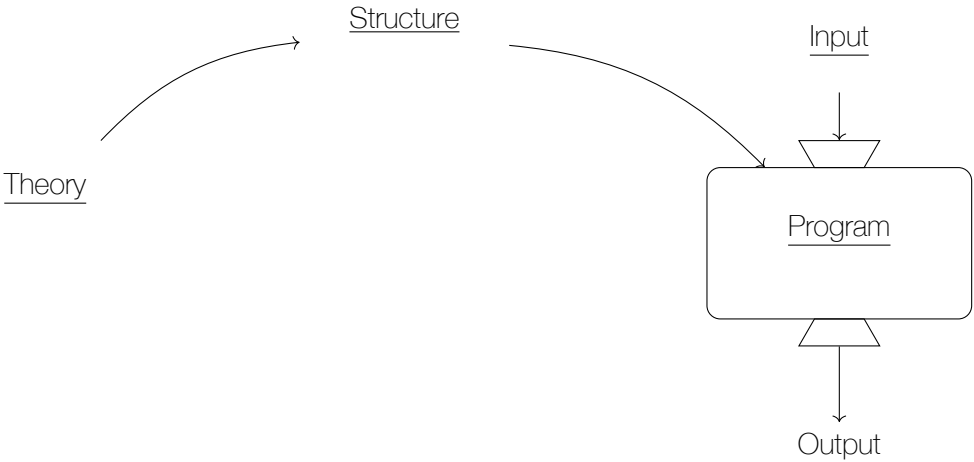
$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$

$P'' :=$ λRófÄÒêÑ5È|Àxñ=∞ù ÿmWf286ëy’SÒú>v&ìÂ ¬2 óÉ7öç∞{ã>2flB°µG#À9ÇU ∞btYBô¥ Ù ë‰3;5 å[l–èuô¨Ü ˚ 7–Ù.λ:ˆ4mÓØ¥ ´è—+ÌsÖ,$+gï„B™÷o–#ïŸê Ûv –gÓÿ/ëiijO‡Œfi • J1«€ø¸Ï hãêt‡æY$ˆ6 FïW»RÙKg¢¨.λ‡d¯… D2÷¸ò ˚ xêÈy. Ó”¢b Bé£NÈ1Ê‡/Û9Ñµ–/JYÇõË9ÿÀÈ.λÁÍ À^öÇ,»fq∞±îˇB5Ì>O˜g™“6Ωe“æëC/ã… Ö · fÓ Å]ÑåyÊN°Ê ›¨.λÆà€fUòfEÙ´Í m#„¸4\r√¬÷Îpò»y*vtÄJÃF1ûÁóz«ñM”DjŒ BËèÍT _ Êa‰ÁÇΩ @\Ø^˜]Î›h‡.:ˆ4mÓØ¥´è—+ÌsÖ,$+gï„B™÷o–#ïŸê Ûv–gÓÿ /ëiijO‡Œfi • J1«€ø¸Ï hãêt‡æY$ˆ6 FïW»RÙKg¢¨ÁÍ À^öÇ,»fq∞±îˇB5Ì>O˜g™“6 Ωe“æëC/ã… Ö · fÓ Å]ÑåyÊN°Ê ›¨(‡d¯… D2÷¸ò ˚ xêÈy. Ó”¢bBé£NÈ1Ê‡/Û9Ñµ–/ JYÇõË9ÿÀÈÁÍ À^öÇ,»fq∞±îˇB5Ì>O˜g™“6Ωe“æëC/ã… Ö · fÓ Å]ÑåyÊN°Ê ›¨Æ à€fUòfEÙ´Í m#„¸4\r√¬÷Îpò»y*vtÄJÃF1ûÁóz«ñM”DjŒBËèÍT _ Êa‰ÁÇΩ @\ Ø^˜]Î›h‡)(ËÍÛeí4WµÍ }w„$Ω“K5èÃ¶%3[m ’ˇBÃfifÒ; óJçCËîõŸØcB,ñ$Åå}ÓÅØ3¡ ˇ?ö¯oŒ@ fl8ˇR CÆoˇ* &<Ÿ ¬o12À‰ àÓÜ#ï ”,ú¨ «ô¸„∞Îâä“øÃd|ˆÑ’Ê yØ;ˆW ›”wò[}\»ÖÈúw’6<ùˇ=ãÕ ¯Ì‘Dz?2±|é‘ 3Â/rxµ∞µ$ÀëÂ*lf˜îû’+ÍVìyªGæßägô/,uÑ)

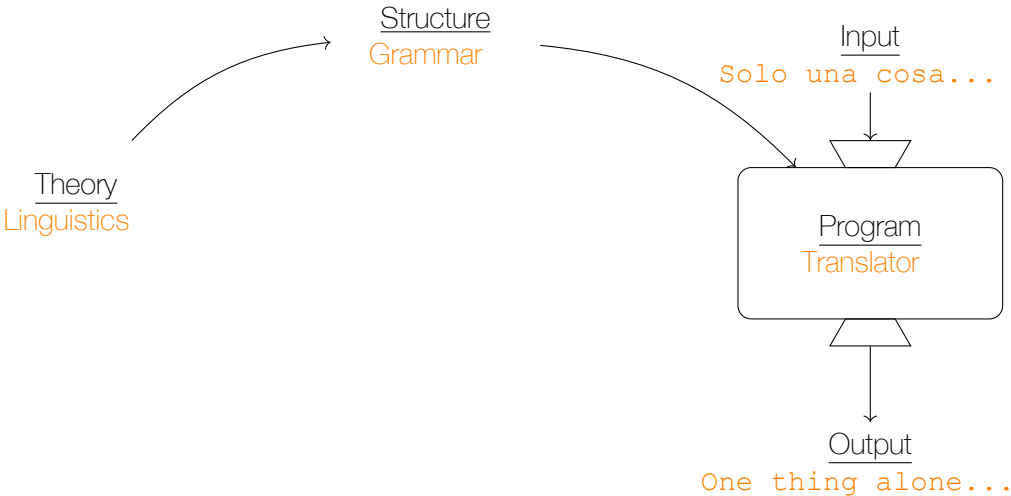
Making It Explicit



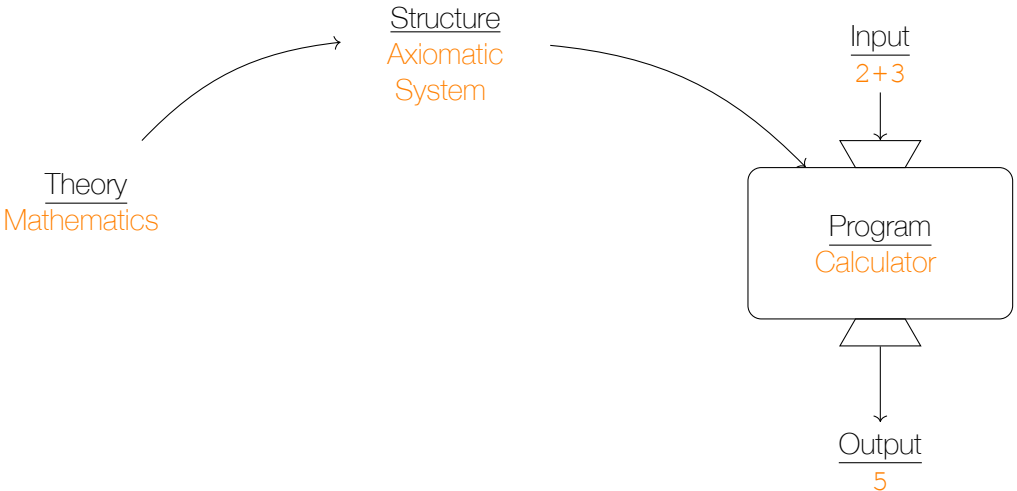
Making It Explicit



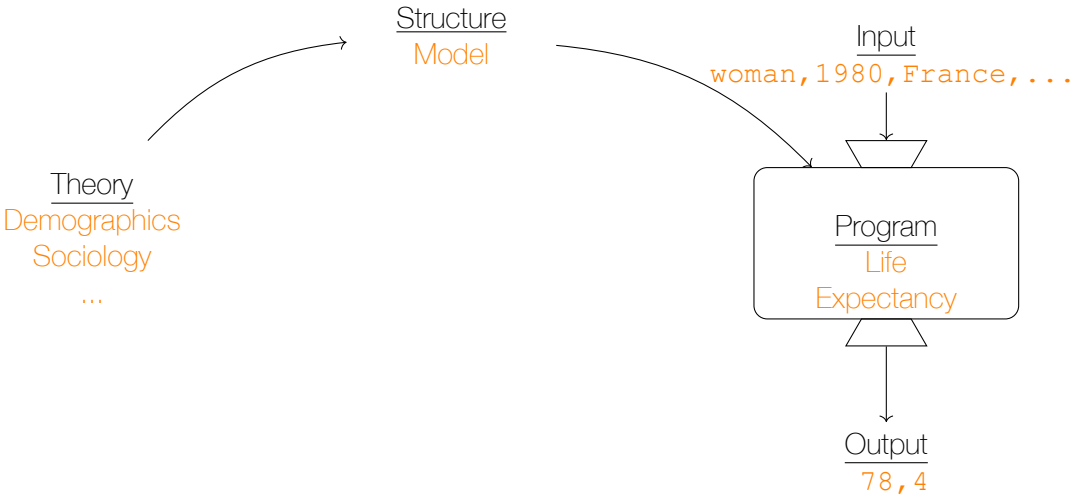
Making It Explicit



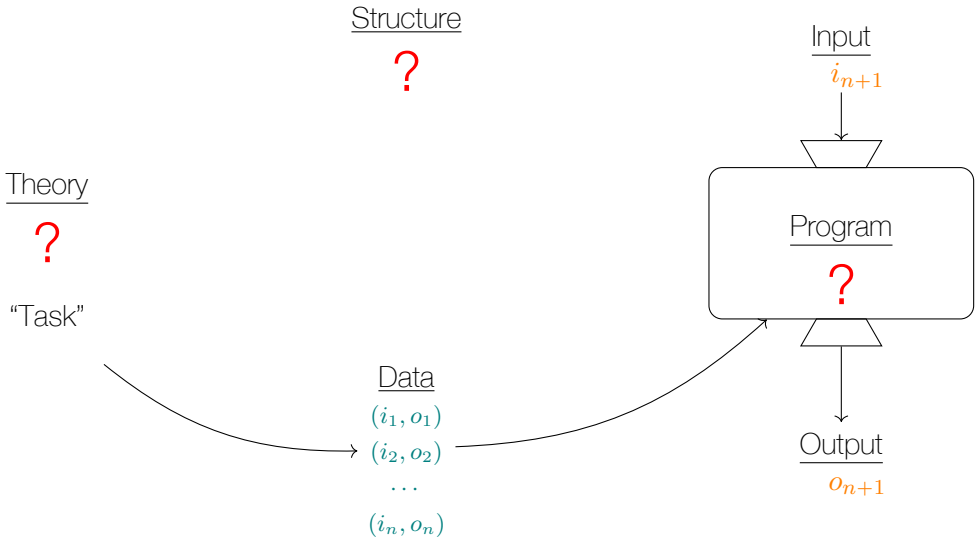
Making It Explicit



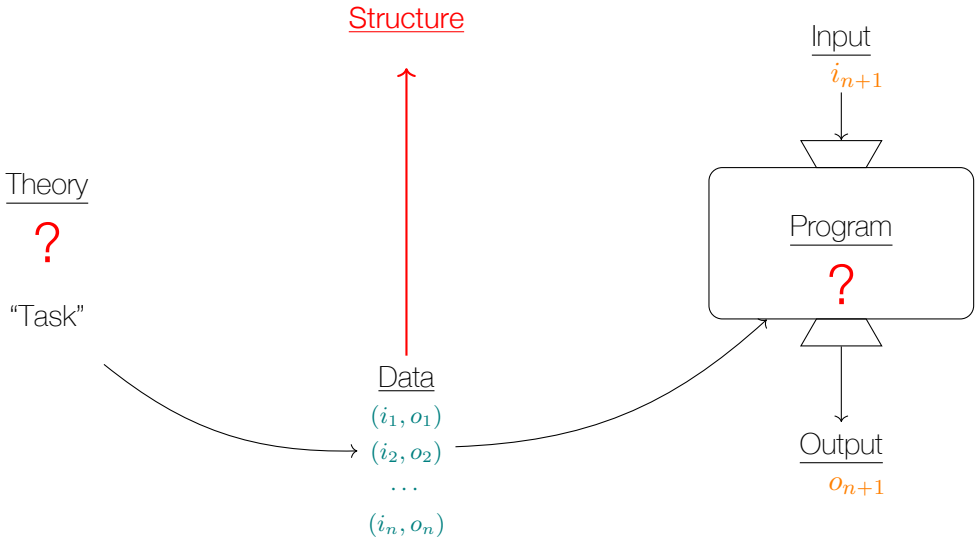
Making It Explicit



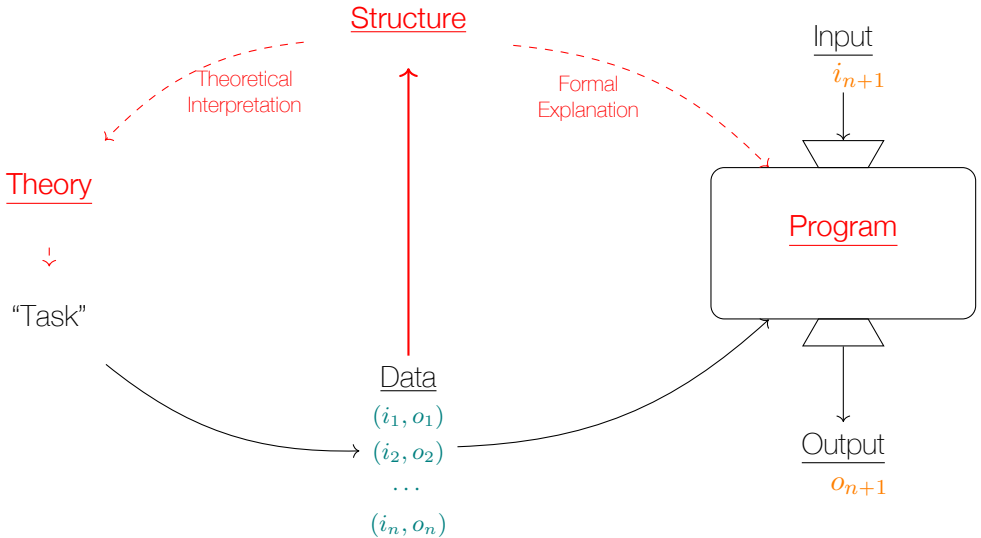
Making It Explicit



Making It Explicit



Making It Explicit



Outline

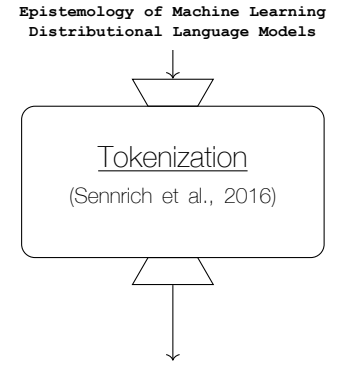
Intro: Critique and Formalism

Epistemological Critique: LLMs as Formal Objects

Theoretical Critique: Formal Explainability

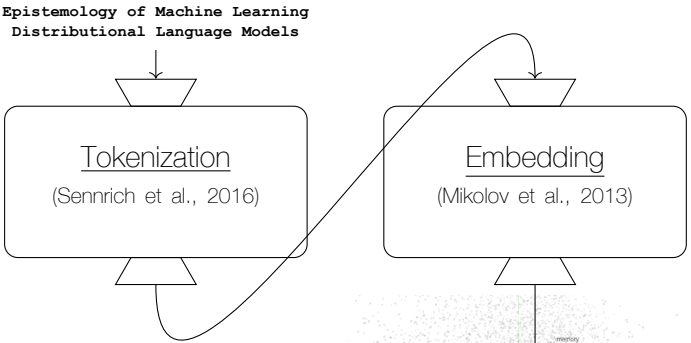
- The Algebra Behind the Embeddings
- The Structure Behind the Algebra
- The Categories Behind the Structure

Conclusion



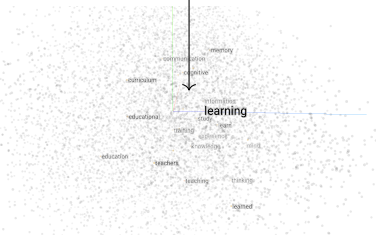
(https://tiktokenizer.vercel.app)

Formal Explainability



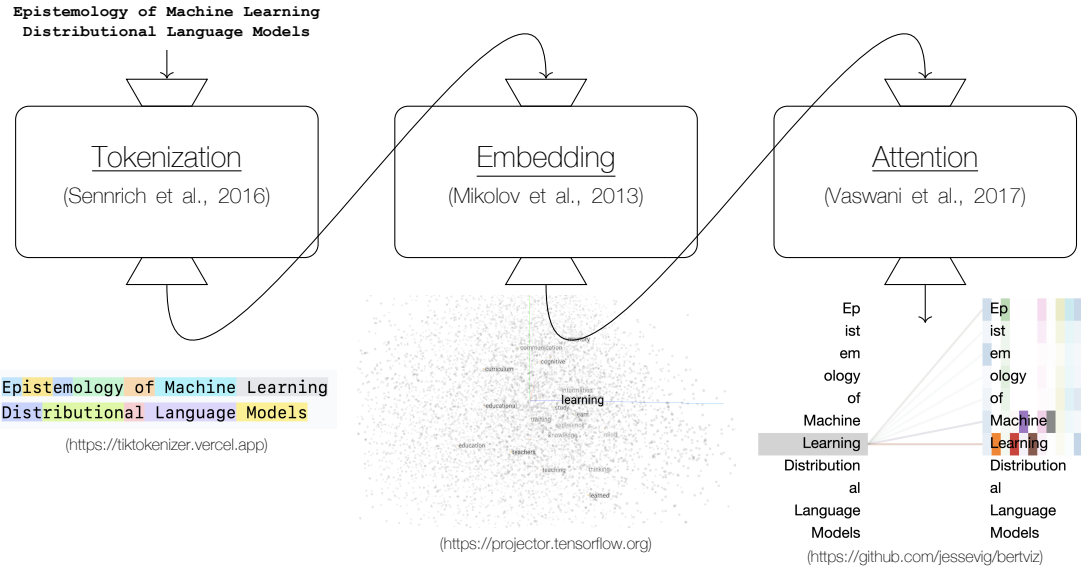
Epistemology of Machine Learning
Distributional Language Models

(https://tiktokenizer.vercel.app)

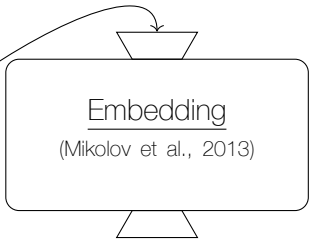


(https://projector.tensorflow.org)

Formal Explainability

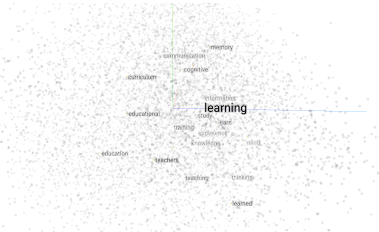


Formal Explainability



Epistemology of Machine Learning
Distributional Language Models

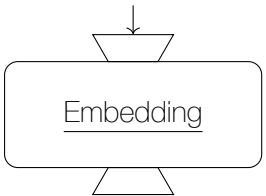
(https://tiktokenizer.vercel.app)

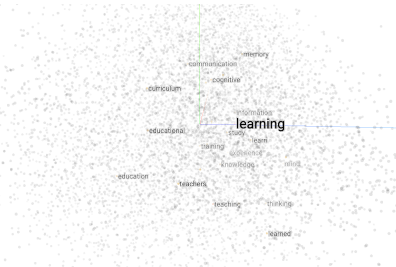


(https://projector.tensorflow.org)

The Structure of Embeddings

Epistemology of Machine Learning
Distributional Language Models





The Structure of Embeddings

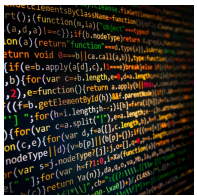
Structure

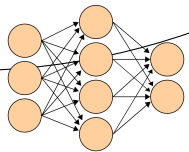
?

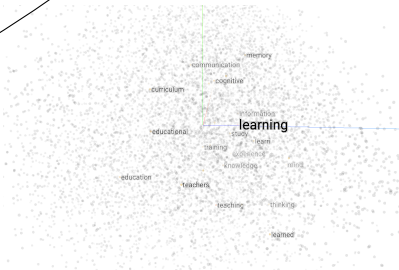
Epistemology of Machine Learning
Distributional Language Models

Embedding

Data







The Structure of Embeddings

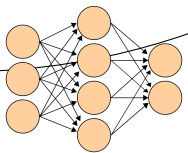
Structure

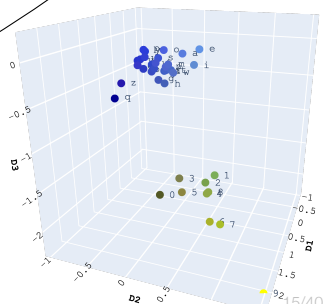
?

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

Embedding

Data





Outline

Intro: Critique and Formalism

Epistemological Critique: LLMs as Formal Objects

Theoretical Critique: Formal Explainability
- The Algebra Behind the Embeddings
- The Structure Behind the Algebra
- The Categories Behind the Structure

Conclusion

word2vec Explained

(Levy and Goldberg, 2014)

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

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

- Word2vec performs an implicit, low-dimensional factorization of a pointwise mutual information (pmi), word-context matrix.

word2vec Explained

(Levy and Goldberg, 2014)

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

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

- Word2vec performs an implicit, low-dimensional factorization of a pointwise mutual information (pmi), word-context matrix.

word2vec Explained

(Levy and Goldberg, 2014)

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

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

- Word2vec performs an implicit, low-dimensional factorization of a pointwise mutual information (pmi), word-context matrix.

word2vec Explained

(Levy and Goldberg, 2014)

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

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

- Word2vec performs an implicit, low-dimensional factorization of a pointwise mutual information (pmi), word-context matrix.

word2vec Explained

(Levy and Goldberg, 2014)

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

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

- Word2vec performs an implicit, low-dimensional factorization of a pointwise mutual information (pmi), word-context matrix.

word2vec Explained

(Levy and Goldberg, 2014)

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

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

- Word2vec performs an implicit, low-dimensional factorization of a pointwise mutual information (pmi), word-context matrix.

word2vec Explained

(Levy and Goldberg, 2014)

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

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

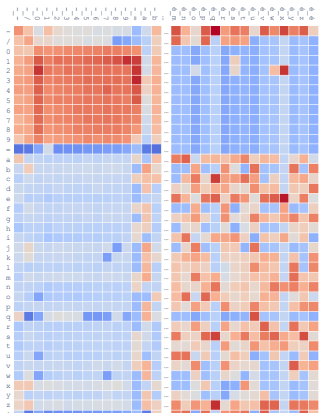
- Word2vec performs an implicit, low-dimensional factorization of a pointwise mutual information (pmi), word-context matrix.
- The Singular Value Decomposition (SVD) provides an exact solution to this problem.

Example: Characters in Wikipedia

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

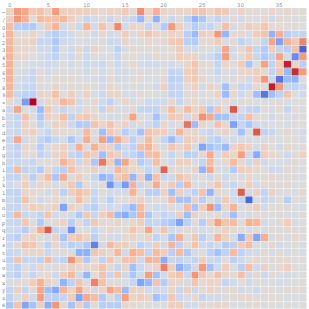
$$C = X \times X = \{\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\}$$

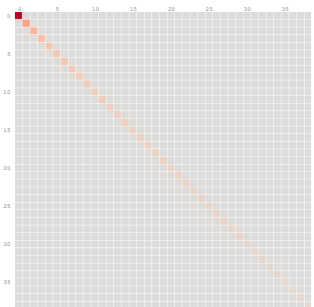
$$\begin{aligned} M_{wc} &= \operatorname{pmi}(w, c) \\ &= \log \frac{p(w, c)}{p(w)p(c)} \end{aligned}$$

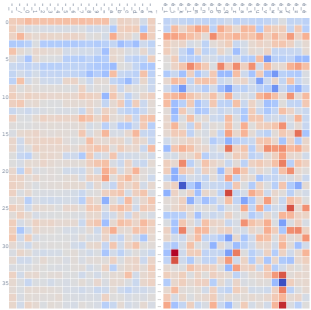


SVD of Wikipedia Character PMI Matrix

$$U \qquad \Sigma \qquad V^T$$

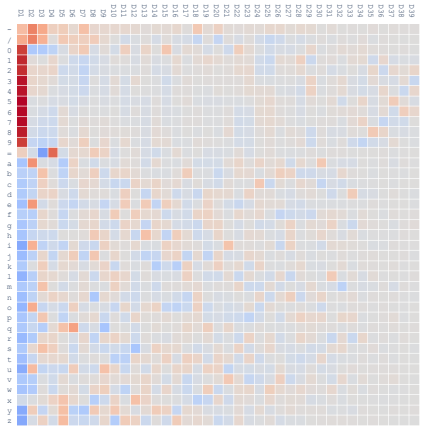






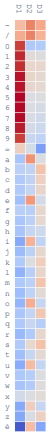
Truncate

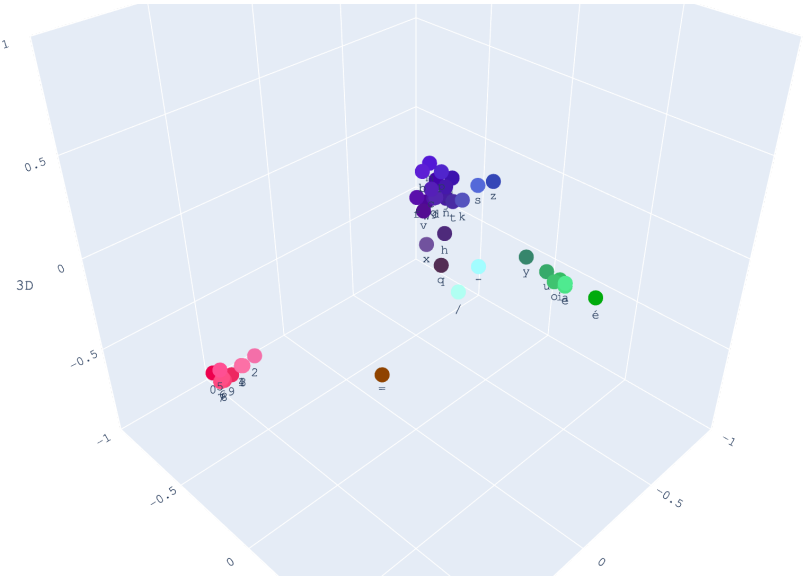
$$U \times \Sigma$$

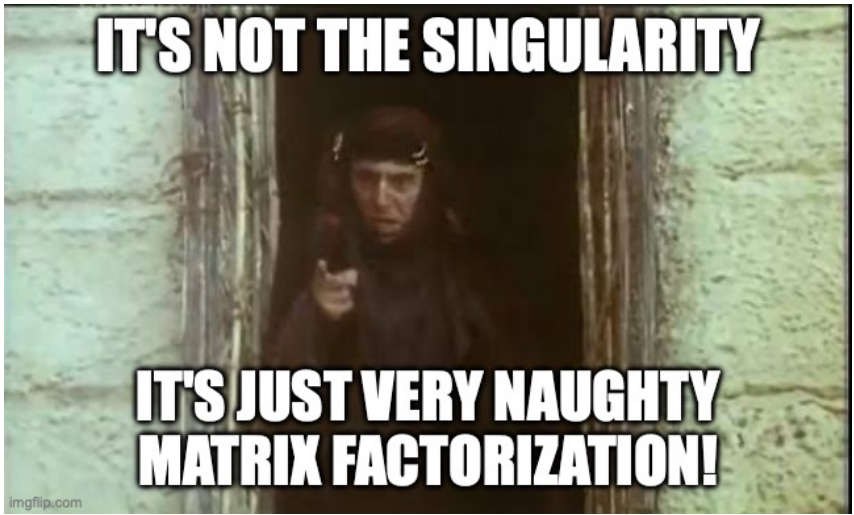


$$\hat{U} \times \hat{\Sigma}$$

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






IT'S NOT THE SINGULARITY
IT'S JUST VERY NAUGHTY
MATRIX FACTORIZATION!
imgflip.com

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 Embeddings

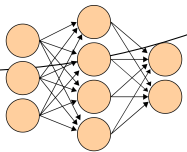
Structure

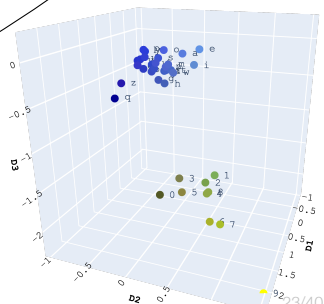
?

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

Embedding

Data





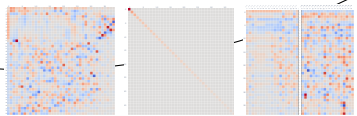
The Structure of Embeddings

Structure

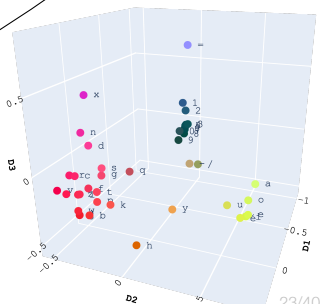
{-, /, 0, 1, 2, . . . , 8, 9, =,
a, b, c, . . . , w, x, y, z, é}

Embedding

Data



SVD



Outline

Intro: Critique and Formalism

Epistemological Critique: LLMs as Formal Objects

Theoretical Critique: Formal Explainability

- The Algebra Behind the Embeddings
- The Structure Behind the Algebra
- The Categories Behind the Structure

Conclusion

Embeddings as Functions Over Sets

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

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

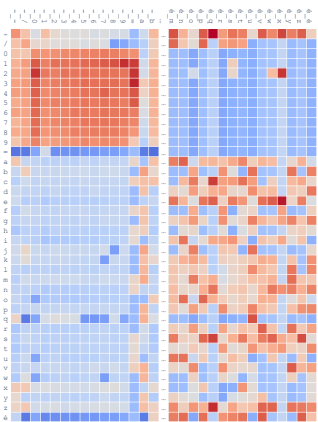
Embeddings as Functions Over Sets

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

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

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

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



Embeddings as Functions Over Sets

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

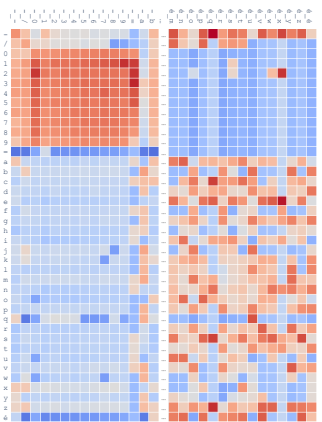
$$Y = X \times X = \{\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\}$$

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

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

$$M_x \colon X \to \mathbb{R}^Y$$

$$x \mapsto M(x, -)$$



Embeddings as Functions Over Sets

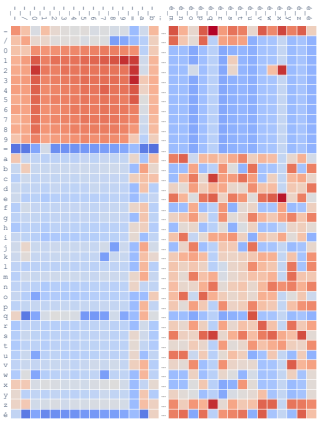
$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

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

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

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

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



Embeddings as Functions Over Sets

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

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

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

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

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

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

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

Embeddings as Functions Over Sets

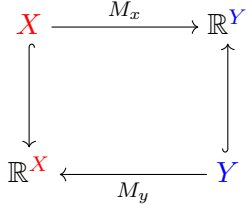
$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

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

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

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

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



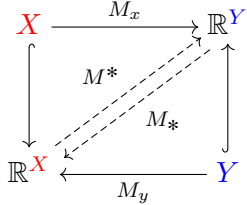
Embeddings as Functions Over Sets

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

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

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

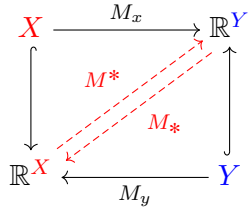
$$M_y\colon Y \to \mathbb{R}^X$$
$$y \mapsto M(-, y)$$



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

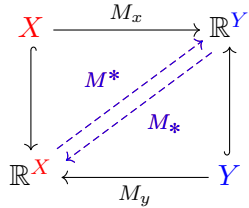
Embeddings as Functions Over Sets

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



Embeddings as Functions Over Sets

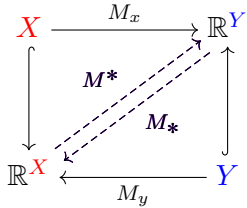
$$M_* M^* : \mathbb{R}^X \to \mathbb{R}^X$$
$$M^* M_* : \mathbb{R}^Y \to \mathbb{R}^Y$$



Embeddings as Functions Over Sets

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

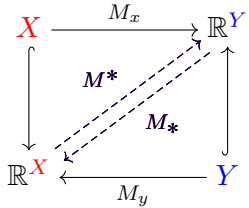
$$\{u_1, \ldots, u_m\} \subset \mathbb{R}^X$$
$$\{v_1, \ldots, v_n\} \subset \mathbb{R}^Y$$
$$\{\lambda_1, \ldots, \lambda_{\min(m,n)}, 0, \ldots, 0\}$$



Embeddings as Functions Over Sets

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

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



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

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

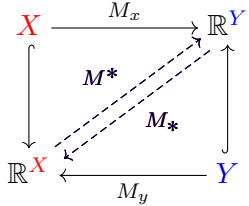
Embeddings as Functions Over Sets

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

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

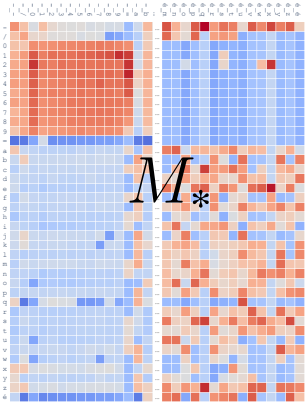
$$M_* M^* u_i = \lambda_i u_i$$
$$M^* M_* v_i = \lambda_i v_i$$

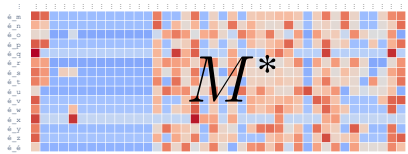
The u_i and v_i are (linear) fixed points!

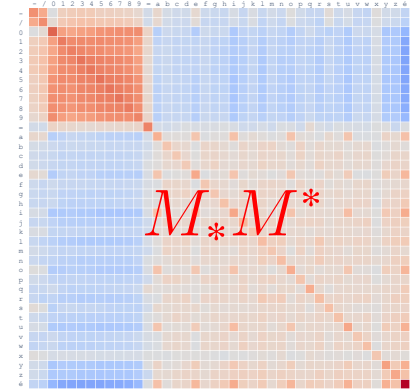


$$M = U \Sigma V^T \qquad U := [u_1, \ldots, u_m] \qquad V := [v_1, \ldots, v_n] \qquad \Sigma := \begin{bmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\lambda_r} \end{bmatrix}$$

M_*M^* as a Covariance Matrix

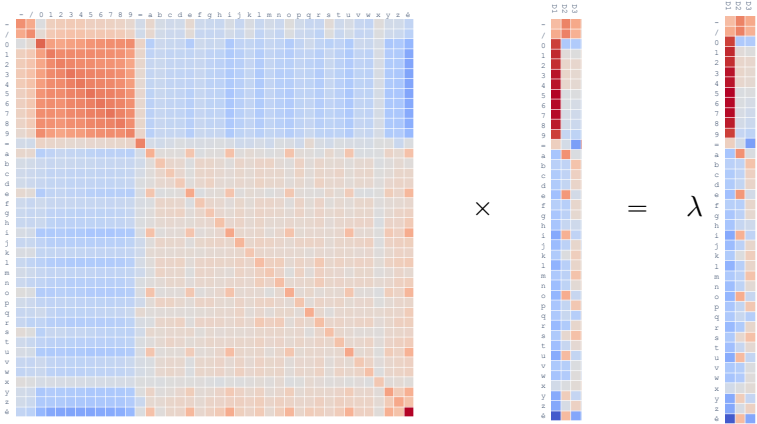




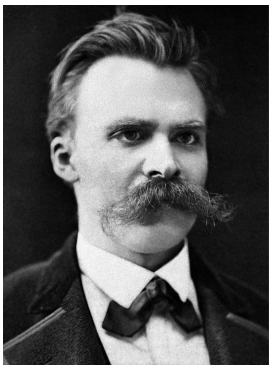


Eigenvectors as Fixed Points

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



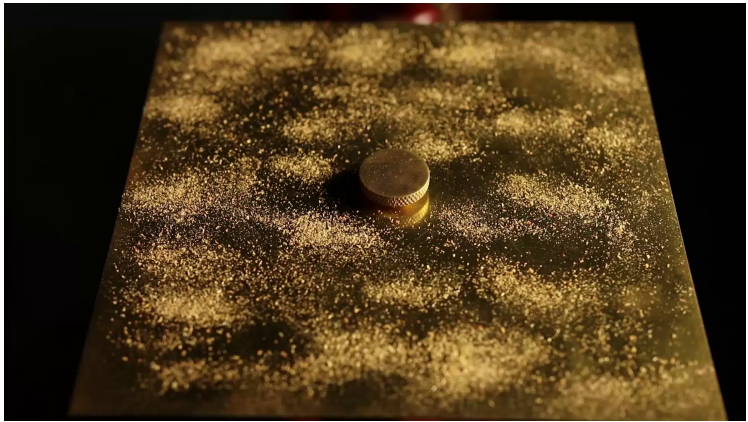
Chladni Figures



"One can conceive of a profoundly deaf human being who has never experienced sound or music; just as such a person will gaze in astonishment at the Chladnian sound-figures in sand, find their cause in the vibration of a string, and swear that he must now know what men call sound — this is precisely what happens to all of us with language."

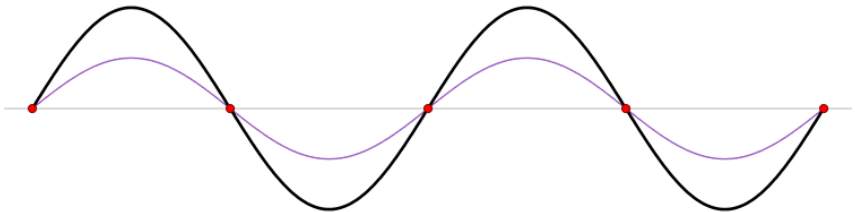
(Nietzsche, 1873)

Chladni Figures

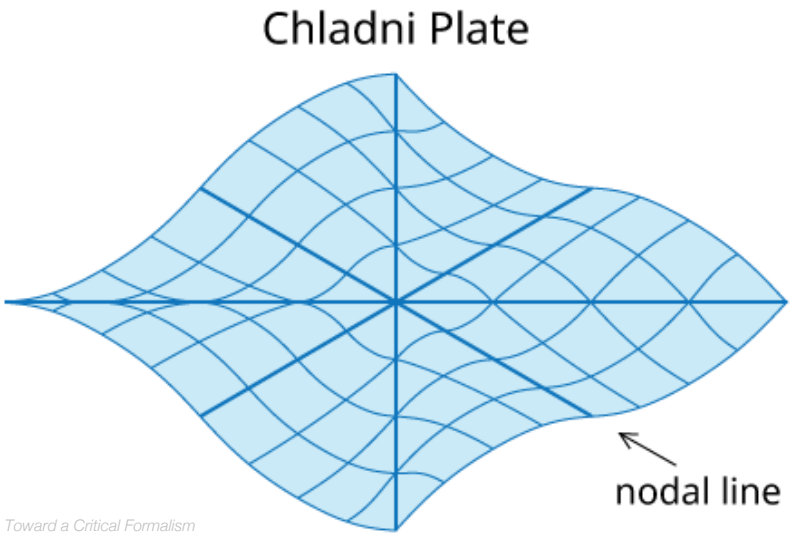


Chladni Figures

Chladni Figures

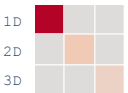


Chladni Figures

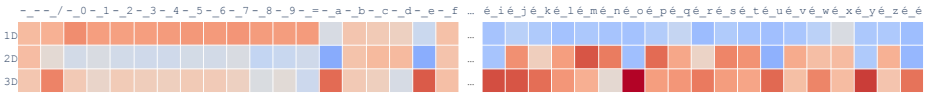


Eigenvectors of $M_* M^*$:

Eigenvalues of $M_* M^*$ and $M^* M_*$:

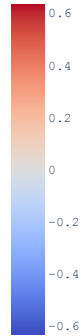


Eigenvectors of $M^* M_*$:



Words

	-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



Outline

Intro: Critique and Formalism

Epistemological Critique: LLMs as Formal Objects

Theoretical Critique: Formal Explainability

- The Algebra Behind the Embeddings
- The Structure Behind the Algebra
- The Categories Behind the Structure

Conclusion

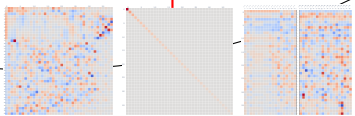
The Structure of Embeddings

Structure

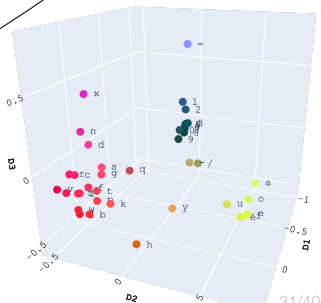
{-, /, 0, 1, 2, . . . , 8, 9, =,
a, b, c, . . . , w, x, y, z, é}

Embedding

Data

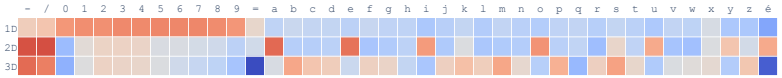


SVD



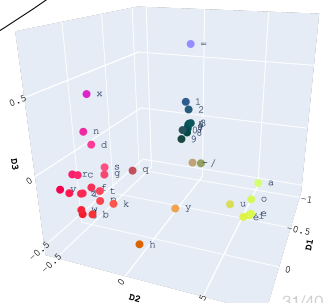
The Structure of Embeddings

Structure

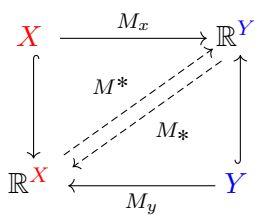


$\{-, /, 0, 1, 2, \ldots, 8, 9, =,$
$a, b, c, \ldots, w, x, y, z, é\}$

Embedding



Data



The Structure of Embeddings

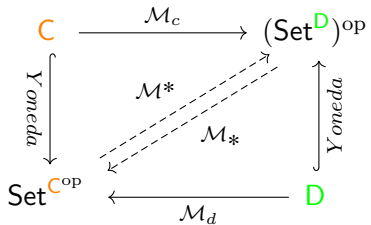
Structure

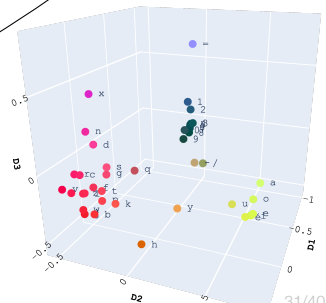
?

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

Embedding

Data





The Structure of Embeddings

Structure

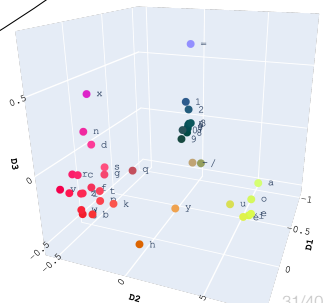
?

Data

$C^{op} \times D \rightarrow Set$

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

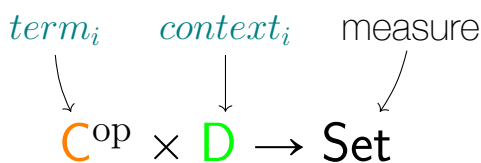
Embedding



Nucleus and Types

Structure

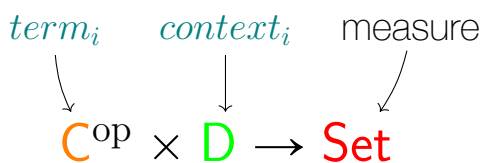
?



Nucleus and Types

Structure

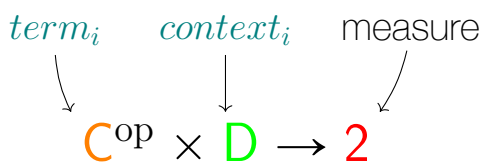
?



Nucleus and Types

Structure

?

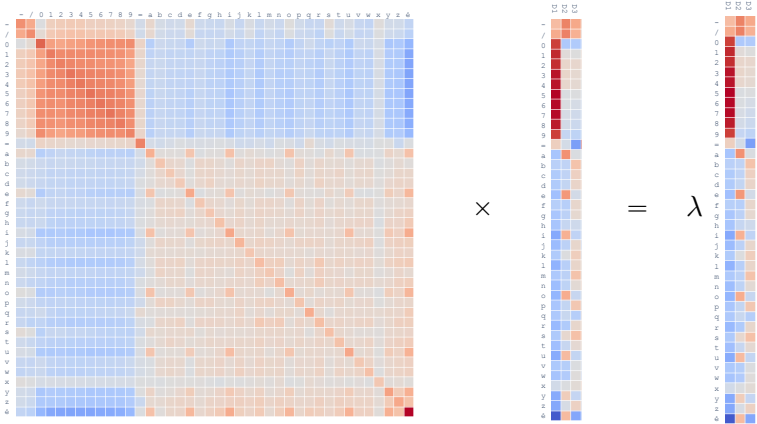


Structure

$$\begin{array}{c} \mathsf{C}^{\mathrm{op}} \times \mathsf{D} \longrightarrow 2 \\ \Downarrow \\ \mathcal{M}^* : 2^{\mathsf{C}^{\mathrm{op}}} \leftrightarrows (2^{\mathsf{D}})^{\mathrm{op}} : \mathcal{M}_* \end{array}$$

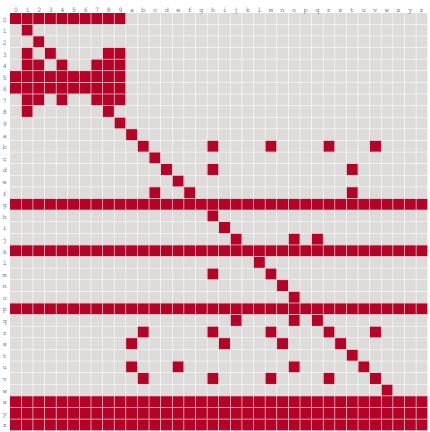
Binary Fixed Points

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

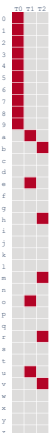


Binary Fixed Points

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



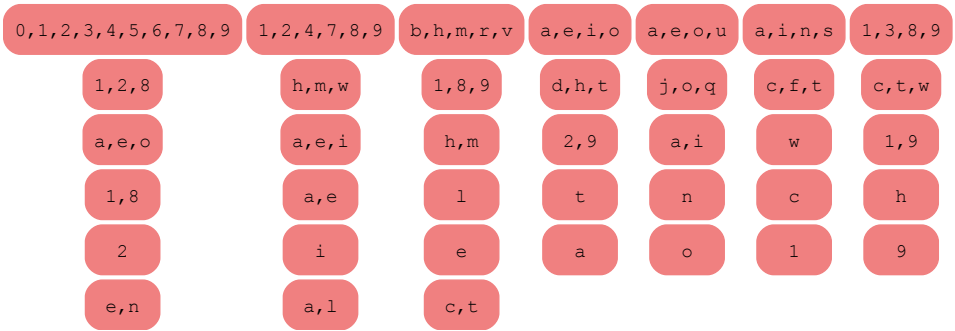
$\star$



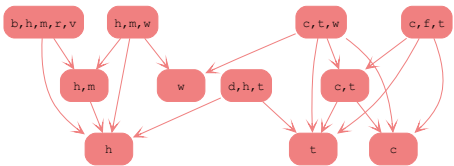
$\stackrel{?}{=}$

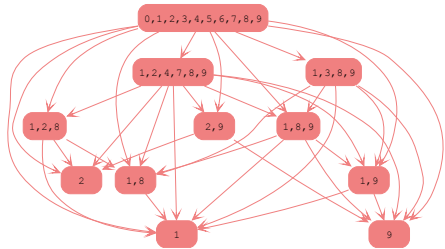
“Eigensets”

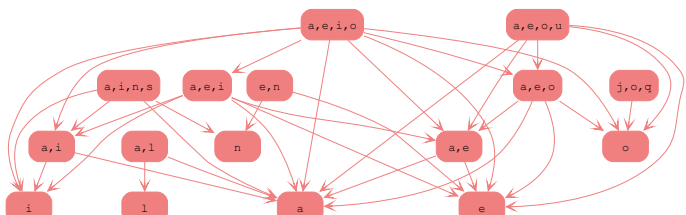
$$\mathcal{M}_* \mathcal{M}^* f = f$$



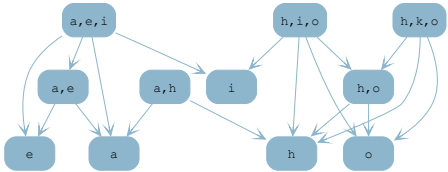
Partial Order Structure

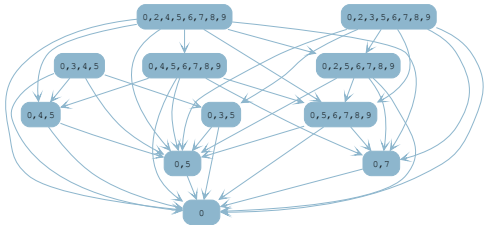


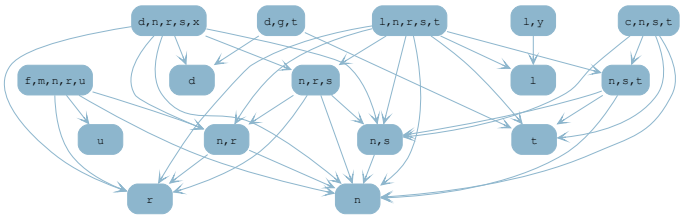




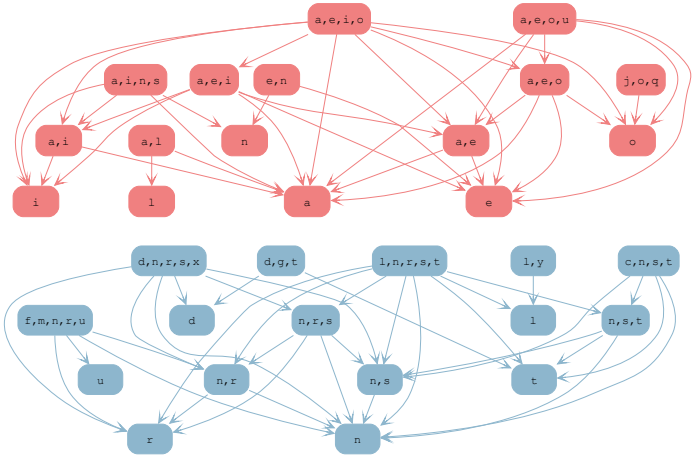
Dual Partial Order



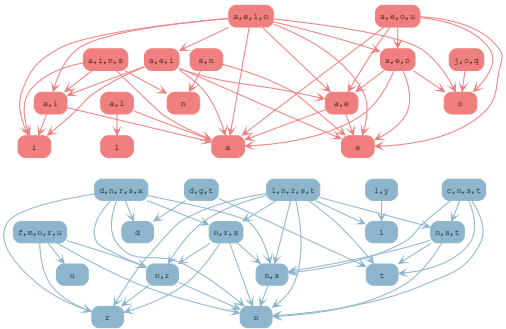




Paring of Partial Ordered Fixed Points



Structure



$$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \longrightarrow 2$$

$$\Downarrow$$

$$\mathcal{M}^* \colon 2^{\mathsf{C}^{\mathrm{op}}} \leftrightarrows \left(2^{\mathsf{D}}\right)^{\mathrm{op}} \colon \mathcal{M}_*$$

Structure

?

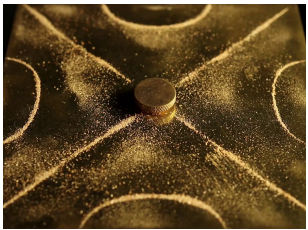
$$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \longrightarrow \bar{\mathbb{R}}$$

$$\Downarrow$$

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

Structure

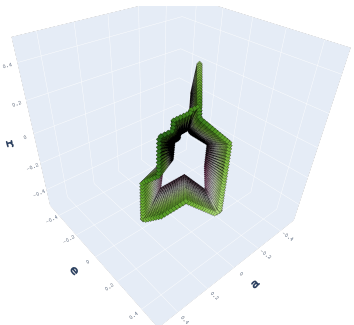
?



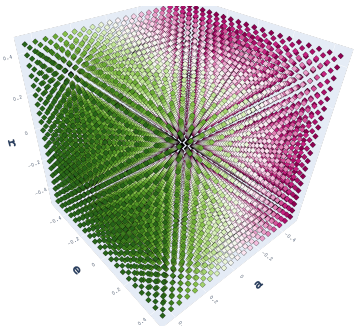
$$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \longrightarrow \bar{\mathbb{R}}$$

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

Structure



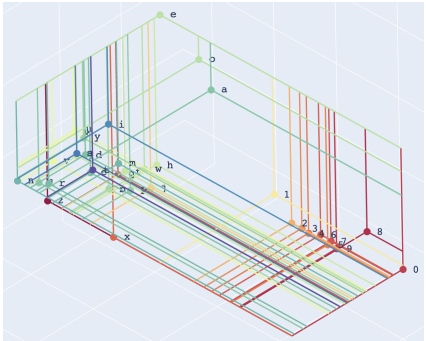
$\mathcal{M}_* \mathcal{M}^*$



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

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

Structure

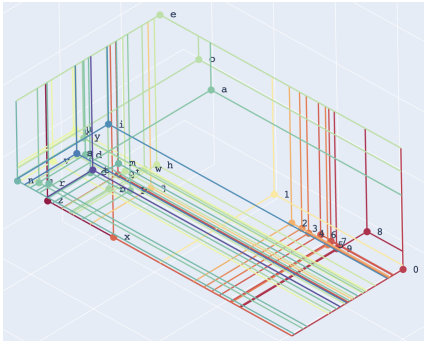


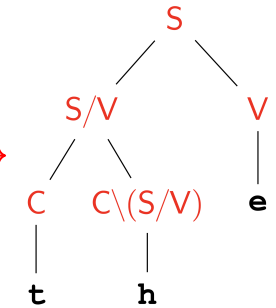
$$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \longrightarrow \bar{\mathbb{R}}$$

$$\Downarrow$$

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

Structure





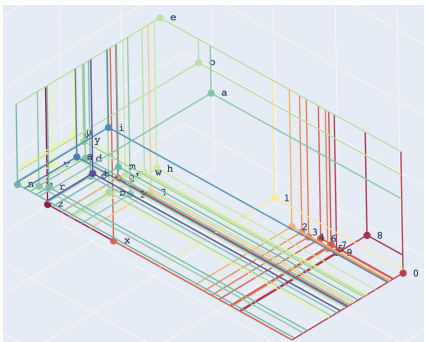
$$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \longrightarrow \bar{\mathbb{R}}$$

$$\Downarrow$$

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

Formal Explainability

Structure

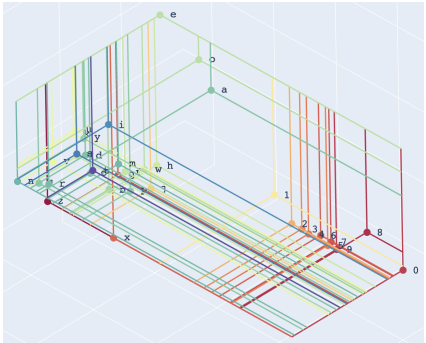


Embedding

Nucleus and Types

Formal Explainability

Structure



Tokenization
Renormalization

Embedding
Nucleus and Types

Attention
Tensor Product

Outline

Intro: Critique and Formalism

Epistemological Critique: LLMs as Formal Objects

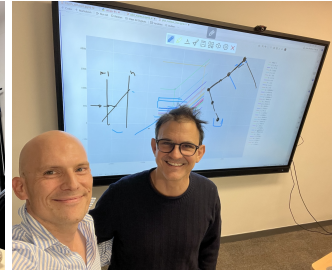
Theoretical Critique: Formal Explainability
- The Algebra Behind the Embeddings
- The Structure Behind the Algebra
- The Categories Behind the Structure

Conclusion

Conclusion: For a Critical Formalism

- It is urgent to address the epistemological dimension of the critical project in its own terms
- This requires to develop a critical approach within formal sciences where formalization is not assumed to lead to naturalization
  - The new role of data within formal sciences is crucial in this sense
- A critical formalism will be incomplete if it remains disconnected from the political, and even the artistic dimension of the critical program
  - We need a new alliance between the formal sciences, the human sciences, and the arts.

Collaborations



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

Références I

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021).On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bender, E. M., & Koller, A. (2020).Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

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.

Church, A. (1936).An unsolvable problem of elementary number theory. *American Journal of Mathematics*, *58*(2), 345–363.

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

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

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

Gödel, K. (1934). On undecidable propositions of formal mathematical systems. In *Collected works* (pp. 346–371). Clarendon Press Oxford University Press.

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

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

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

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.

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.

Kirschenbaum, M. (2023). *Again theory: A forum on language, meaning, and intent in the time of stochastic parrot*. https://critinq.wordpress.com/2023/06/26/again-theory-a-forum-on-language-meaning-and-intent-in-the-time-of-stochastic-parrots/

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.

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.

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

Nietzsche, F. (1873). On truth and lying in a non-moral sense [Originally unpublished; written in 1873.]. (R. Speirs, Trans.). In R. Geuss & R. Speirs (Eds.), *The birth of tragedy and other writings* (pp. 141–153). Cambridge University Press.

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.

Références III

Turing, A. (1937).On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *s2-42*(1), 230–265. https://doi.org/10.1112/plms/s2-42.1.230

Underwood, T. (2023, October 15). *The empirical triumph of theory* [Accessed: 2023-10-15]. https://critinq.wordpress.com/2023/06/29/the-empirical-triumph-of-theory/

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

Chat Token Vector
Università Ca' Foscari
Venice, Italy

Toward a Critical Formalism

Philosophical and Theoretical Effects of a Mathematical Critique of LLMs

Juan Luis Gastaldi

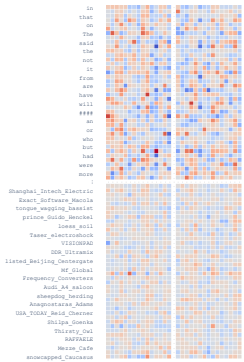
www.giannigastaldi.com

ETH zürich

June 12, 2025

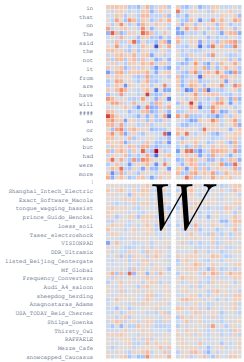
word2vec as Implicit Matrix Factorization

(Levy and Goldberg, 2014)



word2vec as Implicit Matrix Factorization

(Levy and Goldberg, 2014)



W

$C?$

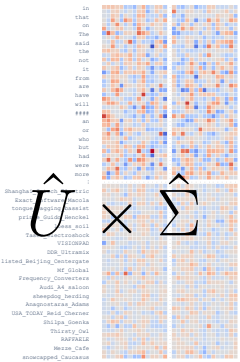
$M =$

$\text{pmi}(w, c) - \log k$

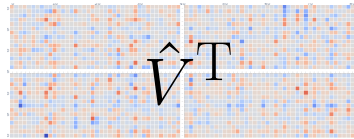
$W \times C \approx M$

Word Embeddings as Truncated SVD

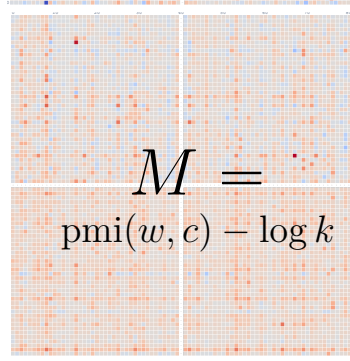
(Levy and Goldberg, 2014)



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



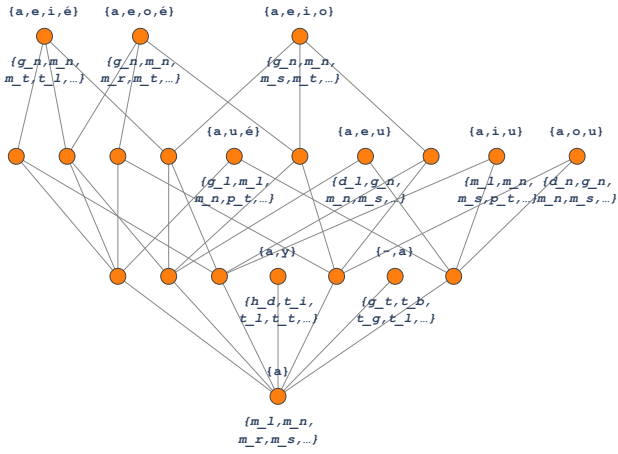
$$\hat{V}^{\mathrm{T}}$$

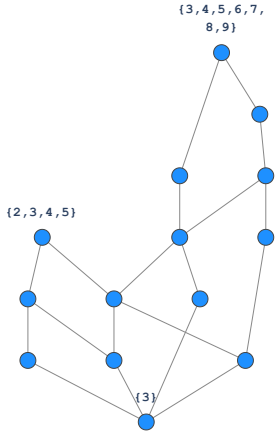


$$M = \mathrm{pmi}(w, c) - \log k$$

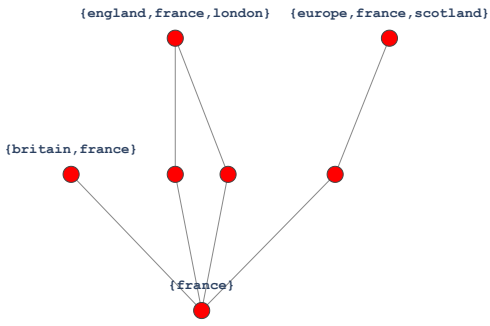
$$M \approx \hat{U} \times \hat{\Sigma} \times \hat{V}^{\mathrm{T}}$$

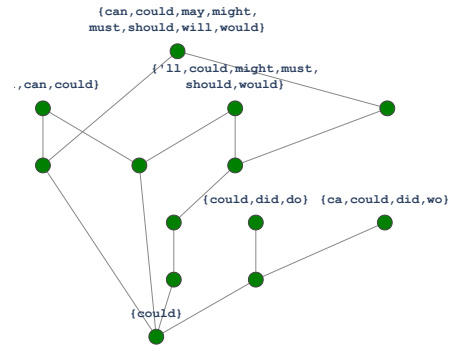
Formal Concepts



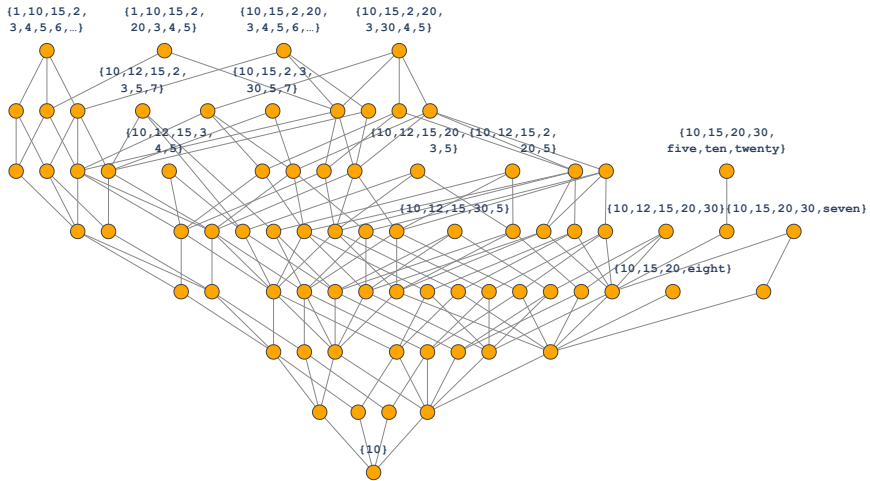


Formal Concepts (words)



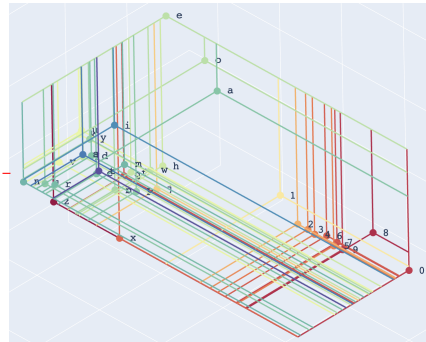


Formal Concepts (words)



Theoretical Interpretability

Structure



Theory

"Task"

?

Theoretical Interpretability

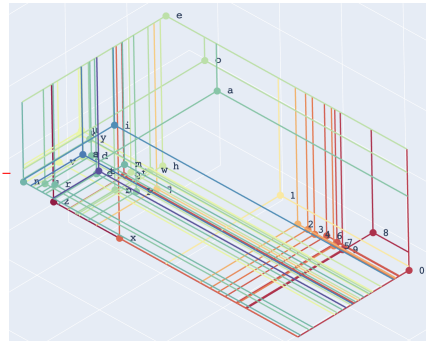
$$C^{op} \times D \rightarrow \bar{\mathbb{R}}$$

Distributional Hypothesis

The content of linguistic units is determined by their *distribution* in a corpus.

Theory
"Task"

Structure



Theoretical Interpretability

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

Theory
"Task"

Distributional Hypothesis

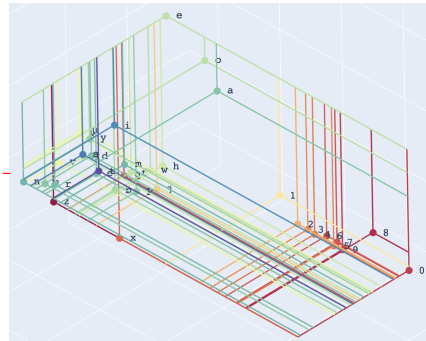
The content of linguistic units is determined by their *distribution* in a corpus.

Structuralist Hypothesis

Linguistic content is the effect of a virtual *structure* underlying linguistic practices within a community

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

Structure

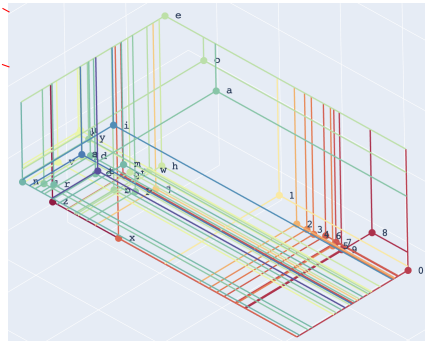


Theoretical Interpretability

Units
Classes
Relations

Semantics
Syntax
Morphology
Phonology

Structure

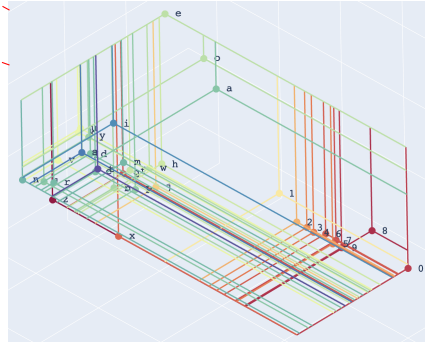


Theoretical Interpretability

Units
Classes
Relations

Semantics
Syntax
Morphology
Phonology

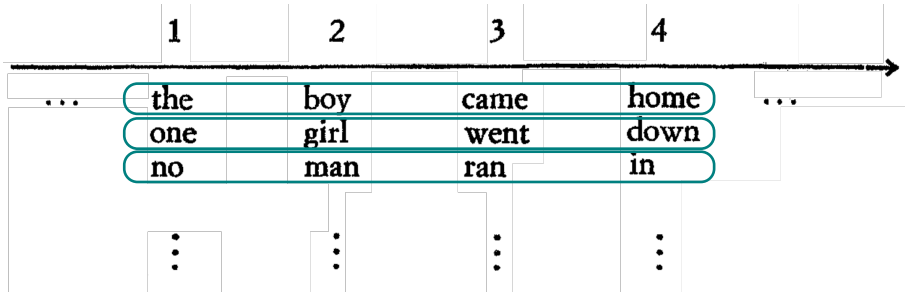
Structure



	o	a	e	u	ɑ	i	ŋ	ʃ	k	ʒ	ð	q	m	f	p	v	b	n	s	θ	t	z	d	h	ŋ
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									+	−		+	−							+	−	+	−		

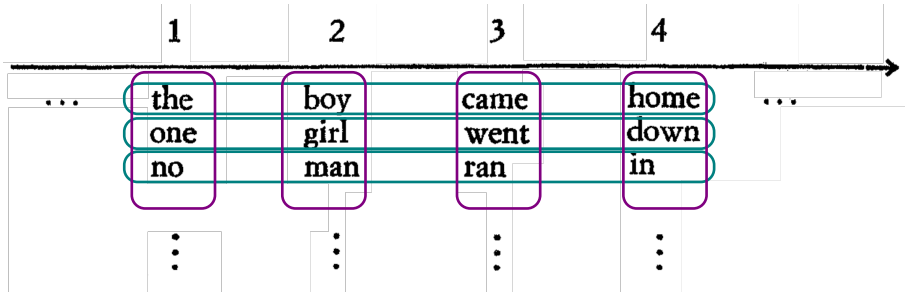
(Jakobson et al., 1952)

Syntagmes et Paradigmes



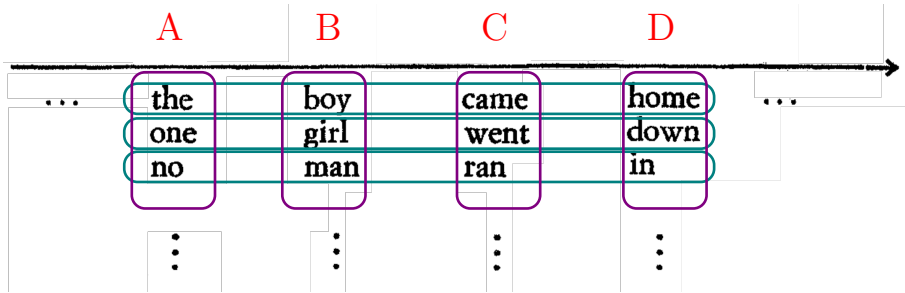
(Hjelmslev, 1971)

Syntagmes et Paradigmes



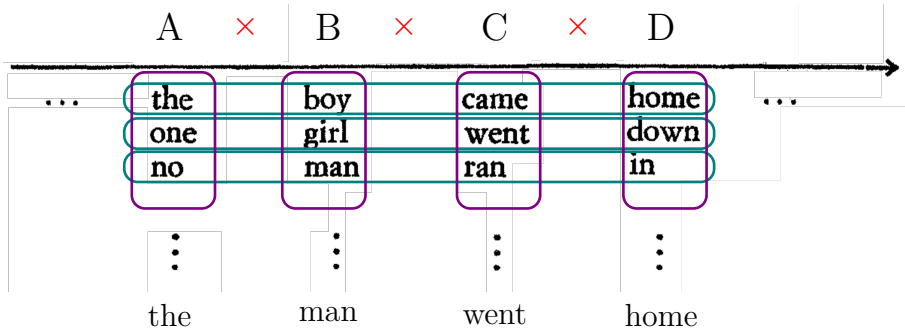
(Hjelmslev, 1971)

Syntagmes et Paradigmes

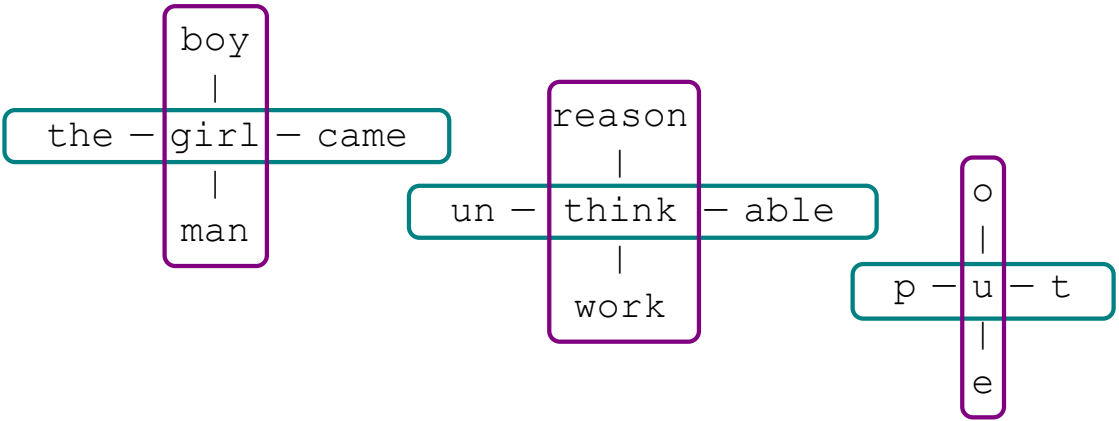


(Hjelmslev, 1971)

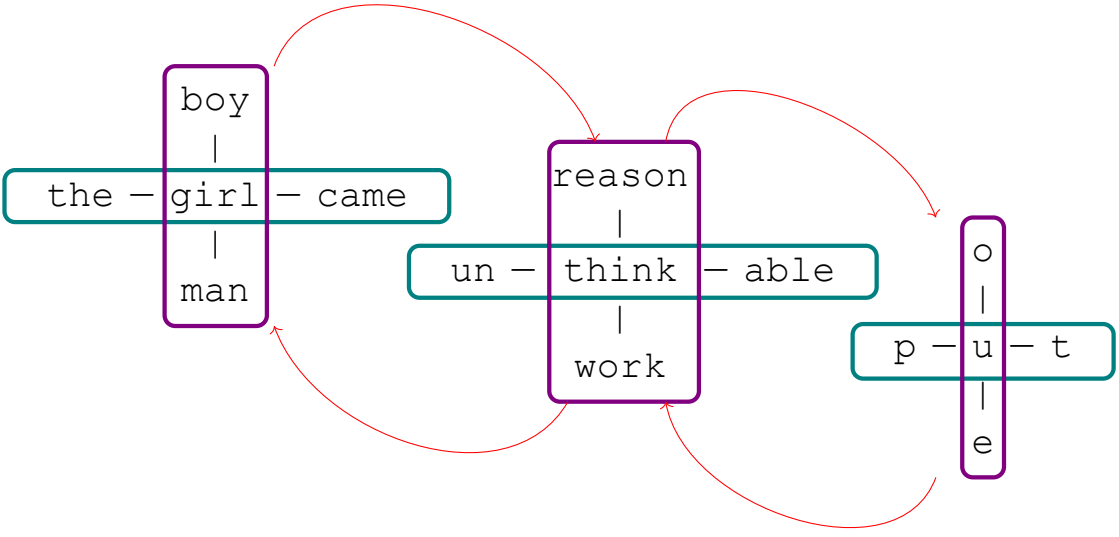
Syntagmes et Paradigmes



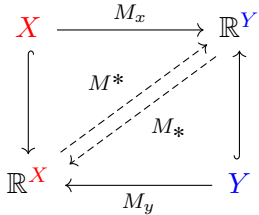
(Hjelmslev, 1971)

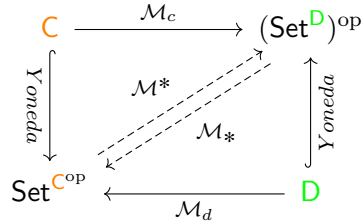

boy
the — girl — came
man
reason
un — think — able
work
o
p — u — t
e

Stratification



De l'algèbre linéaire aux catégories





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

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

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

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

$\mathcal{M}_*\mathcal{M}^* \colon \mathsf{Set}^{\mathsf{C}^{\mathrm{op}}} \to \mathsf{Set}^{\mathsf{C}^{\mathrm{op}}}$

$\mathcal{M}^*\mathcal{M}_* \colon (\mathsf{Set}^{\mathsf{D}})^{\mathrm{op}} \to (\mathsf{Set}^{\mathsf{D}})^{\mathrm{op}}$

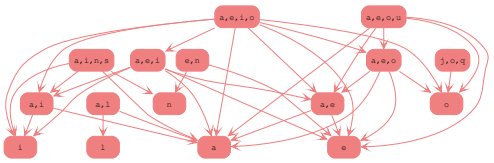
$\mathrm{Fix}(\mathcal{M}_*\mathcal{M}^*) := \{f \in \mathsf{Set}^{\mathsf{C}^{\mathrm{op}}} | \mathcal{M}_*\mathcal{M}^*(f) \cong f\}$

$\mathrm{Fix}(\mathcal{M}^*\mathcal{M}_*) := \{g \in (\mathsf{Set}^{\mathsf{D}})^{\mathrm{op}} | \mathcal{M}^*\mathcal{M}_*(g) \cong g\}$

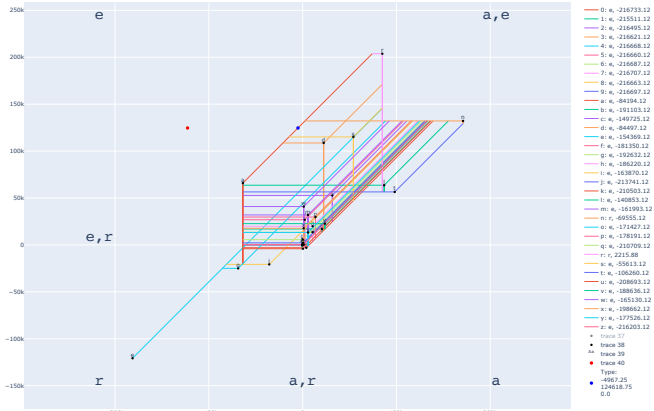
Structures catégoriques

Algèbre linéaire

$\mathcal{M}^*\colon 2^{C^{\mathrm{op}}} \leftrightarrows (2^D)^{\mathrm{op}}\colon \mathcal{M}_*$



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



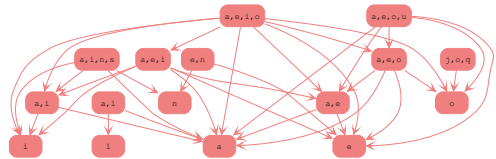
Théorie des types computationnels

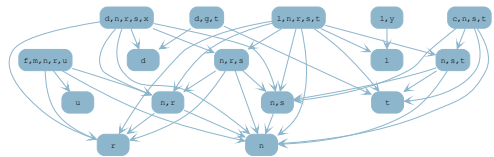
Definition (Polaire/Orthogonal - Girard, 2006)

[É]tant donnée une fonction binaire $a, b \rightsquigarrow \langle a|b\rangle \colon A \times B \to C$ 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 :

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

- 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, $X^{\perp\perp\perp} = X^\perp$.





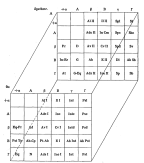
Axe 2: Interprétabilité théorique

Hypothèse distributionnelle

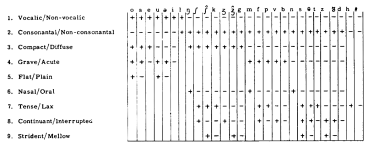
Le contenu des unités linguistiques est déterminé par leur **distribution** dans un corpus.

Hypothèse structurale

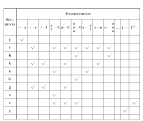
Le contenu linguistique est l'effet d'une **structure** virtuelle dérivée des pratiques linguistiques dans une communauté.



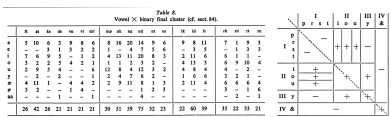
(Hjelmslev, 1935)



(Jakobson et al., 1952)



(Harris, 1960)



(Spang-Hanssen, 1959)

(Lévi-Strauss, 1949)

(Bourdieu, 1979)

(Foucault, 1966)

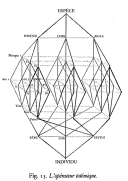
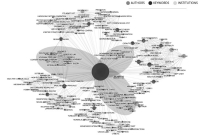


Fig. 11. *L'opération totémique.*

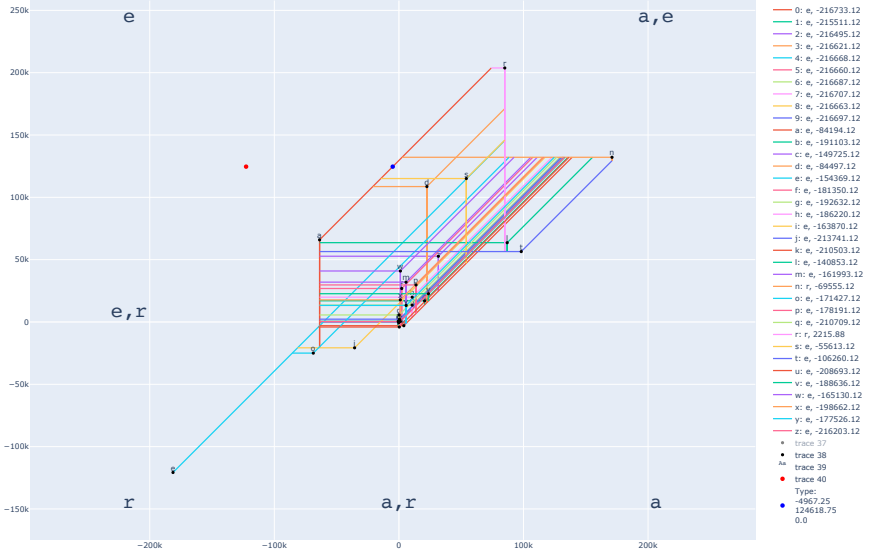
(Lévi-Strauss, 1962)

(Bourdieu, 1994)

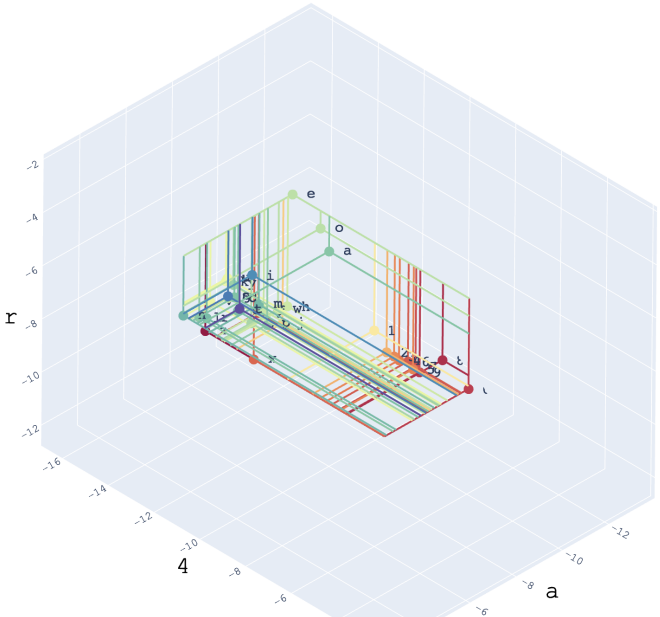


(Latour et al., 2012)

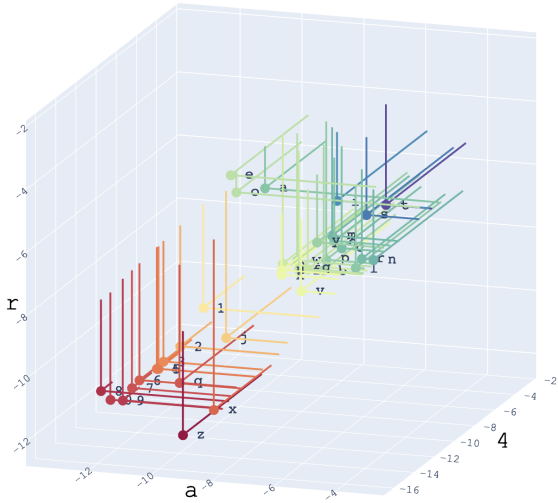
Structure interne du noyeau



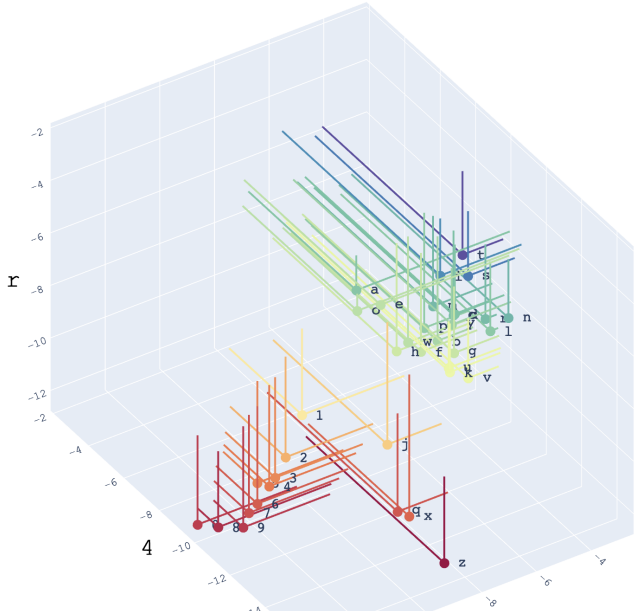
Structure interne du noyeau



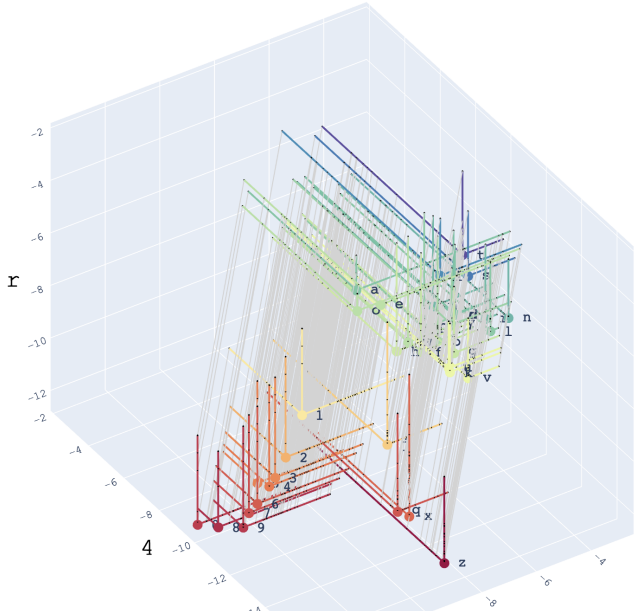
Structure interne du noyeau



Structure interne du noyeau



Structure interne du noyeau



Matrice et analogie

a = your c = my	w = apartment x = house y = chair z = stool

	…	w	x	y	z	…
…	…	0	0	0	0	…
a	…	0	1	1	0	…
b	…	0	0	1	1	…
c	…	1	0	0	1	…
…	…	0	0	0	0	…

your : house
my : apartment

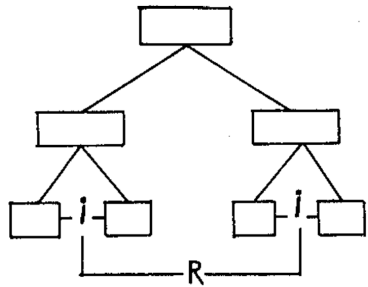
Semiotique Structurale

Une sémiotique [...] est une hiérarchie dont chacune des composantes admet une analyse ultérieure en classes définies par relation mutuelle, de telle sorte que chacune de ces classes admette une analyse en dérivés définis par mutation mutuelle.

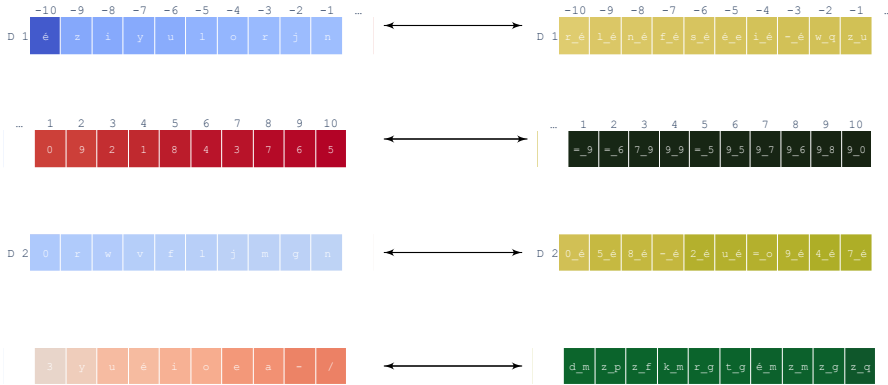
(Hjelmslev, 1975, Df. 24)

Une mutation [...] est une fonction existant entre des dérivés du premier degré d'une seule et même classe, une fonction qui a une relation à une fonction entre d'autres dérivés de premier degré d'une seule et même classe et appartenant au même rang.

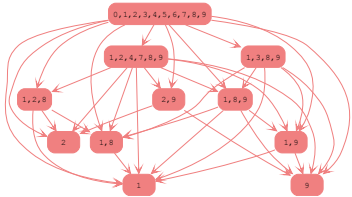
(Hjelmslev, 1975, Df. 23)

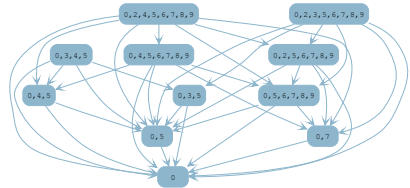


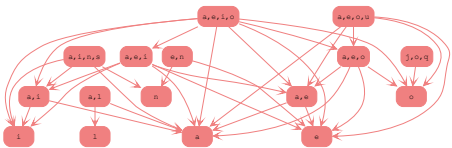
Syntagmatique et Texte (Vecteurs)

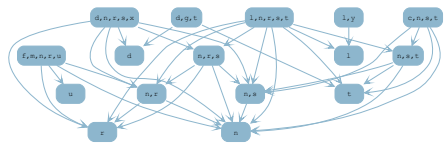


Syntagmatique et Texte (Noyeau/Types)

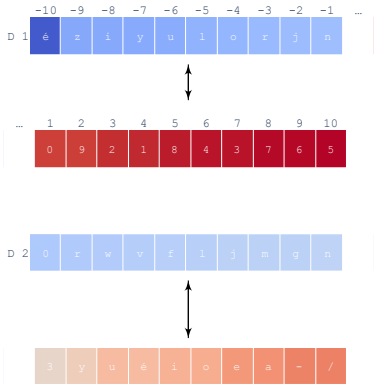


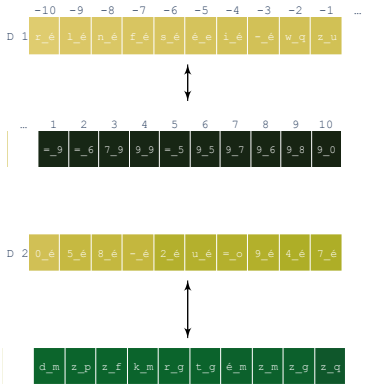




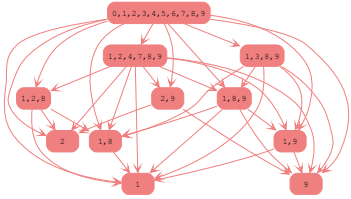


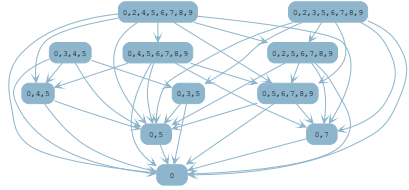
Paradigmatique et Langue (Vecteurs)

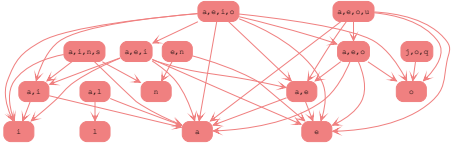




Paradigmatique et Langue (Noyau/Types)







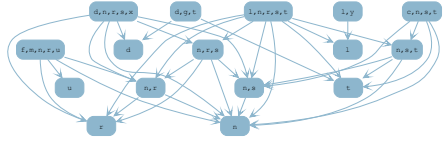


Illustration du contenu formel

(Gastaldi and Pellissier, 2021)