Le langage est-il encore le propre des humains?

Linguistique computationnelle et modèles probabilistes

Pascal Amsili

novembre 2025

Brève histoire de la linguistique computationnelle

- Cadre : Intelligence Artificielle (depuis 1950)
- Approche symbolique : explicitation des règles de la langue
- Approche non symbolique : apprentissage machine

Bascule des années 2010 :

- Approche distributionnelle :
  les mots sont des points dans l'espace
- Approche neuronale :
  méthode inspirée d'une idéalisation du neurone biologique

Problèmes de la sémantique lexicale

- Quel est le sens d'un mot ?
- Quelles relations les mots entretiennent-ils ?

Hypothèse distributionnelle

- If A and B have almost identical environments we say that they are synonyms. *(Harris, 1954)*
- You shall know a word by the company it keeps. *(Firth, 1957)*
- The degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear. *(Lenci, 2008)*

Matrice documents-termes

	QuatreVT 119 Kw	Voyage Bal 82 kw	Bête Hum. 128 kw	Bovary 117 kw
bataille	35	4	6	2
clair	105	26	96	52
facile	12	19	6	10
politique	11	0	9	5
voyage	17	196	94	44
idiot	2	1	2	6
amour	19	0	47	94

Table – Matrice documents-termes pour quelques mots et 4 romans (Quatrevingt-treize (Hugo) ; Le voyage en ballon (Verne) ; La bête humaine (Zola) ; Mme Bovary (Flaubert)).

Des romans dans l'espace

QuatreVT	(17,19)
Voyage Bal	(196,0)
Bête Hum.	(94,47)
Mme Bovary	(44,94)

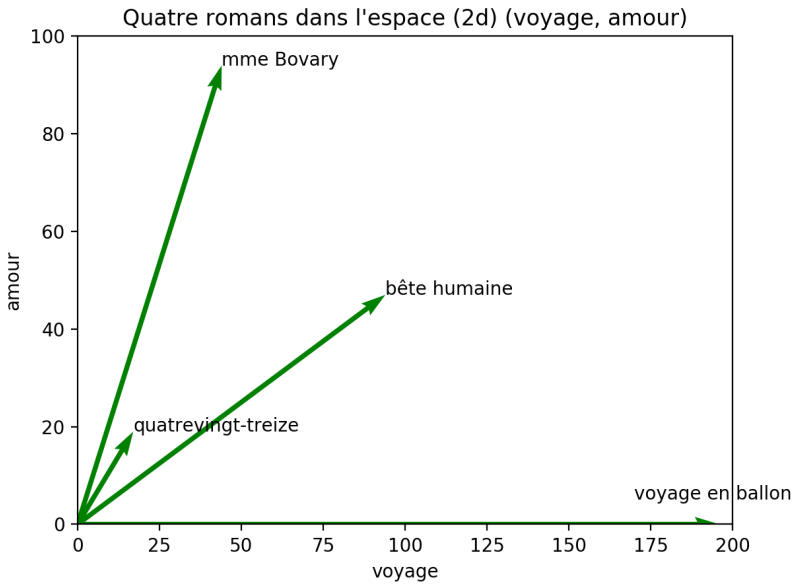


Figure – Représentation graphique de quelques romans dans le plan (voyage, amour)

Plongement dans un espace

- Système dense (il y a toujours un point entre deux points)
- Notion de déplacement (vecteur)
- Notion de distance (euclidienne)
- Notion de proximité (similarité cosinus)

Recherche d'information

Moteur de recherche (type google) :

- Tous les documents indexés sont des points dans un espace
- La requête (après traitement) peut être plongée dans le même espace
- Les documents pertinents sont les documents les plus « proches » de la requête

Inversion de la matrice

	QuatreVT 119 Kw	Voyage Bal 82 kw	Bête Hum. 128 kw	Bovary 117 kw
bataille	35	4	6	2
clair	105	26	96	52
facile	12	19	6	10
politique	11	0	9	5
voyage	17	196	94	44
idiot	2	1	2	6
amour	19	0	47	94

amour	(19,94)
bataille	(35,2)
facile	(12,10)
politique	(11,5)
voyage	(17,44)

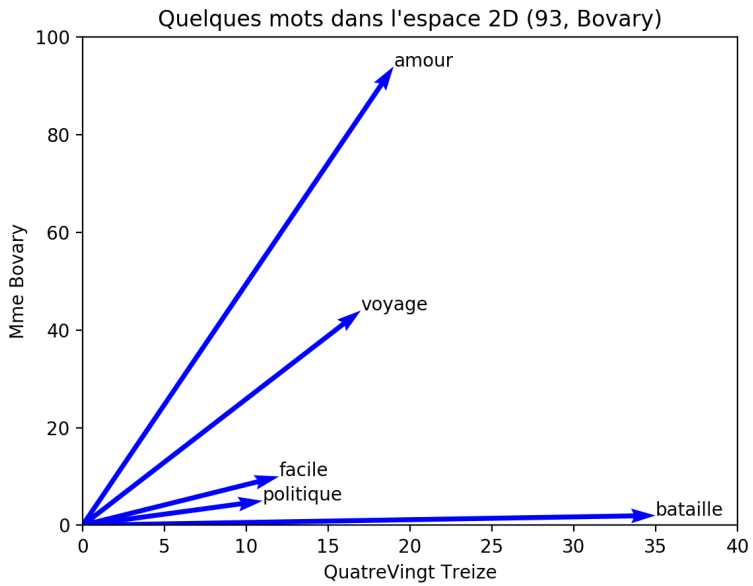


Figure – Représentation graphique de quelques mots dans le plan (93, Bovary)

Voisinages comme dimensions

je connaissais le	voyagiste	qui les amène
voyages : Différent du	voyagiste	ou tour-opérateur . En
compagnie à un	voyagiste	. Animaux domestiques : Lorsqu'
que disent certains	voyagistes	, légalement , aucune n'
Transport / Hébergement Le	voyagiste	FRAM propose les
plus vite votre	voyagiste	. Le Comité scientifique
le monde ... 6ème	voyagiste	européen , Kuoni est
autres , les grands	voyagistes	essayent de prendre
fév 2008 Ces	voyagistes	, qui opèrent principalement
parmis les meilleurs	voyagistes	en ligne . Service
clients chez un	voyagiste	- Développement d' un
les 15 principaux	voyagistes	en ligne en
ses amies des	voyagistes	locaux qui fêtent
foyers ...) et des	voyagistes	. La commercialisation Le
voyage à cheval .	Voyagiste	spécialiste du voyage
hôtel , à votre	voyagiste	, à votre ambassade

Extrait kwic de la concordance [-3,3] autour de voyagiste dans frWaC, consultation 2021-08-11

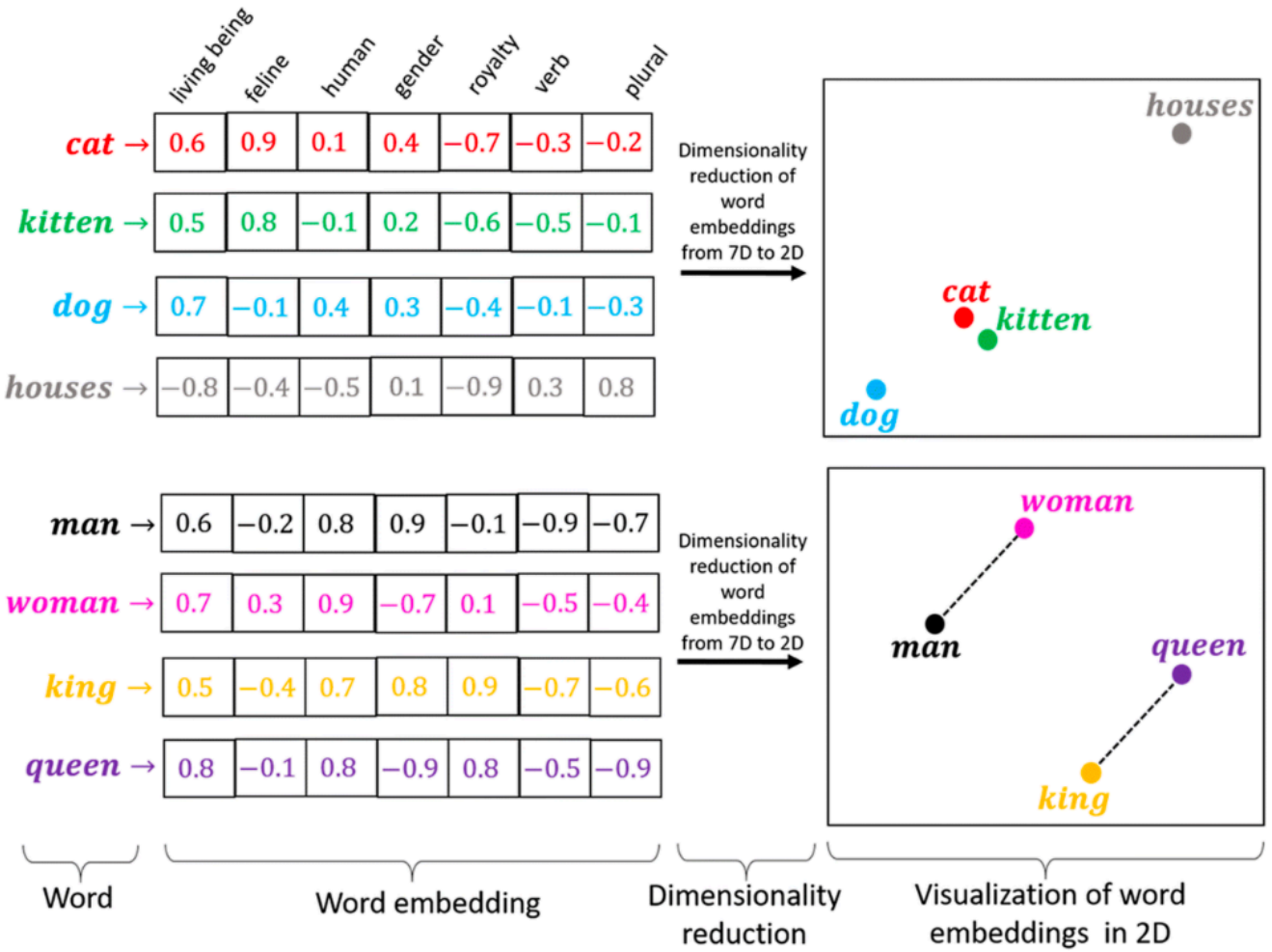
Sorbonne Nouvelle

Matrice termes-termes

	bataille	voyage	homme	femme
arriver	246	470	1 819	890
tomber	100	83	1 205	384
habiller	2	4	339	384
mourir	180	116	339	1 088
	55 331	208 520	668 289	346 093

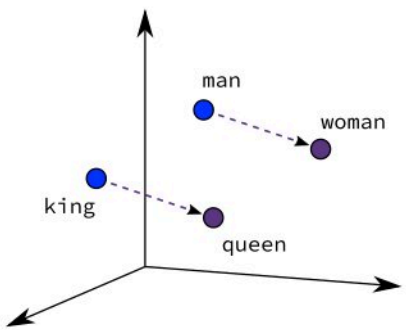
Table – Matrice Terme-Terme obtenue dans frWaC. Contextes en ligne.

Similarité

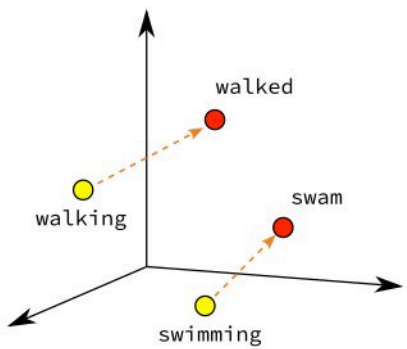


(Huang, 2025)

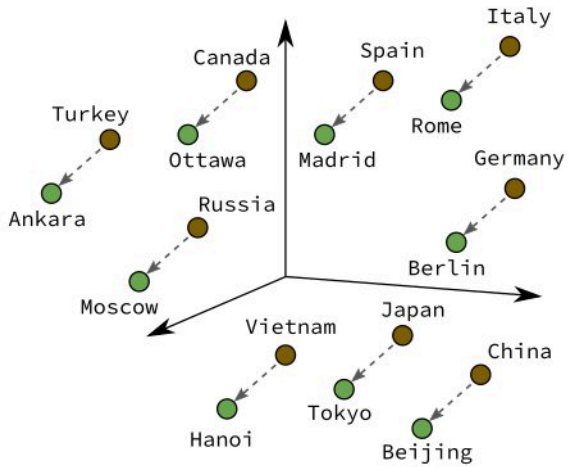
Analogie



Male-Female



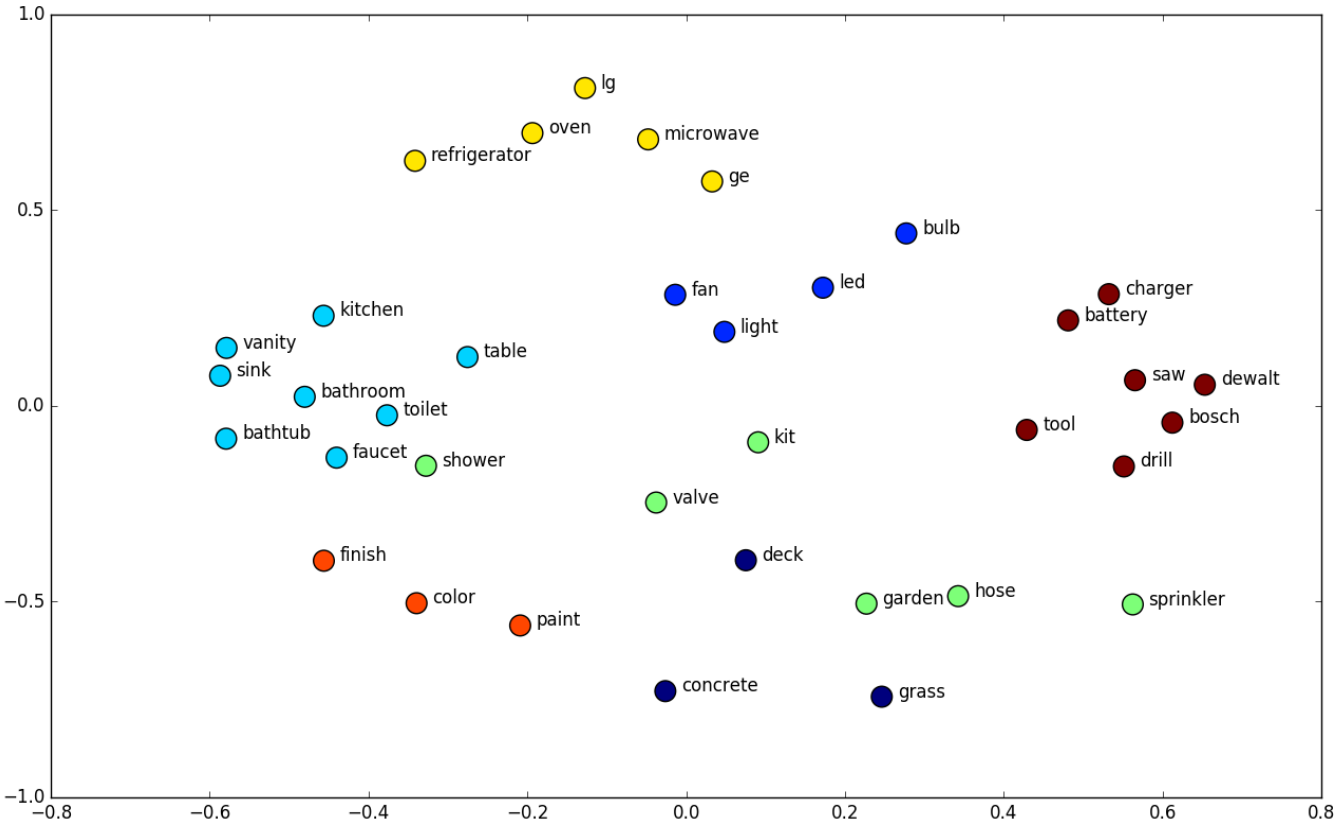
Verb Tense



Country-Capital

(Huang, 2025)

Visualisation



(Huang, 2025)

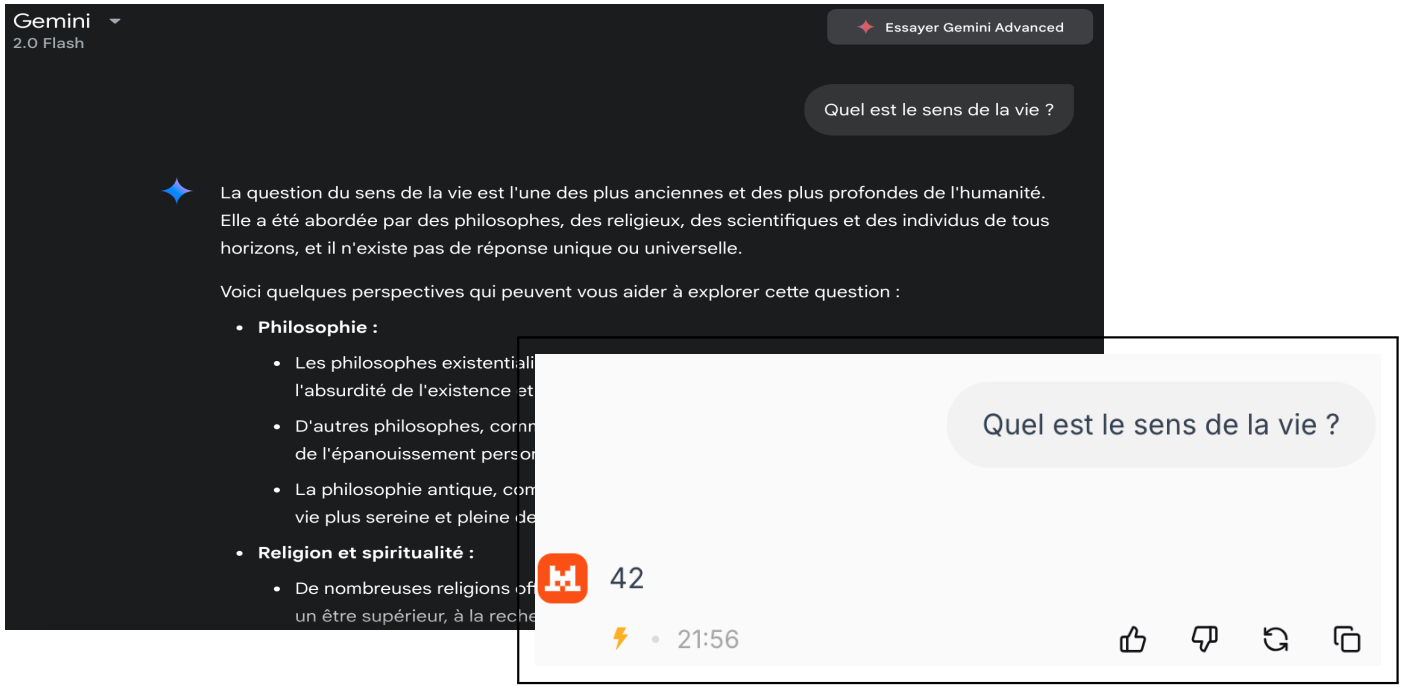
Synthèse

En représentant les mots comme des $\left\{ \begin{array}{l} \text{points} \\ \text{vecteurs} \\ \text{suites de nombres} \end{array} \right.$

on renouvelle la façon de faire de la sémantique lexicale,

et on permet de représenter les mots et les phrases de la meilleure façon possible pour l'entrée des systèmes neuronaux.

Introduction



- **Grands modèles de langue** : brique essentielle des *chatbots*

Sorbonne Nouvelle

Modèle

- **Modèle** : représentation (simplifiée) d'un phénomène réel,
- capable de **décrire** et de **prédire** des observations.
- Exemples :
  - modèles mathématiques (ex : équations de la physique)
  - modèles numériques (ex : simulations nucléaires, météo)
  - modèles conceptuels (ex : schémas, analogies)
  - modèles animaux (ex : axone de calmar, drosophile)
  - ...

Le cas particulier des modèles de langue

- Un **modèle de langue** est un modèle statistique entraîné à prédire des mots (dans un contexte).
- Il repose sur
  - un grand volume de textes (exemples)
  - un algorithme d'apprentissage (entraînement)
  - une représentation (numérique) des mots
  - une distribution de probabilités
    pour prédire un mot, il faut être capable d'attribuer une probabilité à l'ensemble des mots du vocabulaire.
- Modèle de langue ou de langage ?

Modèles de langue simples

(1) D'un effort, Jacques tenta de se

baisser
lever
moucher
tourner
dormir
camion
vivace
vraiment

Attribution d'une probabilité à tous les mots du vocabulaire
attribution d'un score à n'importe quelle phrase

Modèle n-gramme

Principe :

- Collecte et comptage des *n*-grammes dans un corpus d'"entraînement".
- Transformation des comptages en probabilités
- Utilisation du modèle :
  - mot à mot : complétion sur le téléphone
  - correction orthographique
  - aide à la reconnaissance (de caractères, de parole)
- Générateur de texte : génération auto-régressive

Exemples

amoureux
vous ayez été noyé dans le port, et qui, étant novice, l'avait eue d'une
d'oeuvre de la ville, et qui, étant novice, l'avait eue d'une vieille comtesse, qui
des étoiles, comme on sortait de table, Cunégonde et la vieille, vous ne savez pas quelle
dit Candide au baron ; pardon, mon cher Candide, qui était à côté de bras et de la
La vieille avait très bien deviné que ce monde, et qui, étant novice, l'avait eue d'une vieille comtesse
promettait déjà d'être aussi belle que vous n'ayez été violée par deux Bulgares, et qui, étant novice
Candide, si Pangloss était ici, il y a plus de trois cents lieues, sans doute, dit Candide, qui
un peu de pain ; et quand Pangloss fut refait : « Je ne vous a inspiré tant de bonté ?

Figure – Génération avec choix greedy (1er max) dans un modèle n-gramme avec $n = 3$, entraîné sur un extrait de Candide (1541 tokens)

amoureux
vous ayez été noyé dans le meilleur des mondes possibles, vous en dire davantage ; je
d'oeuvre de la beauté des tableaux.
des étoiles, comme partout ailleurs ; mais j'ai été musicien de la province, on vendit les esclaves
dit Candide au baron ; je délivrerai aisément Cunégonde.
La vieille avait très bien deviné que ce qui lui restait dans ses poches, et dit : « Si ce
promettait déjà d'être aussi belle que vous ne savez pas quelle est l'inhumanité affreuse de faire des épingles ; mais je ne
Candide, si Pangloss était ici, il fallut faire un choix ; il me fit cette horrible opération.
un peu de pain ; et quand Pangloss fut refait : « Nous allons certainement être rôtis ou bouillis.

Figure – Méthode « sampling »

Apprentissage machine supervisé

- On fournit à un modèle des couples (entrée, sortie)
- Pour chaque entrée, le modèle prédit une sortie,
- compare cette sortie prédite à la sortie attendue,
- ajuste ses paramètres pour réduire l'erreur,
- et recommence.

Ce qui compte :

- La quantité de données d'entraînement ;
- La quantité de paramètres interne du modèle ;
- Le nombre de tours d'apprentissage (*epochs*)

Modèles neuronaux

- Mots manipulés comme des vecteurs (*embeddings*)
- Réseau de neurones avec poids synaptiques : opérations sur des représentations déterminées par des matrices de poids
- Apprentissage auto-supervisé : toutes les suites de mots d'un texte sont des exemples d'apprentissage
- Tâche : prédire le mot suivant

Perroquets stochastiques

Années 2020 : architectures *transformers*, grandes quantités de données, puissance de calcul :
⇒ arrivée des premiers « perroquets stochastiques »

- produisent facilement du texte à partir de contextes (*prompts*)
- ont assez de paramètres pour produire des textes sans fautes de langue
- "recrachent" des connaissances cachées dans les textes (2)

(2) La capitale de la France est ...

Problèmes

- vraisemblable $\neq$ vrai
  (*"lack of grounding"*)
- les modèles ne sont pas spécialistes
  (affinage (*fine-tuning*))
- continuer une phrase n'est pas dialoguer
  (modèles dits *"instruct"*)
- les textes produits peuvent être biaisés, offensants, dangereux, interdits...
  (*"lack of alignment"*)
  Entraînement supplémentaire : RLHF

Etats des lieux

- Les modèles sont environnés :
  - entraînement au dialogue
  - alignement par renforcement
  - contrôle de la véracité : *retrieval augmented generation* (RAG)
  - raisonnement explicité : *chain of thought* (COT)
- Les modèles rencontrent des difficultés :
  - hallucinations
  - biais, discours toxiques
  - fuites de données
  - désinformation

N'oubliez pas : les LLM sont simplement entraînés à prédire le mot suivant !

Conclusion : leçon pour la linguiste

- Renouvellement des études sur le lexique (sémantique, morphologie, diachronie...)
- Un argument en faveur des approches usage-based
- Des limites nombreuses à prendre en considération

Merci

Références

Firth, John Rupert. 1957. *Papers in Linguistics (1934-1951)*. Oxford : Oxford University Press.

Harris, Zellig S. 1954. Distributional structure. *Word*, **10**(2-3), 146–162.

Huang, Liang. 2025. *Course in Machine Learning. Embeddings*. `https://web.engr.oregonstate.edu/~huanlian/teaching/ML/2025fall/unit4/word_embeddings.html`. Oregon State University.

Lenci, Alessandro. 2008. Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, **20**(1), 1–31.

Sorbonne Nouvelle