



ENSEEIH

APPRENTISSAGE PROFOND  
RAPPORT

---

## Projet Immorthon

---

*Élèves :*

Alex BECHU  
Laerian BONTINCK  
Clément DEMAZURE  
Vianney HERVY  
Yige YANG

*Enseignant :*  
Axel CARLIER

26 mai 2025

## Table des matières

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Description du sujet</b>	<b>1</b>
<b>3</b>	<b>Données d'entraînement</b>	<b>2</b>
3.1	Corpus . . . . .	2
3.2	Définitions . . . . .	2
3.3	Scrapping . . . . .	3
3.4	Préparation des données . . . . .	3
<b>4</b>	<b>Modèle</b>	<b>4</b>
4.1	Création du modèle . . . . .	4
4.2	Entraînement . . . . .	4
4.3	Utilisation du modèle . . . . .	5
<b>5</b>	<b>Analyse des résultats</b>	<b>5</b>
5.1	Modèle non affiné . . . . .	5
5.2	Modèle affiné . . . . .	6
5.3	Sondage . . . . .	10
<b>6</b>	<b>Conclusion</b>	<b>11</b>
<b>7</b>	<b>Annexes</b>	<b>11</b>

## 1 Introduction

Lorsque nous entendons pour la première fois un mot en français, il nous est parfois possible d'en comprendre le sens à l'aide du contexte ou de l'étymologie. C'est cette seconde méthode que nous allons essayer de reproduire à l'aide d'un modèle de langage.

Les données d'entraînement, le notebook, le script de scrapping et le fichier `.tex` du rapport sont disponibles sur le dépôt GitHub public du projet <sup>1</sup>

## 2 Description du sujet

L'objectif de notre projet tient en peu de mots : *un modèle de langage capable de générer des définitions plausibles pour des mots inventés*.

Toute la difficulté tient dans le terme "définition plausible". En effet, il faut non seulement que le texte généré soit grammaticalement correct et ait la forme d'une définition, mais il faut également que le sens donné au mot soit cohérent avec sa structure étymologique. Par exemple, le mot "étymologie" est formé des radicaux grecs *ἔτυμον* (vrai sens) et *λόγος* (discours). Il est donc logique que sa définition fasse référence à l'étude du "vrai" sens des mots. De manière générale, la plupart des mots scientifiques ont des sens assez clairs pour qui a fait des langues anciennes.

1. <https://github.com/sully-vian/immorthon>

Le nom "Immorthon" est un mot-valise entre "immortel" et "python". "Immortel" est le surnom<sup>2</sup> donné aux académiciens de l'Académie Française, dont une partie de la responsabilité est de définir les mots de la langue française.

### 3 Données d'entraînement

Les données nécessaires sont sous forme de paire mot-définition. Le format JSON est parfait pour cela.

Nous avons initialement prévu un corpus de mots en français. En effet, l'orthographe des mots fait souvent explicitement référence à leur étymologie (majoritairement grecque ou latine). Cependant, Il nous fallait aussi trouver un modèle français préentraîné à affiner. Il nous est apparu finalement plus facile de partir d'un modèle anglais et de l'affiner sur un corpus anglais.

#### 3.1 Corpus

Différents corpus ont été testés : Les 3000 mots<sup>3</sup> les plus fréquents en anglais selon le site EF, une liste<sup>4</sup> de 10 000 mots utilisée par un professeur du MIT pour entraîner ses modèles et finalement un corpus de presque 500 000 mots<sup>5</sup>.

C'est ce dernier qui a donné les meilleurs résultats. À la fois parce qu'il contient le plus de mots mais aussi parce qu'il contient de nombreux mots rares, scientifiques ou techniques. Un extrait du corpus est donné au listing 1.

Sur certains corpus, il nous est apparu que la proportion de prénoms dans notre corpus était anormalement élevée. Environ 10% des définitions étaient "a first name for boys/girls". Cela a conduit le modèle à souvent reconnaître des mots comme des prénoms. Le corpus finalement choisi n'a que 1% de prénoms.

```
triphthong
serosynovial
syndesmoma
anton
intoxication
vajra
stue
loupcerviers
psychosis
dirtiest
```

Listing 1 – Extrait du corpus (généré avec `shuf -n 10 words-alpha.txt`)

#### 3.2 Définitions

Une fois le corpus choisi, il nous fallait un moyen de récupérer les définitions des mots. Nous avons choisi de les extraire du site Oxford Learner's Dictionaries<sup>6</sup>. Ce choix est dû en partie à

2. <https://w.wiki/EDk9>

3. <https://www.ef.com/wwen/english-resources/english-vocabulary/top-3000-words/>

4. <https://www.mit.edu/~ecprice/wordlist.10000>

5. <https://github.com/dwyl/english-words>

6. <https://www.oxfordlearnersdictionaries.com>

la qualité des définitions fournies, à la quantité de mots disponibles et surtout à la facilité de scrapper le site.

```
"fornicators", "a person who has sex with somebody they are not married to"
"spins", "to turn round and round quickly; to make something do this"
"carbonization", "the process of becoming or being made into carbon"
"agonies", "extreme physical or mental pain"
"militarize", "to send armed forces to an area"
"raptor", "any bird of prey"
"biding", "to stay or live in a place"
"sterility", "the fact of not being able to produce children or young animals"
"armada", "a large group of armed ships sailing together"
"ambiguities", "the state of having more than one possible meaning"
```

Listing 2 – Extrait des définitions (généré avec `shuf -n 10 dico-alpha.csv`)

### 3.3 Scrapping

La difficulté principale du scrapping est de comprendre la structure du site. Il faut trouver un algorithme qui fonctionnera pour toutes les pages de définitions. On pense que ces pages sont elles-mêmes générées automatiquement à partir d'un modèle et d'une base de données de définitions. Il est donc possible de rétro-ingénierer le modèle pour trouver comment extraire une définition d'une page.

Malgré cela, de nombreux mots du corpus choisi ont soit une page dont nous n'avons pas pu extraire la définition, soit n'ont pas de définition sur le site.

Nous nous trouvons donc avec 71010 définitions (7 fois plus que le corpus du MIT), ce qui est suffisant comme nous le verront ensuite.

Le résultat du scrapping C++ accéléré par parallélisation OpenMP<sup>7</sup> est au listing 3.

```
$ ./main corpora/words-alpha.txt dictionaries/dico-alpha.csv
OpenMP is enabled.
Fetching definitions for 370105 words...
[#####-----] 71010/370105 (19%)
71010 definitions added to dictionaries/dico-alpha.csv
299095 definitions failed to be added.
Time taken: 18255 seconds.
```

Listing 3 – Résultat du scrapping

### 3.4 Préparation des données

Le modèle doit être entraîné sur des chaînes de la forme <prompt> <réponse>. Il faut donc unifier les colonnes `word` et `definition` de nos données.

7. <https://www.openmp.org>

```
df["text"] = "Define: " + df["word"] + "\n" + df["definition"]
```

Listing 4 – Préparation des données

## 4 Modèle

La solution que nous avons choisie pour le modèle de langage est globalement très proche de celle réalisée au TP5 (Transformers pour la génération de texte).

### 4.1 Création du modèle

Nous avons essayé plusieurs modèles préentraînés différents (DistilGPT2<sup>8</sup>, GPT-2<sup>9</sup>).

Le modèle est importé depuis Hugging Face<sup>10</sup> ou depuis une sauvegarde de notre modèle déjà affiné sur les définitions. Le tokenizer associé est également importé.

On tokenise ensuite la colonne `text` de notre base de données de définitions. Nous avons choisi un padding de 200 tokens étant donné la répartition des longueurs de définitions (figure 1).

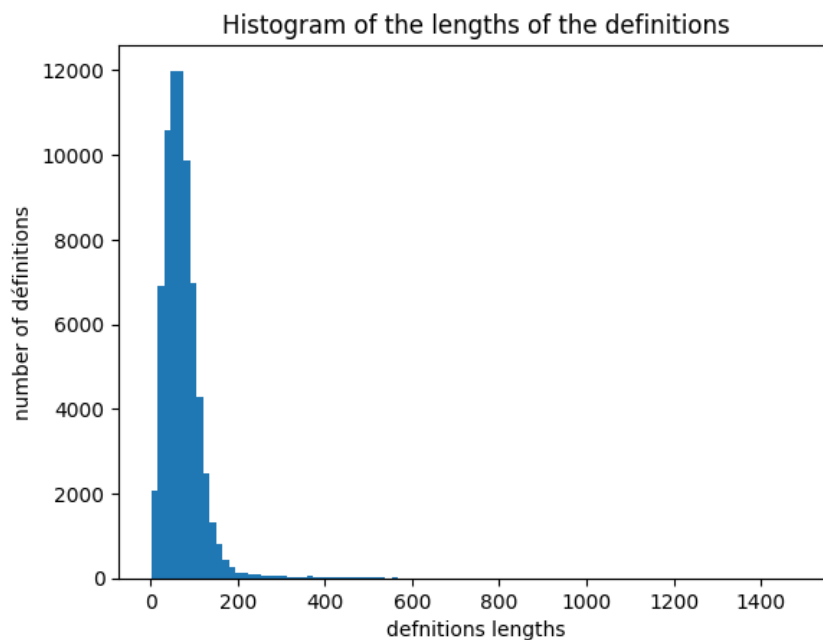


FIGURE 1 – Histogramme de la répartition des longueurs de définitions

### 4.2 Entraînement

L'entraînement est réalisé en une seule époque (ou plus si on réentraîne le modèle ultérieurement). La taille du batch est de 8 pour éviter la surcharge mémoire des GPU.

Pour le modèle GPT-2, le temps total d'entraînement est d'environ 1h.

8. <https://huggingface.co/distilbert/distilgpt2>

9. <https://huggingface.co/openai-community/gpt2>

10. <https://huggingface.co/>

### 4.3 Utilisation du modèle

Comme lors du TP5, nous utilisons ici la fonction `pipeline` de la librairie Transformers pour générer le texte. Le modèle est chargé avec le tokenizer associé.

Le modèle étant entraîné sur des exemples de la forme `Define: <mot>\n<definition>`, il faut donc préparer le prompt d'entrée (listing 5).

```
def generate(prompt, numDef):
    fullPrompt = f"Define: {prompt}\n"
    results = generator(fullPrompt, num_return_sequences=numDef, ...)
    return [result["generated_text"] for result in results]

for result in generate("zoophobia", 3):
    print(result, end="\n\n")
```

Listing 5 – Fonction de génération de texte

## 5 Analyse des résultats

En commençant ce projet nous espérions que le modèle apprenne à reconnaître des préfixes, suffixes et radicaux et à les associer à un sens précis. Cette approche est parfaitement en accord avec la notion de jetons que le modèle de langage utilise pour générer du texte.

Malheureusement, le fait de partir d'un modèle préentraîné impose le tokenizer et notre affinage ne permet pas de le modifier. Il faut donc se contenter du tokenizer du modèle de base.

### 5.1 Modèle non affiné

Pour être certain de l'efficacité de notre apprentissage, nous avons réalisé quelques tests sur le modèle GPT-2 non affiné.

Le mot `zoophobia` nous a donné les 2 résultats du listing 6. Le premier montre que le modèle croit être au milieu d'un dialogue entre A et B. Il comprend toutefois bien le thème. La seconde génération correspond déjà à ce que l'on aurait pu attendre de notre modèle affiné, mais avec une tendance encyclopédique.

```

1 Define: zoophobia
2 A: You're a zoo fag, aren't you?
3 B: No, I'm not.
4 C: I know.
5 A: I'm talking about you.
6 B: Yeah, I'm talking about it.
7 C: Well
8
9 Define: zoophobia
10 The zoophilia is a condition in which a group of people exhibit behaviors that
    ↳ are considered offensive or offensive. People with a zoophilia typically
    ↳ have difficulty moving around and are often seen in the streets or in
    ↳ parks. The zoophilia can be a sign of a poor judgment,

```

Listing 6 – Générations du modèle GPT-2 non affiné pour le mot zoophobia

Pour le mot `polyphyt` (listing 7), la première génération semble correspondre à de la documentation d'une fonction informatique. C'est amusant puisque "polyphyt" est homophone de la fonction "polyfit" de la bibliothèque NumPy. La seconde génération laisse penser que le modèle se croit dans la documentation d'un jeu vidéo dont un élément (ennemi, arme, sort...) serait appelé "polyphyt".

```

1 Define: polyphyt
2 (optional)
3 The number of times the value of a given value must be specified in any given
    ↳ condition.
4 The expression
5 $sum = "1"
6 will produce the sum of all the values specified in the condition.
7 Examples
8 This
9
10 Define: polyphyt
11 Affects: all
12 Damage Type: damage
13 Attack Type: movement speed
14 Attack Speed Modifier: 1.0
15 Attack Modifier Modifier: 0.9
16 Shield Modifier: 0.7
17 Damage Modifier: 0.

```

Listing 7 – Générations du modèle GPT-2 non affiné pour le mot polyphyt

## 5.2 Modèle affiné

Le listing 8 nous montre un exemple du modèle qui arrive à décomposer le mot en radical "dog" et suffixe "oid". On peut retrouver dans les données d'entraînement les définitions qui ont

pu influencer cette génération : *anthropoid*<sup>11</sup>, *spheroid*<sup>12</sup>, *ovoid*<sup>13</sup>

```
1 Define: dogoid
2 having the form of a dog
```

Listing 8 – Génération du modèle GPT-2 affiné pour le mot **dogoid**

Le listing 9 montre un cas similaire où le sens du mot a été obtenu en isolant des préfixes bien connus ("re-" et "anti-") et en les associant à un radical ("kiss" et "sleep"). Le modèle a donc réussi à comprendre la structure du mot et à en donner une définition cohérente.

```
1 Define: rekiss
2 to kiss somebody/something again
3
4 Define: antisleep
5 not asleep
```

Listing 9 – Génération du modèle GPT-2 affiné pour les mots **rekiss** et **antisleep**

Le listing 10 montre un cas particulier où le sens est obtenu par comparaison orthographique avec des mots des données d'entraînement. Les deux premières définitions sont inspirés par *giggle*<sup>14</sup> et la troisième par *jiggle*<sup>15</sup>

```
1 Define: siggle
2 to laugh out in a very excited way, especially when you are feeling nervous or
  ↳ have fun
3
4 Define: siggle
5 to let out a soft sound, usually because you are nervous or nervous
6
7 Define: siggle
8 to move or pull somebody/something in different directions
```

Listing 10 – Génération du modèle GPT-2 affiné pour le mot **siggle**

Le listing 11 montre un cas où le modèle a réussi à donner le sens d'un terme qui n'a rien à voir avec sa structure étymologique ou avec les mots de son corpus d'entraînement. On remarque donc que si l'IA a appris le format d'une définition, elle n'a pas perdu les connaissances apprises lors de son entraînement préalable.

11. looking like a human

12. a solid object that is approximately the same shape as a sphere

13. an object that is like an egg in shape

14. to laugh in a silly way because you are embarrassed or nervous or you think that something is funny

15. to move or make something move up and down or from side to side with short quick movements



```

1 Define: Kanye West
2 (in the US) a star or a celebrity, especially a famous name
3
4 Define: Kanye West
5 a black and brown American singer, singer, etc.

```

Listing 11 – Génération du modèle GPT-2 affiné pour le terme Kanye West

Le listing 12 montre les résultats obtenus lorsque le prompt à compléter vaut `Define: \n` (aucun mot à définir). On constate que le modèle a appris à générer des définitions même sans mot à définir. Sans rien pour le guider, il génère les définitions les plus génériques possibles. Ce n'était pas forcément prévisible, mais bien cohérent avec son entraînement.

```

1 Define:
2 a first name for girls
3
4 Define:
5 (in the past) a period of time when the people who lived in the past lived in a
  ↳ different place
6
7 Define:
8 a noun or adjective that has a different meaning

```

Listing 12 – Génération du modèle GPT-2 affiné pour une entrée vide

Le listing 13 contient des exemples de génération où l'IA a suivi une tendance encyclopédique. Il est en effet souvent difficile de définir un mot sans décrire précisément le concept qui y est associé, raconter son histoire ou expliquer son utilisation. Le problème est que le contexte complet du mot n'est pas nécessairement lié à son étymologie - au contraire de son sens. Le modèle se retrouve donc à inventer des dates, des personnages et des informations complémentaires (tâche bien plus dure que de simplement définir un mot).

```

1 Define: maxitruck
2 a large vehicle for the driver in the car. A short distance between the
  ↳ driver's side and the car's engine makes it more difficult to travel.
3
4 Define: UK
5 East Anglia
6 (in the UK in the past) the capital of England. It is the largest city, with
  ↳ its large docks and shops. It became an important centre in the 18th and
  ↳ early 19th centuries. It was the capital of the British Invasion in the
  ↳ 18th century and

```

Listing 13 – Génération du modèle GPT-2 affiné pour les mots maxitruck et UK

Le listing 14 montre un cas particulier de ce qui est présenté au dessus. Avec cette fois la spécificité que le modèle a "puisé" dans les données de son préentraînement pour relier le

mot "adam" au personnage de Genèse. On a l'impression d'interroger quelqu'un qui a perdu la mémoire, mais qui accède à des souvenirs très flous et morcelés.

```

1 Define: adam
2 (16-17) a British man who was sent to fight with God in the Middle Ages. He
  ↳ died in the 4th chapter of the Bible with his wife. His son was the son of
  ↳ Luther.
```

Listing 14 – Génération du modèle GPT-2 affiné pour le mot adam

Le listing 15 présente un cas rare (rencontré une seule fois sur tous nos tests) où le modèle a inventé<sup>16</sup> un second mot ("blibo") afin de définir le premier ("blibol"). Il ne sait donc pas uniquement faire des liens entre un mot présenté et un mot de sa base de données mais aussi créer un mot qui serait de la même famille s'il existait.

```

1 Define: blibol
2 a British company that produces blibos and other alcoholic drinks
```

Listing 15 – Génération du modèle GPT-2 affiné pour le mot blibol

Le listing 16 réunit des exemples de génération pour des prompts non-alphabétiques (symboles, chiffres etc). Les résultats ne sont pas décevants du tout. On lit bien que le modèle puise dans les données de son préapprentissage masi respecte bien le format d'une définition.

```

1 Define: 21
2 21
3
4 Define: :)
5 to say hello, or something in a friendly way
6
7 Define: $
8 a unit of money or a unit of money paid to somebody by a business, school, etc.
9
10 Define: %
11 a unit for measuring the volume or amount of something
```

Listing 16 – Génération du modèle GPT-2 affiné pour des prompts non-alphabétiques

Malgré tous ces résultats intéressants, une part non négligeable des générations sont juste des phrases sans fin qui regroupent des termes liés au sens attendu du mot. Le listing 17 en montre quelques unes.

16. Le mot "blibo" apparaît sur le wiktionnaire avec une référence vers la page IPA. Cette page concerne l'Alphabet Phonétique International mais peut avoir été interprétée au préentraînement comme la bière IPA, reliant donc avec la définition comme alcool

```

1 Define: eye
2 the part of the eye that is not there
3
4 Define: relooping
5 to make repeated or repeated a series of repeated or repeated repeated steps
6
7 Define: naruto
8 a first name for boys, short for Jeffrey
9
10 Define: blabliblu
11 a language spoken in a North American people
12
13 Define: faker
14 a person who gives an honest answer to the question that somebody asks
15
16 Define: %
17 the number of people or things that are considered to have a higher than the
    ↳ number of people or things that are considered to have a higher than the
    ↳ number of people or things that are considered to have a higher than the
    ↳ number of people or things that are considered to have a higher than the
    ↳ number of

```

Listing 17 – Génération du modèle GPT-2 hallucinant

### 5.3 Sondage

En complément de nos analyses, nous avons testé nos définitions à l’aide d’un questionnaire que nous avons fait passer à des amis. Le but était de deviner, pour chaque définition, si elle était issue d’un dictionnaire, générée par une IA, ou inventée par nous. La majorité des participants avaient des connaissances en informatique et en intelligence artificielle.

Malgré cela, le score moyen obtenu est de 11,31/20, ce qui suggère que nos définitions sont globalement crédibles. Une analyse question par question montre que les définitions inventées par des humains sont celles qui ont le plus souvent trompé les répondants. Par ailleurs, certains mots d’argot, comme “dekko : to look at something”, ont une structure syntaxique inhabituelle, ce qui les fait passer pour des inventions, même s’ils sont authentiques.

Pour le reste des définitions, les réponses étaient assez équilibrées, autour de la moitié de bonnes réponses. Nous avons également testé ce questionnaire sur ChatGPT (version GPT-4.5), qui a répondu correctement à toutes les questions, identifiant sans erreur les définitions humaines et celles générées par IA.

Cependant, le questionnaire présente un biais : en créant nos propres définitions, nous avons tenté d’imiter le style de notre IA. Cela a pu rendre nos définitions plus confuses, ce qui expliquerait les scores plus faibles obtenus pour celles-ci. Il aurait été préférable de séparer le questionnaire en deux parties, ou de se concentrer uniquement sur les définitions générées par IA et celles issues du dictionnaire, afin de mieux évaluer leur crédibilité.

## 6 Conclusion

Ce projet nous a permis de mettre en œuvre un apprentissage supervisé original autour de la génération de définitions pour des mots inventés, en s'appuyant sur leur structure étymologique. L'idée de départ, inspirée des capacités humaines à inférer le sens des mots à partir de leurs racines, s'est avérée pertinente, mais aussi délicate à formaliser pour un modèle de langage.

L'une des principales difficultés rencontrées fut la constitution d'un corpus adéquat. Les premiers jeux de données incluaient une forte proportion de prénoms, ce qui biaisait les générations du modèle. Le choix d'un corpus plus large et plus varié, issu d'un dictionnaire anglais, a permis de réduire cet effet, même si des traces de ce biais subsistent dans certaines sorties. Le scrapping des définitions a également représenté un défi technique, que nous avons relevé grâce à un outil en C++ parallélisé avec OpenMP, malgré un taux de réussite limité (environ 19%).

Du point de vue du modèle, l'utilisation de GPT-2 affiné a donné des résultats globalement satisfaisants. Nous avons observé que le modèle réussissait parfois à décomposer les mots en préfixes, radicaux et suffixes pour en tirer du sens — comme avec *dogoid*, *rekiss* ou *antisleep*. Toutefois, les générations restent hétérogènes : certaines sont tout à fait plausibles, d'autres plus encyclopédiques ou incohérentes, et certaines dérivent vers des hallucinations ou des répétitions absurdes. Cela peut s'expliquer à la fois par les limites du tokeniseur, figé par le modèle préentraîné, et par le fait que les mots inventés ne correspondent à rien dans les données d'origine du modèle.

Plusieurs pistes d'amélioration s'ouvrent à nous. Sur le plan des données, un filtrage plus rigoureux des entrées (suppression des prénoms, validation de la nature grammaticale) pourrait renforcer la cohérence des définitions générées. En outre, constituer un corpus multilingue ou étymologiquement annoté ouvrirait des perspectives intéressantes, notamment pour entraîner le modèle à associer systématiquement racines et sens. Du côté du modèle, entraîner un tokenizer spécifique ou fine-tuner un modèle multilingue préexistant pourrait améliorer la gestion des néologismes. Enfin, entraîner un modèle de contrôle capable de noter la cohérence d'une phrase pour filtrer les générations hallucinées serait une avancée significative.

En somme, si le modèle ne parvient pas toujours à reproduire la finesse humaine dans la compréhension étymologique, il montre des signes encourageants d'apprentissage structurel. L'idée qu'un modèle puisse inventer des mots et leur donner du sens reste une piste prometteuse pour explorer les frontières entre langage, sens et imagination algorithmique.

## 7 Annexes

Quelques générations heureuses que nous n'avons pas pris le temps de commenter

```

1 Define: mythophile
2 a person who believes that stories and stories are true
3
4 Define: mythophile
5 a person who uses a lot of magic powers
6
7 Define: mythophile
8 a person who believes in the existence of gods and their teachings
9
10 Define: hypertoxic
11 producing a high level of harmful or harmful substances
12
13 Define: hopun
14 a wild tree with a long thick root with leaves often used for making clothes
15
16 Define: blibol
17 to talk in a strange way about your husband, wife or partner, especially when
   ↪ you are feeling ashamed or upset by it

```

Listing 18 – Générations heureuses du modèle GPT-2 affiné