# Comment une machine écrit-elle ?

## Prédire le mot suivant

**Mise en situation.** ChatGPT, Le Chat, Gemini... Ces agents conversationnels semblent écrire comme des humains : ils répondent à des questions, rédigent des textes, résument des documents. Pourtant, leur fonctionnement repose sur une idée étonnamment simple : **deviner, encore et encore, le mot qui a le plus de chances de venir ensuite**.
Dans cette activité, vous allez fabriquer, avec des enveloppes et des cartes, une petite machine capable d'écrire des phrases toute seule. Vous découvrirez ainsi ce qu'est un **modèle de langage**, et pourquoi une telle machine peut écrire des phrases très crédibles... sans rien comprendre à ce qu'elle écrit.

## 1. Le jeu du mot suivant

Vous jouez déjà à ce jeu tous les jours : en tapant un message, le clavier de votre téléphone vous propose les mots qui ont le plus de chances de suivre.




**▶ À vous de jouer !**

Complétez chaque début de phrase par le premier mot qui vous vient à l'esprit, **sans réfléchir** :

1. Joyeux .....
2. Il pleut, je prends mon .....
3. Le chat attrape la .....
4. Ce midi, à la cantine, j'ai mangé .....

1. Comparez vos réponses avec celles de vos voisins. Pour quelles phrases tout le monde a-t-il écrit le même mot ? Pour lesquelles les réponses sont-elles très variées ? Comment l'expliquer ?

**Définition**

Un **modèle de langage** est un programme qui, à partir des mots déjà écrits, estime la **probabilité** de chaque mot possible pour la suite. Il choisit un mot, l'ajoute au texte, puis recommence avec le mot suivant, et ainsi de suite. C'est le principe de base des agents conversationnels comme ChatGPT.

## 2. Apprendre à partir d'un corpus

**Définition**

Un **corpus** est l'ensemble des textes à partir desquels la machine apprend. Notre machine va simplement compter, dans le corpus, quels mots viennent juste après chaque mot : on les appelle les **successeurs** de ce mot.
Conventions : on écrit tout en minuscules, le point « . » est considéré comme un mot (il marque la fin de la phrase), et on note **début** le début de chaque phrase.

**Notre corpus : 7 phrases**

**P1.** le chat mange la souris .
**P2.** le chat mange la viande .
**P3.** le chat dort sur le tapis .
**P4.** le chien chasse la souris .
**P5.** le chien mange la viande .
**P6.** la souris mange le fromage .
**P7.** la souris dort sur le lit .

**Exemple**

**Les successeurs du mot le.** On repère chaque le du corpus, et on note le mot qui le suit :

- P1, P2 et P3 : le chat P3 : le tapis P4 et P5 : le chien P6 : le fromage P7 : le lit.

On obtient donc : chat (3 fois), chien (2 fois), tapis (1 fois), fromage (1 fois), lit (1 fois), soit $3+2+1+1+1=8$ successeurs au total.

**▶ À vous de jouer !**

Compléter le tableau des successeurs (les deux premières lignes sont déjà remplies).

| Mot | Successeurs (et nombre de fois) | Total |
|---|---|---|
| début | le (5) la (2) | 7 |
| le | chat (3) chien (2) tapis (1) fromage (1) lit (1) | 8 |
| la | | |
| chat | | |
| chien | | |
| souris | | |
| mange | | |
| dort | | |
| chasse | | |
| sur | | |
| viande, tapis, fromage, lit | | |

**Coup de pouce**

Pour le mot la, n'oubliez pas les phrases P6 et P7, qui **commencent** par la souris. Pour vérifier votre tableau : en additionnant tous les totaux, on doit retrouver le nombre total de mots du corpus (points compris), c'est-à-dire 44.

## 3. Des effectifs aux probabilités

**Définition**

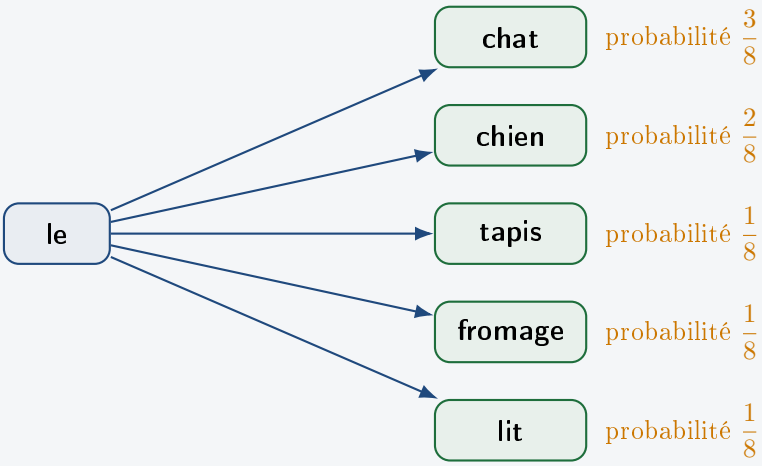
Après un mot donné, la **probabilité** qu'un successeur soit choisi se calcule ainsi :

$$\text{probabilité} = \frac{\text{nombre de fois où ce successeur apparaît}}{\text{nombre total de successeurs du mot}}$$

**Exemple**

Après le mot **le**, il y a 8 successeurs au total, dont 3 fois **chat**. La probabilité que le mot suivant soit **chat** est donc :

$$\frac{3}{8} = 0{,}375$$




On vérifie que la somme des probabilités vaut 1 : $\frac{3}{8} + \frac{2}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{8}{8} = 1$.

**▶ À vous de jouer !**

À l'aide de votre tableau des successeurs :

1. Calculer la probabilité que le mot qui suit **la** soit **souris**. Donner le résultat sous forme de fraction simplifiée.
2. Calculer la probabilité que le mot qui suit **mange** soit **le**.
3. Après le mot **souris**, quel est le successeur le plus probable ? Avec quelle probabilité ?
4. Après le mot **sur**, combien de successeurs différents sont possibles ? Que vaut la probabilité de **le** ?

## 4. Une machine à écrire... avec des enveloppes

Le principe ressemble à celui des gobelets du jeu de Nim : **chaque enveloppe correspond à un mot**, et elle contient les cartes des mots qui peuvent venir **juste après** lui.

## Fabriquer la machine

- Sur chaque enveloppe, on écrit un mot du tableau : **début**, **le**, **la**, **chat**, ..., **lit** (14 enveloppes).
- Dans l'enveloppe d'un mot, on met les cartes de ses successeurs, en recopiant **sa ligne du tableau**. Par exemple, la ligne **la** indique **souris** (4) et **viande** (2) : on met 4 cartes « souris » et 2 cartes « viande » dans l'enveloppe « la ».




souris souris souris souris

viande viande

**6 cartes : 4 « souris » et 2 « viande »**

## Faire fonctionner la machine

1. On ouvre l'enveloppe **début** et on tire une carte **au hasard**, sans regarder.
2. On écrit le mot de la carte sur sa feuille, puis on **remet la carte** dans son enveloppe.
3. On ouvre l'enveloppe qui porte **le mot qu'on vient d'écrire**, et on tire une carte. Et ainsi de suite.
4. Quand on tire la carte « **.** », la phrase est terminée.

## Exemple

**Une phrase écrite par la machine, pas à pas.**

| Étape | Enveloppe | Cartes dedans | Carte tirée | Phrase écrite |
|---|---|---|---|---|
| 1 | **début** | 5 le, 2 la | **le** | **le** |
| 2 | **le** | 3 chat, 2 chien, 1 tapis, ... | **chien** | **le chien** |
| 3 | **chien** | 1 chasse, 1 mange | **chasse** | **le chien chasse** |
| 4 | **chasse** | 1 la | **la** | **le chien chasse la** |
| 5 | **la** | 4 souris, 2 viande | **souris** | **le chien chasse la souris** |
| 6 | **souris** | 2 « . », 1 mange, 1 dort | **.** | **le chien chasse la souris .** |

La machine ne réfléchit pas : elle regarde seulement le **dernier mot écrit** pour savoir quelle enveloppe ouvrir.

## ▶ À vous de jouer !

Par groupes, fabriquez les enveloppes et les cartes à partir de votre tableau des successeurs.

1. Faites fonctionner la machine pour écrire trois phrases. Ces phrases sont-elles toutes dans le corpus ?
2. La machine peut-elle écrire la phrase **le chat dort sur la souris .** ? Justifier.
3. Trouver une phrase que la machine peut écrire, mais qui n'est **pas** dans le corpus. Vérifier, carte par carte, qu'elle est bien possible.
4. On change la règle : au lieu de tirer au hasard, la machine choisit **toujours le successeur le plus fréquent**. Quelle phrase écrit-elle ? Que se passe-t-il si on la fait fonctionner plusieurs fois ?

**Coup de pouce**

Pour la question 2, regardez dans votre tableau quels mots peuvent suivre **sur**. Pour la question 4, partez de **début**, puis prenez à chaque fois le successeur qui a le plus grand nombre de cartes.

## 5. Une enveloppe en Python

**Exemple**

En Python, une enveloppe peut être représentée par une **liste** de cartes. La fonction `choice` du module `random` tire au hasard un élément de la liste, comme on tire une carte dans l'enveloppe.

```
from random import choice

enveloppe_la = ["souris", "souris", "souris", "souris", "viande", "viande"]
carte = choice(enveloppe_la)
print("la", carte)
```

**▶ À vous de jouer !**

1. Quels sont les deux affichages possibles de ce programme ? Lequel est le plus probable, et avec quelle probabilité ?
2. Compléter la liste qui représente l'enveloppe **mange**, à l'aide du tableau des successeurs :

```
enveloppe_mange = [...]
```

3. On exécute le programme ci-dessous. Quels sont les affichages possibles ? Avec quelle probabilité obtient-on `le chat mange la` ?

```
carte = choice(enveloppe_mange)
print("le chat mange", carte)
```

**Coup de pouce**

Dans le tableau, le mot **mange** est suivi 3 fois par **la** et 1 fois par **le** : la liste doit donc contenir 4 éléments, dont 3 fois `"la"`.

**Remarque**

En écrivant une liste pour chacune des 14 enveloppes et en répétant le tirage avec une boucle, on obtiendrait un programme qui écrit des phrases entières, exactement comme la machine à enveloppes.

## 6. Et les vrais modèles de langage ?

**Remarque**

Notre machine ne regarde que le **dernier mot** écrit. Elle peut donc produire des phrases bizarres comme le chien mange la souris mange le lit . : chaque paire de mots consécutifs existe dans le corpus, mais la phrase n'a aucun sens.
Les vrais modèles de langage fonctionnent sur le même principe (prédire le mot suivant), mais avec d'énormes différences :

- ils tiennent compte d'un très long **contexte** (des milliers de mots précédents, et pas seulement le dernier) ;
- ils apprennent sur un corpus gigantesque (des milliards de textes issus du Web, de livres...) ;
- les probabilités ne sont pas obtenues en comptant des cartes, mais calculées par un immense **réseau de neurones** ;
- ils découpent le texte en morceaux de mots appelés **tokens**.

Comme notre machine, ils ne vérifient jamais si ce qu'ils écrivent est **vrai** : ils écrivent ce qui est **probable**. C'est pourquoi ils peuvent produire des affirmations fausses mais très crédibles, que l'on appelle des **hallucinations**.

**★ Question étoilée**

On améliore la machine en regardant les **deux** derniers mots au lieu d'un seul.

1. Dans le corpus, repérer tous les endroits où l'on trouve les deux mots mange la à la suite. Quels mots les suivent ?
2. Calculer la probabilité que le mot suivant soit viande : d'abord en ne regardant que le mot la, puis en regardant les deux mots mange la. Que constate-t-on ?
3. Quels mots peuvent suivre chasse la ? Que peut-on en conclure sur l'intérêt d'un contexte plus long ?