

# Pourquoi mon téléphone comprend le français mais pas l'alsacien?

Sam Bigeard - INRIA  
14 mai 2024



PINT OF SCIENCE #pint24

Mais d'abord

## Quizz sur les langues du monde



Combien de langues dans le monde?

...  
environ 7000

Combien ont moins de 1000 locuteurs?

...  
environ 40%

(les 23 langues les plus parlées = la moitié de la population mondiale)

Combien ont une écriture?

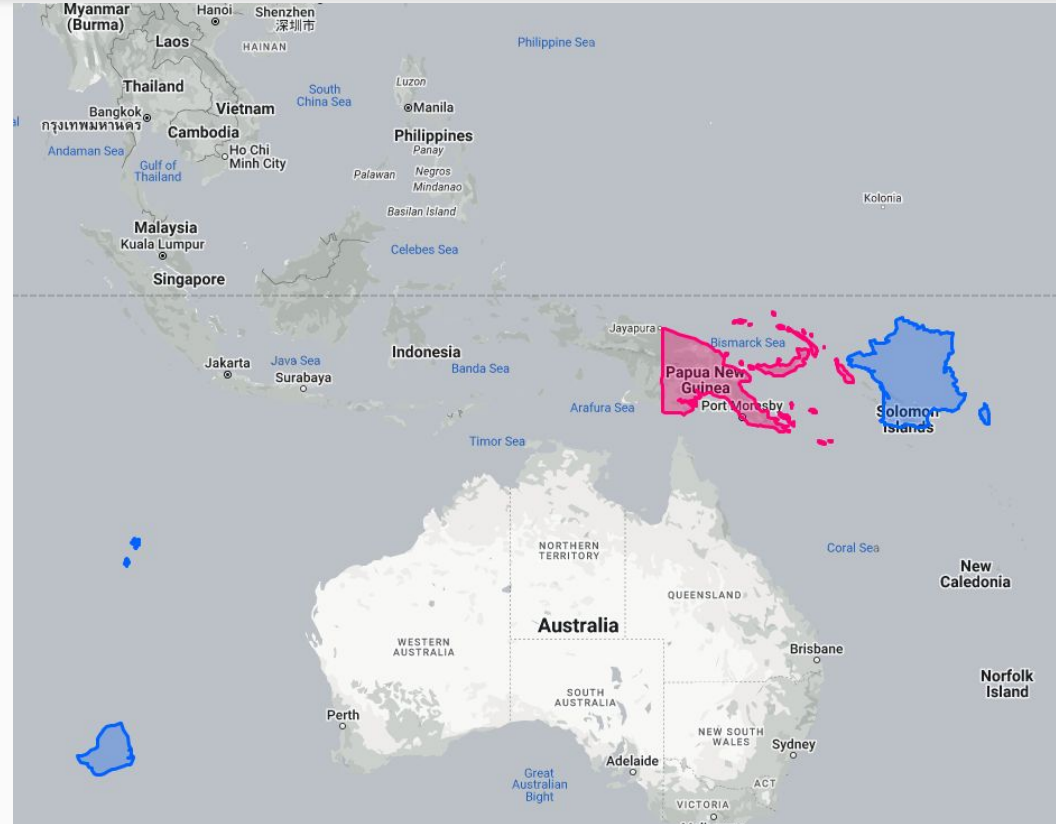
...  
environ la moitié



Le pays où on parle le plus  
de langues en contient  
combien?

...

840 langues  
en Papouasie Nouvelle  
Guinée



Combien de langues des signes dans le monde?

...  
environ 300




Combien de locuteurs des langues des signes dans le monde?




...  
72 millions  
(population française = 67 millions)

4% des français de moins de 20 ans ont une déficience auditive



# C'est quoi les technologies linguistiques?

- Clavier prédictif
- Transcription de la voix 
- Correction orthographique
- Synthèse de la parole
- Traduction automatique 
- Génération de texte 
- ...

fais     
Je fait ce que je veux



1. Comment marchent ces technologies?
2. Quels sont les obstacles pour les langues peu dotées?



## Deux familles de méthodes

### Règles explicites

A la main

### Apprentissage

IA, machine learning, big data, réseaux de neurones...





# Comment l'ordinateur apprend la langue?

## Règles explicites :

Si un mot de cette liste est présent dans le message de l'utilisateur, alors présente lui cette pub.

- + facile à mettre en place
- + explicable et corrigeable
- marche seulement pour des tâches simples

## Exemples :

détection de propos injurieux dans les média sociaux

correction orthographique

génération de texte simple :  
météo, horoscope

"Demain, il fera \_\_\_ degrés à \_\_\_"



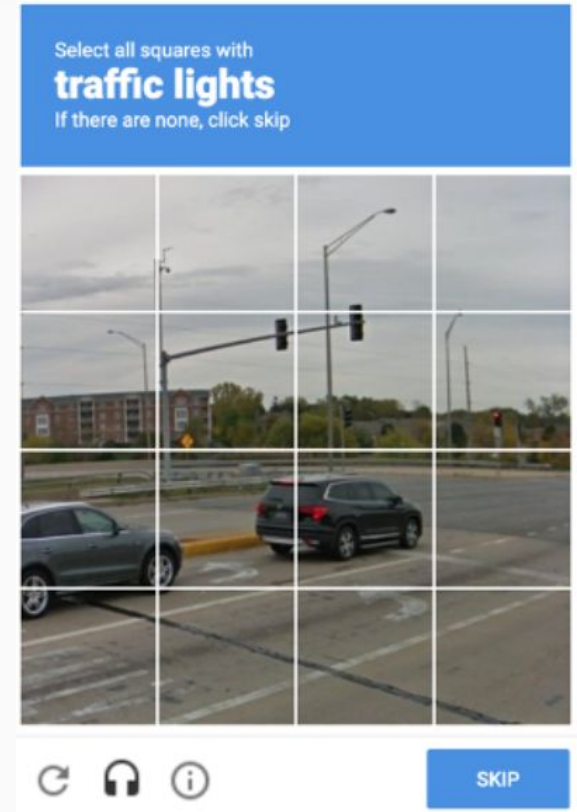
# Comment l'ordinateur apprend la langue?

## Apprentissage :

IA, machine learning, big data, réseaux de neurones...

On donne une grande quantité de données à l'ordinateur et lui dit d'apprendre par l'exemple.

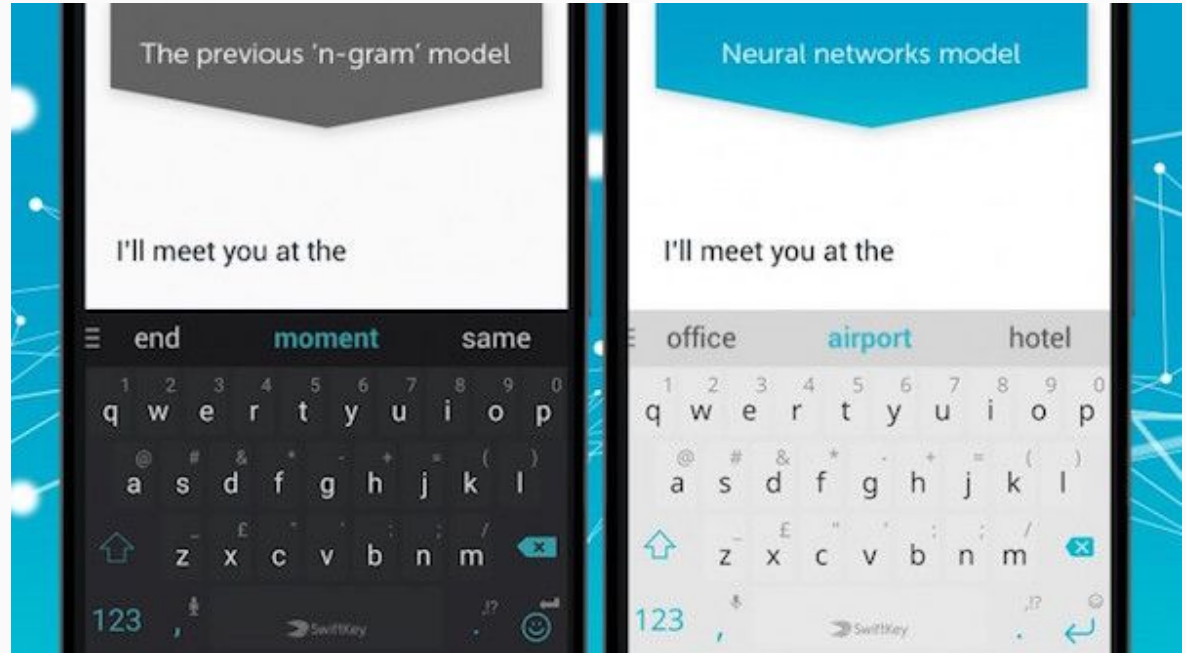
- + peut faire des tâches complexes, non explicables
- difficile à expliquer ou à corriger -> éthique
- nécessite beaucoup de données
- mauvais pour l'environnement



# Exemple de ChatGPT

ChatGPT = générateur de  
texte prédictif

Autres exemple:  
Clavier prédictif



Coût d'apprentissage de ChatGPT :

28,800,000 kW/h = consommation annuelle d'électricité de 12 500 français

6,912,000 kilos de CO<sub>2</sub> = faire rouler une voiture pendant 28 millions km, soit presque 700 fois la circonférence de la Terre

Ça coûte cher -> il faut pouvoir se faire de l'argent avec !

référence <https://www.linkedin.com/pulse/carbon-impact-large-language-models-ais-growing-cost-vaidheeswaran-fcbhc>



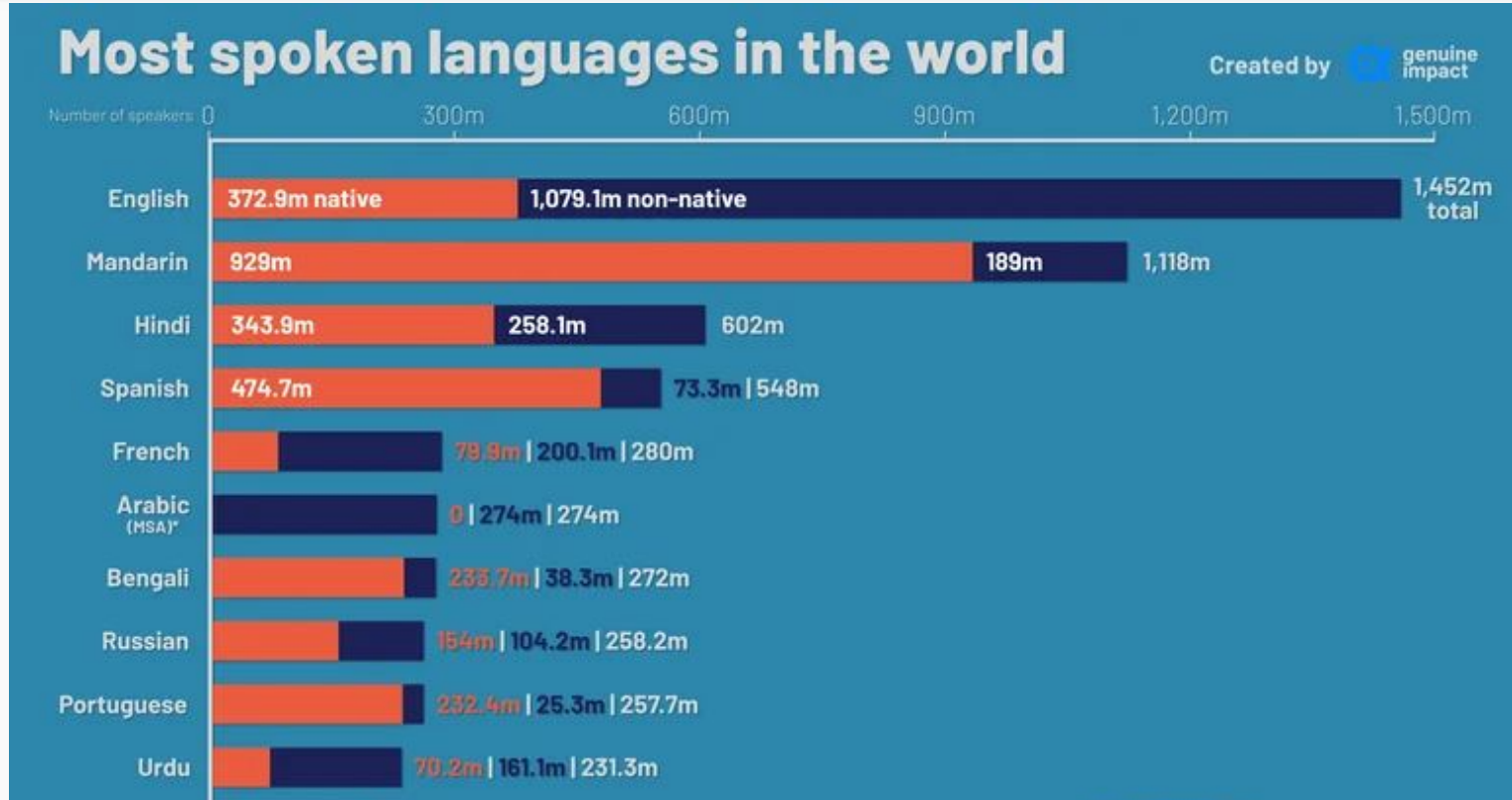
Combien de temps faut-il à un enfant pour apprendre à parler?  
à 14h par jour pendant...  
4 ans : 20 440 heures  
10 ans : 51 100 heures

Pour la machine, pour apprendre la reconnaissance vocale  
+ ou - selon vous?

...

réponse : 438 000 heures





Pour rappel:

environ 7000 langues dans le monde  
40% des langues ont moins de 1000 locuteurs

et combien de langues en France ?  
combien de locuteurs ?



26 sur la carte sans aller plus loin dans la distinction des dialectes

## LES LANGUES RÉGIONALES EN FRANCE MÉTROPOLITAINE







créoles à base  
française





environ  
10 langues natives en  
Guyane

29 en  
Nouvelle-Calédonie



et LSF !



## Quelques langues parlées en métropole

Langue	Nombre de locuteurs	Pourcentage des habitants du territoire qui sont locuteurs
Alsacien	660 000	38.8%
Langues d'Oï (Picard, Normand, Poitevin, etc)	570 000	1.6%
Breton	280 000	18.6%
Catalan	110 000	29.7%
Franco-provençal	80 000	1.3%
Corse	70 000	28.0%
Basque	50 000	38.8%



Locuteurs alsaciens : 660 000 personnes  
Heures d'entraînement Whisper : 680 000 heures

même si tous les locuteurs s'enregistraient 1 heure,  
ce ne serait pas suffisant



On les trouve comment ces données?

- Se servir sur internet (crawl)
- Être un hébergeur de données (google, meta...)
- Crowdsourcing (volontaire ou forcé)
- Archives institutionnelles (TV, journaux...)
- Enquêtes linguistiques



- Plus de variance, moins de standardisation :
  - Orthographe etc non standardisée
  - Moins de média de masse
- Méfiance des locuteurs
  - Erreurs, mauvaise représentation
  - Standardisation, perte de richesse
  - Remplacement des professionnels



## Common Voice : collection de données vocales via crowdsourcing





7000 langues dans le monde

La majorité sont peu dotées

Les langues peu dotées ont:

- Moins de locuteurs donc
  - Moins de données
  - Moins de marché
- Plus de variation
  - Richesse mais aussi obstacle



Merci !

