PERLIĆ Vladana Numéro étudiant : 22213513

Titre : Analyse de l'impact d'une modification unique de mot sur le sens des requêtes reformulées en français.

Sous la direction de Ludovic Tanguy et Claire Ibarboure

Master 1 – Linguistique, Informatique et Technologies du Langage
Année 2022-2023

Table des matières

Table des figures

Table des tableaux

1. Introduction

Dans notre société actuelle, où la gestion et l'accès aux connaissances revêtent une importance cruciale, la recherche d'information occupe une place prépondérante. C'est pourquoi une question fondamentale se pose naturellement : comment les modifications apportées à une requête peuvent-elles influencer son sens et, par conséquent, l'information recherchée ? Afin d'explorer cette problématique complexe, mon mémoire se focalisera sur deux questions de recherche spécifiques :

a) Dans quelle mesure le sens d'une requête est-il modifié lorsque l'on ajoute, modifie ou supprime un seul mot ?

b) Quels sont les facteurs linguistiques et contextuels qui contribuent à ces changements sémantiques ?

En comprenant ces mécanismes, nous pourrons mieux appréhender la manière dont les mots interagissent et influencent le sens global des requêtes.

Cette étude vise à apporter des éclairages sur l'influence des modifications de mots sur le sens des requêtes et à ouvrir de nouvelles perspectives pour améliorer la recherche d'information sémantique en français.

La recherche d'information sémantique constitue le cadre dans lequel s'inscrit mon travail. Contrairement à la recherche d'information basée sur des mots-clés, qui se concentre sur la correspondance exacte entre les termes de recherche et les documents, la recherche d'information sémantique vise à comprendre le sens et le contexte des requêtes et des documents afin de fournir des résultats plus pertinents et précis (Zargayouna et al., 2015).

Mes objectifs de recherche se décomposent en deux axes principaux. Tout d'abord, je cherche à déterminer dans quelle mesure les modifications apportées à une requête affectent sa signification. Je me concentre sur les cas où un seul mot est ajouté, modifié ou supprimé dans une requête. En mesurant le rang auquel le mot ajouté/modifié/supprimé est retrouvé par le modèle de langage CamemBERT, j'évalue l'influence de ces modifications sur le sens global de la requête. En mesurant le rang, je peux déterminer si un changement dans une requête affecte de manière significative sa signification. Si le mot modifié se trouve dans un rang supérieur, cela indique que la modification peut avoir eu un effet sur la compréhension de la question. En revanche, si le mot modifié se retrouve à un rang inférieur, cela indique que la modification n'a pas significativement altéré le sens global de la requête.

Ensuite, je m'intéresse à l'identification des facteurs linguistiques et contextuels qui contribuent aux différences sémantiques entre les requêtes. Par exemple, j'examine si les différences sémantiques entre les requêtes peuvent être attribuées à des facteurs linguistiques tels que les relations lexicales, les changements formels, les domaines sémantiques et les variations morphologiques. Comprendre ces facteurs me permettra de mieux appréhender les mécanismes linguistiques sous-jacents aux changements de sens dans les requêtes.

Pour mener à bien cette recherche, j'utilise le modèle CamemBERT (Martin, Muller, Suárez, Dupont, Romary, de la Clergerie, et al., 2020), spécifiquement conçu pour la langue française. CamemBERT est un modèle de langage pré-entraîné sur une grande quantité de données textuelles en français, ce qui lui confère une compréhension fine des nuances et spécificités de la langue. Je mesure le rang auquel le mot ajouté/modifié/supprimé est retrouvé par CamemBERT, en comparant ces résultats avec une grille annotée manuellement pour identifier les tendances et les relations entre les modifications apportées aux mots et les changements de sens dans les requêtes.

Ce mémoire s'organise comme suit : dans la première partie, je présenterai l'état de l'art en ce qui concerne la recherche d'information sur le Web, les relations sémantiques et les méthodes de mesure

de similarité et distance sémantique. Ensuite, je décrirai les données utilisées dans mon étude, en mettant en évidence le processus de collecte et d'annotation. J'expliquerai ensuite en détail ma méthodologie, en mettant l'accent sur l'utilisation de CamemBERT pour mesurer la similarité et le fonctionnement de mon code.

Les résultats obtenus seront ensuite présentés et analysés, en mettant en évidence les tendances et les relations observées. Je discuterai des limites de mon étude et proposerai des perspectives de recherche futures. Enfin, dans la conclusion, je récapitulerai les principales conclusions de mon travail, fournirai une réponse à ma question de recherche et discuterai de l'apport de ma recherche ainsi que des recommandations et des perspectives pour les travaux futurs dans ce domaine en constante évolution.

Ce mémoire se concentre sur deux questions de recherche spécifiques : l'impact des modifications de mots sur le sens des requêtes et l'analyse des facteurs linguistiques et contextuels qui contribuent à ces changements sémantiques. En utilisant le modèle CamemBERT, il est possible d'évaluer comment l'ajout, la modification ou la suppression d'un mot affecte le rang de prédiction du modèle et ainsi d'évaluer son impact sur le sens global de la requête. Cette étude vise à éclairer l'influence des modifications de mots sur le sens des requêtes et à ouvrir de nouvelles perspectives pour améliorer la recherche d'information sémantique en français. En comprenant mieux comment les mots interagissent et influencent le sens global des requêtes, il devient possible d'améliorer les techniques de recherche d'information pour fournir des résultats plus pertinents et précis.

2. Recherche d'information sémantique : un état de l'art

La section suivante présente l'état de l'art dans le domaine de la recherche d'information sémantique et des modifications de mots dans les requêtes. Elle explore les concepts clés tels que la recherche d'information sur le Web, les taxonomies de reformulations de requêtes, les relations sémantiques, ainsi que les méthodes de mesure de similarité et de distance sémantique. Cette revue de littérature permet de situer mon travail de recherche dans le contexte actuel et de mettre en évidence les avancées réalisées dans ces domaines.

2.1. Recherche d'information sur le Web

La recherche d'information sur le Web consiste à interroger des bases de données et des moteurs de recherche pour trouver des informations pertinentes en réponse à une requête utilisateur. Les systèmes de recherche d'information classiques reposent sur l'indexation par les mots-clés pour représenter le contenu des documents et des requêtes (Azzoug, 2014).

Une requête est une demande d'information formulée par un utilisateur dans le but d'obtenir des résultats pertinents à partir d'une base de données ou d'un moteur de recherche. Les requêtes sont généralement composées de mots-clés et de critères de recherche spécifiques pour aider à filtrer et à organiser les résultats (Azzoug, 2014).

Cependant, les utilisateurs ne trouvent pas toujours les informations souhaitées avec leur requête initiale, ce qui conduit à des reformulations de requêtes. Ces reformulations peuvent consister à ajouter, supprimer ou modifier des mots-clés pour affiner la recherche et obtenir de meilleurs résultats (Awadallah et al., 2013; Hearst, 2009; Huang & Efthimiadis, 2009). La reformulation de requêtes est un processus par lequel un utilisateur modifie sa requête initiale pour améliorer la pertinence des résultats obtenus. Les reformulations peuvent inclure l'ajout, la suppression ou la modification de mots-clés, ainsi que l'utilisation de synonymes, d'abréviations ou de variantes orthographiques (Mothe & Pai, 2017).

Pour qu'une requête soit considérée comme reformulée, elle doit présenter des modifications significatives par rapport à la requête initiale, tout en conservant l'intention de recherche de l'utilisateur (Adam et al., 2013). Les reformulations peuvent être effectuées manuellement par l'utilisateur ou automatiquement par le système de recherche d'information (Mothe & Pai, 2017).

Nettleton (2014) illustre le cycle de recherche de l'utilisateur et de reformulation de la requête dans le contexte des moteurs de recherche dans la figure 1. Selon Nettleton (2014), le processus commence par un besoin d'information, tel que la recherche d'un hôtel à prix raisonnable dans le Maryland (l’exemple de l’auteur). L'utilisateur formule une requête initiale, telle que "hôtels Maryland", que le moteur de recherche traite en recherchant dans son index les documents contenant les termes "hôtels" et "Maryland" avec la fréquence de terme la plus élevée ou la plus grande similarité avec le contenu du texte. Le moteur de recherche prend également en compte le classement des pages, ce qui permet d'afficher les documents dont la fréquence et le classement sont les plus élevés.

Les résultats de la recherche sont classés par ordre décroissant en fonction de la fréquence combinée et de la valeur du classement des pages. Si l'utilisateur estime que les résultats ne correspondent pas à ses critères de recherche d'un hôtel à prix raisonnable, il peut affiner sa requête en la rendant plus spécifique, par exemple "hôtels à prix moyen dans le Maryland". Le moteur de recherche consulte alors à nouveau son index, récupère les documents pertinents et présente les résultats classés. Ce cycle se poursuit jusqu'à ce que l'utilisateur trouve l'information souhaitée ou décide d'arrêter sa recherche

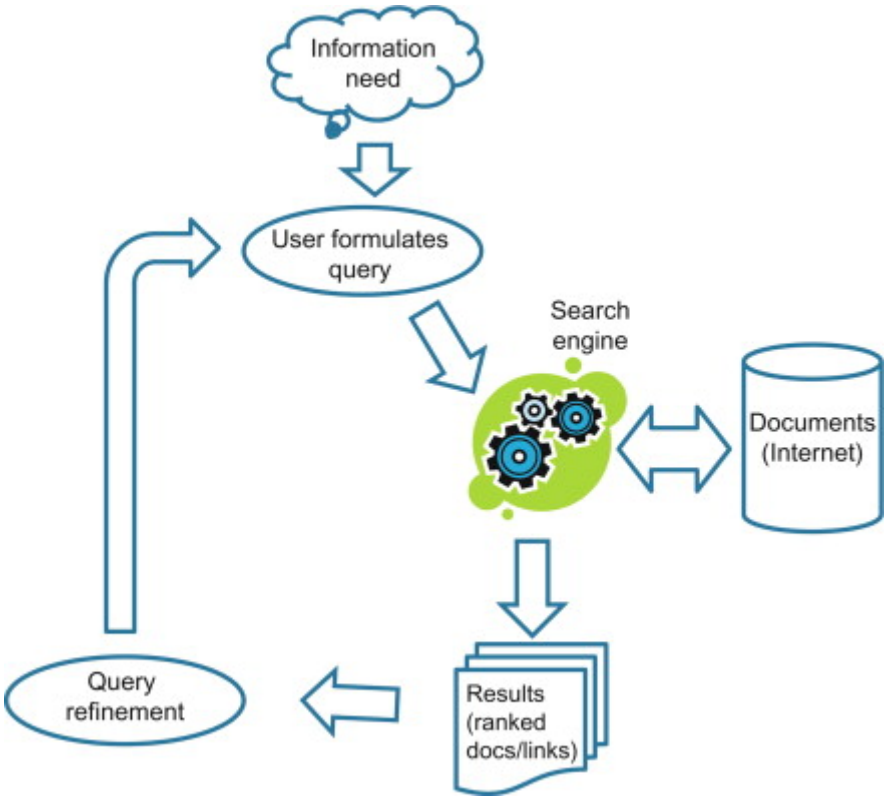

Figure 1 - Représentation schématique du cycle de recherche de l'utilisateur et de reformulation de la requête (Nettleton, 2014)

2.1.1. Taxonomies de reformulations de requêtes

Afin d'améliorer le processus de recherche d'informations, des travaux ont été menés pour développer des taxonomies de reformulations de requêtes. Le but de ces catégories est de décrire les types de modifications possibles pour les requêtes, telles que l'ajout, la suppression ou la modification de termes. Ces catégories sont souvent basées sur des critères fixes, tels que la syntaxe ou la structure de la requête.

Selon Huang et Efthimiadis (2009), il est possible d'analyser et d'évaluer les stratégies de reformulation de requêtes dans les journaux de recherche Web en utilisant des méthodes d'analyse de données textuelles. Ils ont proposé une taxonomie formelle de treize types de reformulations de requêtes, mais je ne présenterai ici que ceux qui sont pertinents pour ma recherche :

- Reformulation 1 : réorganisation des mots — l'ordre des mots de la seconde requête est modifié par rapport à la première requête. Exemple : seattle pizza palace → pizza seattle palace.
- Reformulation 2 : espace et ponctuation — les espaces et les signes de ponctuation varient entre les deux requêtes, des espaces peuvent être ajoutés ou supprimés entre les mots, de même pour la ponctuation. Exemple : wal mart, tomatoprices → walmart tomato prices
- Reformulation 3 : suppression de mots — des mots sont supprimés de la première requête dans la deuxième requête. Exemple : yahoo stock price → price yahoo
- Reformulation 4 : ajout de mots — à l'inverse de la reformulation précédente, de nouveaux termes apparaissent dans la deuxième requête. Exemple : Eastlake home → Eastlake home price index

- Reformulation 6 : stemming — changement morphologique des mots de la première requête. Exemple : running over bridges → run over bridge
- Reformulation 7 : transformation en acronyme — la seconde requête contient un acronyme correspondant à des mots de la première. Exemple : personal computer → PC
- Reformulation 8 : acronyme développé — à l'inverse de la précédente, l'acronyme est présent dans la première requête et développé dans la seconde. Exemple : pda → personal digital assistant
- Reformulation 9 : substring – la seconde requête est un préfixe ou suffixe strict de la première requête. Exemple : is there spyware on my computer → is there spywa
- Reformulation 10 : superstring – la deuxième requête contient la première requête comme préfixe ou suffixe. Exemple : nevada police rec → nevada police records 2008
- Reformulation 11 : abréviation – les mots correspondants de la première et la deuxième requête sont des préfixes l'un de l'autre (ceci diffère de substring qui prend en compte les suffixes et ne compare que les requêtes entières). Exemple : shortened dict → short dictionary
- Reformulation 12 : remplacement de mots – un ou plusieurs mots de la première requête sont remplacés par des mots sémantiquement liés. Deux mots sont liés si l'un d'eux est une relation sémantique (synonyme, hyponyme, hypéronyme, méronyme ou holonyme) de l'autre après avoir été convertis en leur forme morphologique de base
  - Synonyme : easter egg search → easter egg hunt
  - Hyponyme : crimson scarf → red scarf
  - Hypéronyme : personal computer → laptop
  - Méronyme : finger → hand
  - Holonyme : automobile → wheel
- Reformulation 13 : correction orthographique – la deuxième requête contient les mots corrigés orthographiquement de la première requête. Exemple : reformualtion → reformulation

Cependant, il est important de prendre en compte les limites de l'utilisation de taxonomies formelles, car celles-ci peuvent ne pas rendre compte des nuances sémantiques présentes dans les modifications de mots. En outre, leur maintenance et leur mise à jour peuvent être difficiles en raison du besoin d'une expertise linguistique approfondie.

En considérant ces limites, une approche alternative consiste à se concentrer sur les relations sémantiques entre les mots. Plutôt que de se limiter à des modifications formelles prédéfinies par une taxonomie, cette méthode cherche à comprendre comment les différents termes interagissent et influencent mutuellement le sens général des requêtes.

2.2. Relations sémantiques

La recherche d'information sur le web est un processus complexe qui nécessite de comprendre le sens des mots et des phrases pour trouver des résultats pertinents. C'est là que la sémantique entre en jeu.

Les relations sémantiques sont des liens entre les mots qui ont un sens similaire ou complémentaire (Bagha, 2011). Les relations sémantiques peuvent être classiques ou non-classiques (Morris & Hirst, 2004). Il existe de nombreux types de relations classiques et surtout non classiques, mais je ne mentionnerai ici que celles qui seront pertinentes dans le cadre de ma recherche.

2.2.1. Relations classiques

Les relations classiques sont des relations lexicales ou sémantiques qui sont largement reconnues et étudiées dans les domaines de la linguistique et de la lexicologie. Elles sont généralement considérées comme des relations fondamentales et systématiques entre les mots. Les relations classiques incluent des concepts tels que les synonymes, les antonymes, les hyperonymes, les hyponymes, les co-hyponymes, les méronymes et les holonymes.

Synonyme

Un synonyme est un mot qui a le même sens ou un sens très similaire à un autre mot (Lehmann & Martin-Berthet, 2018). Les synonymes permettent d'éviter les répétitions dans un texte et d'enrichir le vocabulaire. Par exemple, les mots "rapide" et "vite" sont des synonymes car ils expriment tous deux l'idée de vitesse.

Antonyme

Un antonyme est un mot qui a un sens opposé ou contraire à un autre mot (Lehmann & Martin-Berthet, 2018). Les antonymes permettent de marquer des oppositions et des contrastes dans un texte. Par exemple, les mots "chaud" et "froid" sont des antonymes car ils expriment des températures opposées.

Hyperonyme

Un hyperonyme est un mot dont le sens englobe le sens d'un ou plusieurs autres mots, appelés hyponymes (Lehmann & Martin-Berthet, 2018). L'hyperonyme est un terme plus général qui englobe des termes plus spécifiques. Par exemple, le mot "animal" est un hyperonyme pour les mots "chien", "chat" et "oiseau".

Hyponyme

Un hyponyme est un mot dont le sens est inclus dans le sens d'un autre mot, appelé hyperonyme (Lehmann & Martin-Berthet, 2018). L'hyponyme est un terme plus spécifique qui est englobé par un terme plus général. Par exemple, les mots "chien", "chat" et "oiseau" sont des hyponymes du mot "animal".

Co-hyponyme

Les co-hyponymes sont des mots qui partagent le même hyperonyme et qui sont donc au même niveau de spécificité (Lehmann & Martin-Berthet, 2018).. Par exemple, les mots "chien", "chat" et "oiseau" sont des co-hyponymes car ils ont tous pour hyperonyme le mot "animal".

Méronyme

Un méronyme est un mot qui désigne une partie d'un tout, dont le nom est appelé holonyme (Lehmann & Martin-Berthet, 2018). Par exemple, le mot "roue" est un méronyme du mot "voiture", car la roue est une partie de la voiture.

Holonyme

Un holonyme est un mot qui désigne un tout dont une partie est nommée par un méronyme (Lehmann & Martin-Berthet, 2018). Par exemple, le mot "voiture" est un holonyme du mot "roue", car la voiture est le tout dont la roue est une partie.

2.2.2. Relations non classiques

Les relations non classiques, par opposition, se réfèrent à des relations moins formelles, moins systématiques ou moins conventionnelles entre les mots ou les concepts. Elles peuvent émerger de l'association d'idées, de l'usage linguistique, de l'expérience individuelle ou d'autres facteurs subjectifs. Les relations non classiques peuvent être plus contextuelles, idiosyncratiques ou spécifiques à certaines expressions ou situations. Elles incluent des concepts tels que l'association d'idées, les instances et les collocations.

Association d'idées

L'association d'idées est une relation entre deux mots ou concepts qui sont liés par une connexion logique, culturelle ou contextuelle, sans être nécessairement synonymes, antonymes ou avoir une relation hiérarchique (Guelfand, 2013, p. 6). Par exemple, les mots "plage" et "vacances" sont associés car ils évoquent souvent des situations similaires, bien qu'ils ne soient pas synonymes.

Instance

Une instance est une relation entre un mot qui désigne une catégorie ou un ensemble et un mot qui désigne un élément particulier de cette catégorie ou ensemble (Matthiessen, 2013). Par exemple, "Paris" est une instance de la catégorie "ville", car Paris est une ville particulière.

Collocation

Une collocation est une combinaison de mots qui apparaissent fréquemment ensemble dans un texte ou un discours, formant une unité sémantique (Lehmann & Martin-Berthet, 2018). Les collocations sont souvent des expressions figées ou des constructions grammaticales spécifiques. Par exemple, l'expression "prendre une décision" est une collocation courante en français, car les mots "prendre" et "décision" apparaissent souvent ensemble dans cette construction particulière.

Les relations sémantiques sont des liens cruciaux entre les mots qui permettent de comprendre leur sens et leur contexte. Les relations classiques, telles que la synonymie et l'antonymie, ainsi que les relations non classiques, comme la collocation et l'association des idées, contribuent à la richesse de la langue et à la compréhension des concepts. En explorant ces relations, nous pouvons mieux appréhender la signification des mots et améliorer nos capacités de recherche d'information et de traitement du langage naturel.

2.3. Méthodes de mesure de similarité et distance sémantique

Pour quantifier et explorer les relations sémantiques entre les mots, différentes méthodes de mesure de similarité et de distance sémantique ont été développées. Ces méthodes permettent de représenter les mots sous forme de vecteurs et de déterminer leur proximité conceptuelle dans un espace vectoriel. En utilisant ces techniques, il est possible de quantifier et de comparer la similarité sémantique entre les mots, ce qui contribue à améliorer diverses tâches de traitement du langage naturel, telles que la recherche d'information, la traduction automatique et la génération de texte.

La similarité, également appelée distance, est une métrique plus souple que la synonymie. Elle mesure combien de caractéristiques de sens les mots partagent. Plus les mots sont similaires, plus ils partagent de caractéristiques de sens (Jurafsky & Martin, 2023). Par exemple, "chat" n'est pas un synonyme de "chien", mais "chats" et "chiens" sont des mots similaires (Jurafsky & Martin, 2023).

Selon Jurafsky & Martin (2023), il est important de distinguer la similarité entre mots (word similarity) des liens entre mots (word relatedness).

La similarité entre mots fait référence à la mesure de la proximité sémantique ou conceptuelle entre deux mots (Jurafsky & Martin, 2023). Elle évalue dans quelle mesure les mots partagent des caractéristiques, des sens ou des contextes similaires. Lorsque deux mots sont similaires, ils peuvent être considérés comme presque synonymes, c'est-à-dire qu'ils ont des significations très proches ou presque identiques. Par exemple, "voiture" et "vélo" sont des mots similaires, car ils partagent des caractéristiques liées au transport, à la mobilité et aux moyens de déplacement.

Les liens entre mots désignent les différentes relations ou associations qu'il peut y avoir entre les mots, sans nécessairement être des synonymes ou des mots similaires (Jurafsky & Martin, 2023). Les mots peuvent être liés de différentes manières, comme des relations de cause à effet, d'appartenance à une même catégorie, d'opposition, ou de co-occurrence fréquente dans le langage. Par exemple, "voiture" et "essence" sont des mots liés, car ils ont une relation de dépendance ou d'association. Dans ce cas, l'essence est souvent utilisée comme carburant pour les voitures.

La distinction entre similarité entre mots et liens entre mots est essentielle lors de l'exploration des méthodes de mesure de similarité et de distance sémantique telles que CamemBERT, Word2Vec et GloVe. Ces méthodes visent à capturer les relations sémantiques entre les mots en utilisant différentes approches.

2.3.1. Word2Vec

Word2Vec est un modèle basé sur un réseau neuronal qui apprend des embeddings de mots en prédisant le contexte d'un mot dans un corpus de texte donné (Mikolov et al., 2013). Il représente chaque mot comme un vecteur dans un espace de dimensions élevées, où les mots ayant des significations similaires sont rapprochés. La similarité entre deux mots peut être mesurée en utilisant la similarité cosinus entre leurs vecteurs correspondants.

2.3.2. GloVe

GloVe (Global Vectors for Word Representation) est un autre modèle basé sur un réseau neuronal qui apprend des embeddings de mots en factorisant une matrice de probabilités de co-occurrence des

mots (Pennington et al., 2014). Il représente chaque mot comme un vecteur dans un espace de dimensions élevées, où les mots ayant des significations similaires sont rapprochés. La similarité entre deux mots peut être mesurée en utilisant la similarité cosinus entre leurs vecteurs correspondants.

2.3.3. CamemBERT

CamemBERT est un modèle de langage pré-entraîné basé sur l'architecture Transformer qui a été spécifiquement entraîné sur du texte français (Martin, Muller, Suárez, Dupont, Romary, de la Clergerie, et al., 2020). Il peut être utilisé pour générer des embeddings de mots contextualisés, où la signification d'un mot dépend de son contexte dans la phrase. La similarité entre deux mots ou deux phrases peut être mesurée en utilisant la similarité cosinus entre leurs embeddings correspondants.

Ces modèles peuvent être utilisés pour mesurer la similarité sémantique ou la distance entre les mots, les phrases ou les documents. Ils sont couramment utilisés dans les tâches de traitement du langage naturel telles que la classification des textes, la recherche d'informations et la réponse aux questions.

2.3.4. Différences entre les modèles Word2Vec, GloVe et CamemBERT

Les modèles Word2Vec, GloVe et CamemBERT diffèrent dans la façon dont ils comprennent la signification des mots et des phrases.

Word2Vec et GloVe sont des modèles basés sur des réseaux neuronaux (Mikolov et al., 2013; Pennington et al., 2014). Les représentations vectorielles des mots sont apprises en prêtant attention au contexte d'un mot ou à la cooccurrence de mots dans un texte. Ces modèles prennent les mots un par un et créent des vecteurs qui regroupent les mots qui ont des significations similaires. Pour mesurer la similarité entre deux termes, on compare les vecteurs correspondants en utilisant la similarité cosinus. Ces modèles sont "statiques" car les vecteurs des mots sont fixes et ne tiennent pas compte du contexte spécifique de la phrase ou du document.

CamemBERT, en revanche, est un modèle de langage pré-entraîné basé sur l'architecture Transformer, spécialement conçu pour le français (Martin, Muller, Suárez, Dupont, Romary, Clergerie, et al., 2020; Martin, Muller, Suárez, Dupont, Romary, de la Clergerie, et al., 2020). Il donne des représentations vectorielles de contexte pour les mots, ce qui signifie que la signification d'un mot dépend de son contexte dans la phrase. CamemBERT considère la structure syntaxique et sémantique de la phrase, lui permettant de capturer des significations subtiles que les modèles statiques tels que Word2Vec et GloVe ne permettent pas.

La principale raison d'utiliser CamemBERT plutôt que les modèles statiques réside dans sa capacité à tenir compte du contexte. Lorsque vous modifiez une requête ou une phrase, CamemBERT peut saisir les changements de sens qui se produisent en fonction du contexte. Par exemple, dans les modèles statiques, les mots "banque" et "rivière" peuvent sembler similaires car ils apparaissent souvent dans des contextes similaires. Cependant, si vous mentionnez le mot "pont" dans une phrase, CamemBERT peut comprendre le contexte spécifique et montrer que "banque" est plus étroitement lié à "pont" qu'à "rivière" en termes de similarité sémantique. Ainsi, CamemBERT est capable de représenter plus précisément la signification changeante des mots en fonction de leur contexte.

En conclusion, l'utilisation de CamemBERT offre une meilleure compréhension des mots et des phrases en prenant en compte le contexte, ce qui le rend très utile pour saisir les nuances sémantiques et

les variations de sens. Cela en fait un choix idéal pour des tâches où il est essentiel de comprendre précisément le sens, comme la recherche d'informations ou la réponse à des questions.

3. Composition, caractéristiques et processus d'annotation des données

Dans ce chapitre, j'examinerai les données utilisées dans mon étude. Dans un premier temps, je détaillerai les données CoST (An annotated Data Collection for Complex Search) que j'ai utilisées, en soulignant leur composition et leurs caractéristiques. Ensuite, j'examinerai les données et effectuerai des étapes de prétraitement pour les préparer à mon analyse. Enfin, je discuterai du processus d'annotation manuelle que j'ai suivi pour enrichir les données avec des informations supplémentaires.

3.1. Description des données CoST (An annotated Data Collection for Complex Search)

CoST est un ensemble de données annotées pour l'évaluation de tâches de recherche complexes (Dosso et al., 2021).

Les données CoST ont été recueillies dans le cadre d'une étude utilisateur impliquant 70 participants francophones experts dans l'un des trois domaines de compétence différents : informatique, médecine et psychologie. Chaque participant a réalisé 15 tâches différentes, avec 5 types de complexité cognitive différents : recherche de faits, apprentissage exploratoire, prise de décision, résolution de problèmes et inférence multicritère (Dosso et al., 2021).

Les données ont été collectées à partir de 630 sessions basées sur des tâches, et au total, 5667 requêtes ont été enregistrées. Outre les données de recherche telles que les requêtes et les clics, CoST fournit également des données relatives aux tâches et aux sessions, ainsi que des annotations de tâches et de requêtes (Dosso et al., 2021).

Par exemple, les participants devaient répondre à la tâche et la sous-tâche de prise de décision suivantes :

Task	TaskName	Description	SubTask
TDPsy	Decision-Making	Dans le cadre d'un projet de conception d'un site Web, vous avez pour objectif de proposer une architecture de qualité pour votre site. Vous souhaitez mettre en place une méthodologie de tri de cartes mais vous hésitez entre un tri de cartes ouvert ou fermé et physique ou informatisé. Après avoir relevé les avantages et les inconvénients de chaque type de tri de cartes, sélectionnez la méthode qui vous paraît la meilleure en justifiant vos choix.	1. Comprendre la méthodologie de tri de cartes 2. Comprendre les spécificités de chaque méthodologie de tri de carte 3. Identifier les avantages et les inconvénients de chaque méthodologie 4. Analyser/Différencier les informations recueillies 5. Juger ces informations selon des critères à établir

Ensuite, les participants formulent des requêtes pour trouver les informations dont ils ont besoin. Toutes leurs sessions sont enregistrées en détail. Voici un exemple de participant cherchant la réponse à la tâche précédente :

QueryId	IdS	Query	Task	QueryActivity	QueryTime	SerpTime
324	Psy5	méthodologie tri de carte	TDPsy	Exploration	1,3702E+11	1,4285E+11
325	Psy5	méthodologie tri carte	TDPsy	SpellingCorrection	3,3173E+11	3,3656E+11
326	Psy5	scholar	TDPsy	Exploration	9,70543E+15	9,71167E+11
327	Psy5	tri de cartes	TDPsy	Exploration	9,76238E+15	9,78162E+11
328	Psy5	scholar	TDPsy	Exploration	99275	99275
329	Psy5	tri carte	TDPsy	Exploration	1,00117E+11	1,0014E+11
330	Psy5	méthode tri carte informatisé	TDPsy	Exploitation	1,36643E+11	1,36668E+11

Les clics des participants sont également enregistrés. L'exemple des clics pour le même participant qui a effectué la recherche précédente :

QueryId	URL	URL Time
324	http://www.google.fr/	2,5005E+11
325	http://www.google.fr/url?q=http://www.ergolab.net/articles/tri-cartes-ergonomie-web.php&sa=U&ved=2ahUKEwi2n8reg5zmAhUH1BoKHUTHDtUQFjACegQIBhAB&usg=AOvVaw1ffiVU_bQ8WOJMaCyplOND	4,6626E+11
325	http://www.google.fr/url?q=http://www.ergolab.net/articles/tri-cartes-ergonomie-web.php&sa=U&ved=2ahUKEwibjo_chZzmAhUQrxoKHR7gAnMQFjACegQIBRAB&usg=AOvVaw2O6Gyma6mYWQ-I_qEgK3JN	5,71435E+11
325	http://www.google.fr/url?q=https://fr.wikipedia.org/wiki/Tri_par_cartes&sa=U&ved=2ahUKEwiLqOjphZzmAhUJx4UKHQGxCGoQFjADegQIBRAB&usg=AOvVaw0FpdIdJYGYGT_rdnl-l0Se	5,96846E+11
325	http://www.google.fr/url?q=https://www.cairn.info/revue-document-numerique-2009-2-page-23.htm&sa=U&ved=2ahUKEwib65j_hZzmAhUF4BoKHXsmCkQ4ChAWMAB6BAgBEAE&usg=AOvVaw1xspEa98UDA_6CdGHePpku	6,49243E+11
326	http://scholar.google.com/	9,73384E+11
327	http://scholar.google.com/	9,90187E+11
328	http://scholar.google.com/	9,951E+11
329	https://www.cairn.info/revue-i2d-information-donnees-et-documents-2017-1-page-62.htm?1=1&DocId=430850&hits=1698+1697+1695+1694+1693+1692+1691+1689+11+10+8+7+6+5+4+2+	1,03296E+11
329	https://www.cairn.info/revue-i2d-information-donnees-et-documents-2017-1-page-62.htm?1=1&DocId=430850&hits=1698+1697+1695+1694+1693+1692+1691+1689+11+10+8+7+6+5+4+2+	1,16583E+11
329	https://www.cairn.info/revue-i2d-information-donnees-et-documents-2017-1-page-62.htm?1=1&DocId=430850&hits=1698+1697+1695+1694+1693+1692+1691+1689+11+10+8+7+6+5+4+2+#s1n3	1,17455E+11
329	https://www.cairn.info/revue-i2d-information-donnees-et-documents-2017-1-page-	1,33497E+11

62.htm?1=1&DocId=430850&hits=1698+1697+1695+1694+1693+1692+1691+1689+11+10+8+7+6+5+4+2+#s1n4

A la fin de leurs recherches, les participants doivent donner une réponse aux tâches demandées. Voici un exemple de réponse du même participant qui a effectué les recherches précédentes :

IdS	Exp	Task	Answer
Psy5	Psy	TDPsy	A l'instar de l'étude de Paillaré (2017), la méthode qui semble être la plus pertinente dans le cadre d'un projet de conception d'un site web est le tri de carte ouvert. Il apparaît être une source d'information riche pour construire un rubriquage pertinent en permettant l'accès aux représentations mentales de l'organisation et de l'interaction entre les contenus proposés plus que le tri fermé. La liberté d'utilisation est plus grande bien que le tri physique et informatisé possède tous deux de grands avantages. Le tri informatisé permet une analyse qualitative supérieure. Elle permet également d'avoir un échantillon plus grand et des analyses plus faciles. De plus, il semble exister un logiciel (lardtort) qui permet de se rapprocher des avantages du tri physique.

L'objectif principal de la collecte de ces données était de créer un ensemble de données richement annotées pour évaluer les tâches de recherche complexes (Dosso et al., 2021). Les chercheurs provenant des domaines de l'informatique et de la psychologie cognitive ont collaboré pour concevoir cet ensemble de données, dans le but de répondre à un large éventail de questions de recherche liées à la recherche basée sur les tâches.

Les données CoST peuvent être utilisées pour évaluer des modèles de classification de requêtes et pour comprendre l'effet de la complexité des tâches et du domaine d'expertise sur le comportement de recherche des utilisateurs (Dosso et al., 2021).

Les données CoST revêtent une grande importance pour ma recherche car elles regroupent un volume considérable de requêtes (5667 au total) correspondant à plusieurs types de tâches complexes provenant de trois domaines d'expertise distincts. De plus, les données ont été recueillies dans un environnement contrôlé, garantissant ainsi une qualité optimale. Une autre caractéristique précieuse de ces données est la richesse de ses annotations. En fait, j'ai des informations détaillées sur la corrélation entre les requêtes et les tâches spécifiques, me permettant de savoir quelle recherche est associée à quelle tâche, quelle est la première requête de la séquence et quelle est la dernière. Cette granularité d'annotation est un avantage significatif par rapport aux journaux de requêtes anonymes, qui ne fournissent pas le même niveau de contexte. Ainsi, grâce aux données CoST, je dispose d'un ensemble de données complet et pertinent pour mener à bien mon étude de recherche.

3.2. Observation des données et prétraitement

Le jeu de données des requêtes CoST est pré-annoté avec l'identifiant du participant (expert en psychologie, en informatique ou en médecine), le domaine de la tâche (psychologie, informatique ou

médecine), l'activité de la requête (correction orthographique, exploration, exploitation ou exploitation restreinte), le temps de la requête et le temps du SERP.

Pour répondre à mes questions de recherche, j'avais seulement besoin de paires de requêtes reformulées, donc le domaine d'expertise des utilisateurs, leur stratégie (exploitation, exploration, etc.) ou le temps qu'ils ont passé sur les pages de résultats (SERP) ou à reformuler leurs requêtes sont sans importance. Ces informations ne sont pas directement liées à mes objectifs de recherche et n'ont donc pas été prises en compte dans mon étude. Mon analyse se concentre uniquement sur les modifications apportées aux requêtes et leur impact sur leur signification. La seule exception à cela était l'annotation du domaine des tâches, que j'ai utilisée pour créer un script Python spécifique. Ce script utilisait l'information sur le domaine des tâches pour effectuer des opérations spécifiques dans le cadre de mon étude. J'expliquerai en détail la fonctionnalité de ce script plus loin dans mon travail.

Pour les besoins de ma recherche, je devais extraire de ce jeu de données uniquement les requêtes dans lesquelles un seul mot était ajouté, supprimé ou modifié. J'ai également décidé d'exclure toutes les requêtes dont les tâches relèvent du domaine de la médecine, car il serait difficile d'analyser correctement ces données compte tenu de mes connaissances limitées dans ce domaine. Enfin, je voulais annoter automatiquement la classification des changements (mot ajouté, supprimé ou modifié), le mot ajouté/supprimé/modifié, la longueur de la requête originale (nombre de mots), ainsi que la position des mots ajoutés/supprimés/modifiés, car cela serait assez simple à faire et permettrait de gagner beaucoup de temps par rapport à l'annotation manuelle.

Afin d'accomplir cela, j'ai écrit un script Python qui effectue les tâches suivantes. Tout d'abord, il lit les données du jeu de données CoST. Ensuite, il compare les requêtes actuelles avec les requêtes suivantes pour repérer les mots qui ont été ajoutés, supprimés ou modifiés.

Pour ce faire, j'utilise l'algorithme de Levenshtein, qui calcule la distance d'édition entre les mots. Cette distance mesure la similarité entre deux mots en quantifiant le nombre minimal d'opérations nécessaires (ajout, suppression, substitution) pour transformer un mot en un autre. Grâce à cet algorithme, mon script est capable de traiter des requêtes contenant des fautes d'orthographe et d'évaluer les modifications apportées aux mots dans ces requêtes.

Lorsqu'une seule modification est détectée, le programme enregistre les détails de cette modification, tels que le mot ajouté, supprimé ou modifié, ainsi que la longueur de la requête d'origine et la position de la modification dans la requête.

Une fois toutes les requêtes traitées, le script génère un jeu de données complet composé de 853 paires de requêtes. Ce jeu de données fournit des informations précieuses sur les modifications, ajouts ou suppressions de mots dans les requêtes, ce qui permet une analyse plus approfondie de la similarité entre les phrases.

3.3. Annotation manuelle des données

Le tableau présenté ci-dessous est utilisé pour l'annotation manuelle des données. Il comprend plusieurs colonnes qui fournissent des informations détaillées sur les requêtes originales et reformulées, ainsi que sur les modifications apportées. Comme expliqué dans la section précédente, les six premières colonnes sont annotées automatiquement à l'aide d'un script Python, tandis que les autres colonnes sont remplies manuellement. Les colonnes comprennent des informations telles que : la requête originale, la requête reformulée, la classification des modifications effectuées, les mots ajoutés, supprimés ou modifiés, leur position et longueur dans la requête, la relation lexicale avec les autres mots, les

commentaires qualitatifs, les changements formels et la langue utilisée. Cette structure permet de capturer des données pertinentes pour l'analyse linguistique des requêtes reformulées.

Ici, pour une meilleure lisibilité, j'ai tourné les colonnes à la verticale pour en faire des lignes et j'ai ajouté deux colonnes à côté : des exemples de données stockées et des explications sur la catégorie donnée et sur la raison pour laquelle cette catégorie a été prise en compte et pas d'autres.

catégorie	exemple	explication
requête originale	algorithme transcription sms	Cette colonne contient la requête initiale telle qu'elle a été formulée par l'utilisateur. Cela permet de connaître le contexte initial et d'analyser les éventuels changements qui ont été apportés.
requête reformulée	développer algorithme transcription sms	Cette colonne représente la version modifiée de la requête originale. En étudiant les modifications apportées, il est possible de comprendre comment la formulation a été modifiée et de déterminer si cela a un impact sur le sens de la requête.
classification	ajouté	Cette colonne indique si des modifications ont été apportées à la requête originale. Pour cette catégorie, je me suis inspirée par la taxonomie de Huang & Efthimiadis (2009), concrètement les reformulations 3, 4 et 12. Cette information est utile pour identifier si la reformulation a ajouté, supprimé ou modifié un mot par rapport à la requête originale.
mot ajouté / supprimé / modifié	développer	Cette colonne spécifie le mot précis qui a été ajouté, supprimé ou modifié dans la requête reformulée. Cela permet d'analyser les changements lexicaux effectués et d'évaluer leur impact sur le sens global de la requête.

position du mot ajouté/supprimé/modifié dans la requête reformulée	au début	Cette colonne enregistre la position relative du mot ajouté, supprimé ou modifié dans la requête reformulée. Les options possibles sont "au début", "au milieu" ou "à la fin". Pour cette catégorie, j'ai été inspirée par Russell (2019), chercheur principal chez Google, qui affirme dans son cours "Power Searching with Google" que l'ordre des mots est important pour la recherche sur Google. Dans ce cours, il donne plusieurs exemples de la manière dont l'ordre des mots peut affecter le changement de sens. Par exemple, la reformulation de la requête initiale "sky" en ajoutant un mot devant, comme "blue sky" (ciel bleu), a un sens différent que si l'on ajoute le même mot à la fin, "sky blue" (bleu ciel). On peut remarquer que, bien que les deux modifications entraînent un changement de sens, dans ce cas, l'ajout du nouveau mot à la fin de la requête modifie le sens de manière plus significative que l'ajout du mot à l'avant, puisqu'il fait passer le sens de "la région de l'atmosphère" à "une nuance de couleur". En français, contrairement à l'anglais, les adjectifs viennent généralement après le nom. Mon hypothèse serait donc que plus le mot modifié est éloigné du début de la requête, moins le changement de sens est important.
longueur de la requête originale	3 mots	Cette colonne indique le nombre de mots dans la requête originale, en excluant les mots vides. Cette information permet de prendre en compte la longueur de la requête

		initiale lors de l'analyse statistique du lien entre la longueur de la requête et le changement de sens. La logique sous-jacente est que plus il y a de mots dans la requête (à condition qu'il s'agisse de mots pleins), mieux le contexte général est établi, de sorte qu'il y a moins de chances qu'un ajout, une suppression ou une substitution d'un mot puisse modifier substantiellement son contexte et, par conséquent, le sens de la requête. En d'autres termes, mon hypothèse est la suivante : plus la requête originale est longue, moins le changement de sens sera important.
même famille morphologique	NON	Cette colonne détermine si le mot ajouté, supprimé ou modifié appartient à la même famille morphologique qu'un mot de la requête originale. Les valeurs possibles sont "OUI" ou "NON". Pour cette catégorie, je me suis inspirée par Moreau (2006), qui affirme que les mots provenant de la même famille morphologique sont sémantiquement très proches. L'auteur utilise les mots "produire", "produit", "producteur" et "productrice" comme exemple de mots ayant la même racine pour affirmer qu'il ne s'agit que de variantes morphologiques différentes du même mot. En suivant cette logique, j'émets l'hypothèse que si le mot modifié appartient à la même famille morphologique qu'un mot de la requête originale, le changement de sens ne sera pas significatif.

le(s) mot(s) le(s) plus proche(s)	algorithme	Cette colonne identifie le ou les mots les plus proches du mot ajouté, supprimé ou modifié dans la requête reformulée. Cela peut aider à comprendre le contexte entourant le mot modifié et à déterminer son impact sur le sens général de la requête, ainsi qu'à indiquer le(s) mot(s) entre lesquel(s) et le mot ajouté/supprimé/modifié existe une relation lexicale. Les difficultés que j'ai rencontrées lors de la détection des mots proches ont été multiples. Tout d'abord, j'ai été confronté à certaines notions utilisées dans certains domaines de la psychologie ou de l'informatique que je ne connaissais pas auparavant. Par exemple, j'ai dû apprendre ce que signifiait "persona ad hoc" dans le domaine de l'ergonomie informatique, ainsi que "PVT test" en psychologie, ou encore "tri de cartes" dans le contexte de l'ergonomie informatique. Ces termes spécifiques étaient nouveaux pour moi et ont nécessité des recherches approfondies pour les comprendre et les intégrer correctement dans mes analyses. Ensuite, j'ai été confronté à des mots qui étaient proches de la limite de la similarité. Parfois, je pouvais percevoir un lien entre deux mots, mais ce lien était plutôt ténu et tiré par les cheveux. Cela rendait la détection des mots proches plus délicate, car il fallait prendre en compte des similitudes subtiles tout en évitant les associations erronées. Ce défi demandait une attention particulière et une

		analyse minutieuse pour distinguer les véritables connexions des coïncidences fortuites.
relation lexicale	association d'idées	Cette colonne spécifie la relation lexicale entre le mot ajouté, supprimé ou modifié et les autres mots de la requête. Les différentes relations lexicales possibles incluent "synonyme", "hyponyme", "association d'idées", "collocation", etc. Pour cette catégorie, je me suis inspirée par la taxonomie de Huang & Efthimiadis (2009), concrètement la reformulation 12 : remplacement de mots. S'il existe une relation lexicale entre deux mots, cela signifie qu'ils se retrouvent dans les mêmes contextes. Si le contexte ne change pas avec l'ajout, la suppression ou la modification d'un mot lexicalement proche, le changement de sens ne devrait pas être significatif. En d'autres termes, j'émets l'hypothèse que s'il existe une relation lexicale, le changement de sens sera moins important. Les difficultés que j'ai rencontrées lors de la détection des relations lexicales étaient nombreuses. Tout d'abord, j'ai été confronté à de nombreux cas où les mots étaient presque synonymes, presque antonymes presque hyponymes, etc., mais pas tout à fait. Il était donc nécessaire de décider quelles relations lexicales je pouvais classifier comme des liaisons classiques et dans quels cas il s'agissait simplement d'une association d'idées. C'était un exercice délicat de jongler entre être trop biaisé en assignant une

		relation trop étirée et négliger une liaison importante. Une autre difficulté résidait dans les termes spécialisés ou techniques que je ne connaissais pas auparavant. Dans de tels cas, j'ai dû effectuer des recherches supplémentaires afin de déterminer s'il existait une relation entre les mots en question. Comprendre le sens précis des termes spécialisés et leur relation avec d'autres mots était essentiel pour une détection précise des relations lexicales. Il était également important de prendre en compte le contexte dans lequel les mots étaient utilisés. Les associations d'idées pouvaient parfois être trompeuses, car les mots pouvaient apparaître ensemble fréquemment sans pour autant avoir une relation lexicale directe. Il fallait donc faire preuve de discernement pour éviter de tirer des conclusions hâtives basées uniquement sur des co-occurrences.
commentaires qualitatifs	mot méta	Cette colonne permet d'ajouter tout commentaire pertinent pour l'annotation, comme par exemple la mention d'un mot méta. L'hypothèse est que si le mot ajouté, supprimé ou modifié est un mot méta, le changement de sens sera moins important, puisque les mots méta n'apportent pas de nouveaux contextes et sont vides de sens.

changements formels	NON	Cette colonne est dédiée à l'enregistrement des changements formels tels que l'utilisation d'opérateurs, les variations flexionnelles, les superstrings, etc. Pour cette catégorie, je me suis inspirée par la taxonomie de Huang & Efthimiadis (2009), concrètement les reformulations suivantes : réorganisation des mots, espace et ponctuation, stemming, transformation en acronyme, acronyme développé, substring, superstring, abréviation et correction orthographique. Ces informations peuvent être utiles pour comprendre les modifications apportées à la requête et pour déterminer si les modifications formelles peuvent affecter le sens de la requête.
langue	fr	Cette colonne indique la langue dans laquelle la requête a été formulée. Les valeurs possibles sont "fr" pour le français ou "en" pour l'anglais. En indiquant si la requête est formulée en français ou en anglais, je peux m'assurer que le modèle CamemBERT, spécialement conçu pour le français, est utilisé de manière appropriée. Bien que CamemBERT puisse toujours prédire des mots en anglais, il ne sera pas aussi précis qu'un modèle conçu spécifiquement pour l'anglais ou que si la requête était en français. Ainsi, si l'on soupçonne que les données produites par CamemBERT pour les requêtes en anglais sont de bien moindre qualité et faussent les statistiques, ces requêtes

		pourraient être retirées de l'analyse.

4. Méthodologie pour mesurer le changement de sens dans les requêtes reformulées

Dans le cadre de mon mémoire, j'ai choisi d'utiliser CamemBERT pour l'exploitation de mes données. CamemBERT est particulièrement adapté pour la tâche de prédiction en masque (MASK) car il peut calculer le rang auquel le mot ajouté/modifié/supprimé apparaît dans son vocabulaire. Cela me permet de déterminer si le sens de la requête a été modifié de manière significative. De plus, CamemBERT est un modèle qui prend en compte le contexte linguistique (Martin, Muller, Suárez, Dupont, Romary, Clergerie, et al., 2020; Martin, Muller, Suárez, Dupont, Romary, de la Clergerie, et al., 2020), contrairement aux modèles statiques tels que GloVe et Word2Vec. Ainsi, il offre une compréhension fine des nuances et spécificités de la langue française, ce qui est essentiel pour mon étude.

CamemBERT est un modèle de langage pour le français basé sur l'architecture RoBERTa et pré-entraîné sur un vaste corpus de textes français, comprenant OSCAR et CCNet (Martin, Muller, Suárez, Dupont, Romary, Clergerie, et al., 2020; Martin, Muller, Suárez, Dupont, Romary, de la Clergerie, et al., 2020). Il présente plusieurs avantages pour mesurer la similarité en français, notamment :

Conception spécifique pour le français : CamemBERT est spécifiquement conçu pour le français, lui permettant de saisir les nuances et les spécificités de la langue française, fournissant des résultats précis et pertinents (Martin, Muller, Suárez, Dupont, Romary, Clergerie, et al., 2020; Martin, Muller, Suárez, Dupont, Romary, de la Clergerie, et al., 2020).

Modèle pré-entraîné de grande taille : CamemBERT est un modèle pré-entraîné de grande taille, ce qui lui permet de mieux comprendre les relations complexes entre les mots et les phrases, ce qui le rend plus efficace dans une variété d'applications en langage naturel, y compris la mesure de similarité (Martin, Muller, Suárez, Dupont, Romary, Clergerie, et al., 2020).

Architecture optimisée : CamemBERT est basé sur l'architecture RoBERTa, qui a été optimisée pour améliorer les performances de BERT en modifiant certains hyperparamètres et en supprimant l'objectif de prédiction de la phrase suivante (Martin, Muller, Suárez, Dupont, Romary, de la Clergerie, et al., 2020). Cela permet à CamemBERT de bénéficier des améliorations apportées par RoBERTa tout en étant spécifiquement adapté au français.

CamemBERT a été évalué dans plusieurs tâches pour le français, telles que l'étiquetage des parties du discours, l'analyse des dépendances, la reconnaissance des entités nommées et l'inférence de langage naturel, et a démontré une amélioration de l'état de l'art pour la plupart de ces tâches par rapport aux approches monolingues et multilingues précédentes. CamemBERT est disponible en différentes versions avec des nombres de paramètres, des quantités de données d'entraînement préalable et des domaines sources de données d'entraînement variables.

Dans le cadre de ma recherche, j'utilise l'approche en "MASK" avec le modèle CamemBERT pour déterminer le rang auquel le mot ajouté/modifié/supprimé est retrouvé. Lors de cette approche, je masque le mot cible dans la requête (ex. "détection plagiat <mask>") et j'observe la prédiction du modèle pour ce masque. Le modèle génère une liste de probabilités pour les mots possibles qui pourraient remplacer le masque. Le rang auquel le mot cible apparaît dans cette liste de probabilités est alors utilisé pour évaluer l'influence de la modification sur le sens global de la requête.

En d'autres termes, si le mot ajouté/modifié/supprimé est prédit par CamemBERT avec un rang plus bas, cela indique que le modèle considère ce mot comme une bonne continuation de la requête, ce qui suggère que le sens global de la requête n'a pas beaucoup changé. En revanche, si le mot se trouve

à un rang élevé, cela suggère que le modèle a du mal à prédire ce mot et que la modification peut avoir altéré le sens de la requête.

Cette approche en "MASK" avec l'utilisation du rang de prédiction dans CamemBERT me permet d'évaluer comment les modifications apportées à une requête influencent le sens global perçu par le modèle.

Plusieurs raisons justifient l'utilisation de Camembert dans le contexte de ma recherche :

1. **Mesure de similarité sémantique** : Le rang d'un mot dans mes prédictions avec CamemBERT peut me donner une idée de la similarité sémantique entre les phrases. Si le mot ajouté/modifié/supprimé se trouve parmi les premiers résultats de CamemBERT, cela indique que le mot est sémantiquement proche des autres mots de la requête et ne change pas beaucoup le sens.
2. **Adaptabilité au français** : CamemBERT est spécifiquement conçu pour le français, ce qui signifie qu'il est capable de capturer les nuances et les spécificités de la langue française. Cela me permet d'obtenir des résultats plus précis et pertinents lors de l'analyse des changements de sens dans les requêtes en français.

J'ai cherché d'automatiser ce processus, vu que mon jeu de données comprend 853 paires de requêtes. Pour réaliser cela, j'ai développé un script Python qui utilise CamemBERT pour analyser des phrases. Le programme vérifie d'abord si le mot recherché est présent dans le vocabulaire de CamemBERT. Ensuite, le script parcourt ensuite chaque phrase, remplace le mot recherché par "<mask>" selon sa classification (ajouté, supprimé ou modifié) et enregistre l'emplacement du mot dans les prédictions. Les résultats sont enregistrés dans un fichier Excel pour une analyse ultérieure.

De 853 paires de requêtes, CamemBERT a généré des prédictions pour 242 paires de requêtes, dont le rang le plus bas est 0 et le rang le plus élevé est 22861. Cela veut dire que Camembert n'a pas trouvé 611 mots dans son vocabulaire.

J'ai utilisé ces 242 paires de requêtes annotées lors des tests statistiques afin de déterminer s'il existe une corrélation ou des tendances significatives entre le rang auquel le mot ajouté/modifié/supprimé et la longueur de la requête originale, position du mot ajouté/supprimé/modifié dans la requête reformulée, l'appartenance à la même famille morphologique, les relations lexicales et changements formels.

5. Résultats et analyses

Les analyses et les résultats jouent un rôle crucial dans cette recherche, car ils permettent de tirer des conclusions pertinentes et d'approfondir la compréhension du sujet étudié. Dans cette partie, j'examinerai et j'interpréterai en détail les résultats obtenus en utilisant des tests statistiques. J'analyserai les tendances et les modèles observés, mettant en évidence les points clés les plus significatifs. De plus, je discuterai des limites de mon étude ainsi que des perspectives de recherche futures.

5.1. Présentation des résultats obtenus : tests statistiques

L'utilisation de tests statistiques est essentielle pour repérer les tendances ou les relations significatives dans les données. Les tests statistiques me permettent d'évaluer si les résultats que j'observe sont simplement le fruit du hasard ou s'ils représentent une association réelle entre les variables. En analysant les données à l'aide de tests statistiques appropriés, je peux obtenir une compréhension plus approfondie des relations sous-jacentes et de leur signification.

Dans le cadre de cette étude, je chercherai à déterminer s'il existe une corrélation entre le rang auquel le mot ajouté/modifié/supprimé apparaît et la longueur de la requête originale, la position du mot dans la requête reformulée, l'appartenance à la même famille morphologique, les relations lexicales et les changements formels. Ces informations peuvent m'aider à mieux comprendre les liens entre les modifications apportées aux requêtes et leur impact sur les prédictions générées par le modèle CamemBERT.

En utilisant les tests statistiques appropriés, je serai en mesure de quantifier l'association entre ces variables et de déterminer si les différences observées sont statistiquement significatives. Cela me permettra d'obtenir des conclusions plus robustes et d'identifier d'éventuelles tendances ou corrélations dans les données.

Plus loin dans le texte, je vais utiliser le terme topk. Le terme "top-k" fait référence à une technique utilisée dans le domaine du traitement du langage naturel pour générer du texte de manière sélective.

La méthode top-k consiste à limiter le choix des mots générés lors de la génération de texte aux k mots les plus probables, où k est un paramètre défini par l'utilisateur. Cela signifie que seuls les k mots les plus probables, selon les prédictions du modèle, sont pris en compte pour la génération du texte suivant.

L'utilisation de la méthode top-k permet de contrôler la diversité et la qualité du texte généré. Si k est un nombre élevé, cela permet d'explorer un plus grand nombre de possibilités, ce qui peut entraîner une plus grande variété de résultats. En revanche, si k est un nombre faible, le modèle se concentrera sur les mots les plus probables, ce qui peut conduire à des textes plus cohérents mais potentiellement moins surprenants. Ici, j'ai attribué la valeur 32005 à top-k, ce qui correspond à la taille totale du vocabulaire CamemBERT.

Pour évaluer la corrélation ou la différence significative entre les variables, des tests statistiques non paramétriques ont été utilisés étant donné que les variables ne suivent pas une distribution normale. J'utiliserai des tests de corrélation de Spearman pour les variables qui ne suivent pas une distribution

normale et des tests de différence appropriés pour évaluer les variations entre les groupes. Les tests choisis et leur justification pour chaque paire de variables mentionnée sont les suivants :

1. Rang auquel le mot ajouté/modifié/supprimé apparaît (variable numérique continue) et longueur de la requête originale (variable numérique continue) :

    - Méthodologie : Le coefficient de corrélation de rang de Spearman est utilisé pour mesurer la corrélation monotone entre ces deux variables continues, sans supposer une distribution normale.
    - Justification : Étant donné que les variables ne suivent pas une distribution normale, le coefficient de corrélation de Spearman est plus approprié pour évaluer la relation monotone entre les deux variables.
    - Résultats :

        Corrélation de Spearman : rho = -0.127, p-value = 0.048

        Il y a une corrélation négative significative entre topk (rang auquel le mot ajouté/modifié/supprimé apparaît) et longueur.

        La figure 2 présente un nuage de points avec une ligne de tendance, permettant d'observer la relation entre les scores topk (rang auquel le mot ajouté/modifié/supprimé apparaît) et la longueur des mots dans les requêtes.

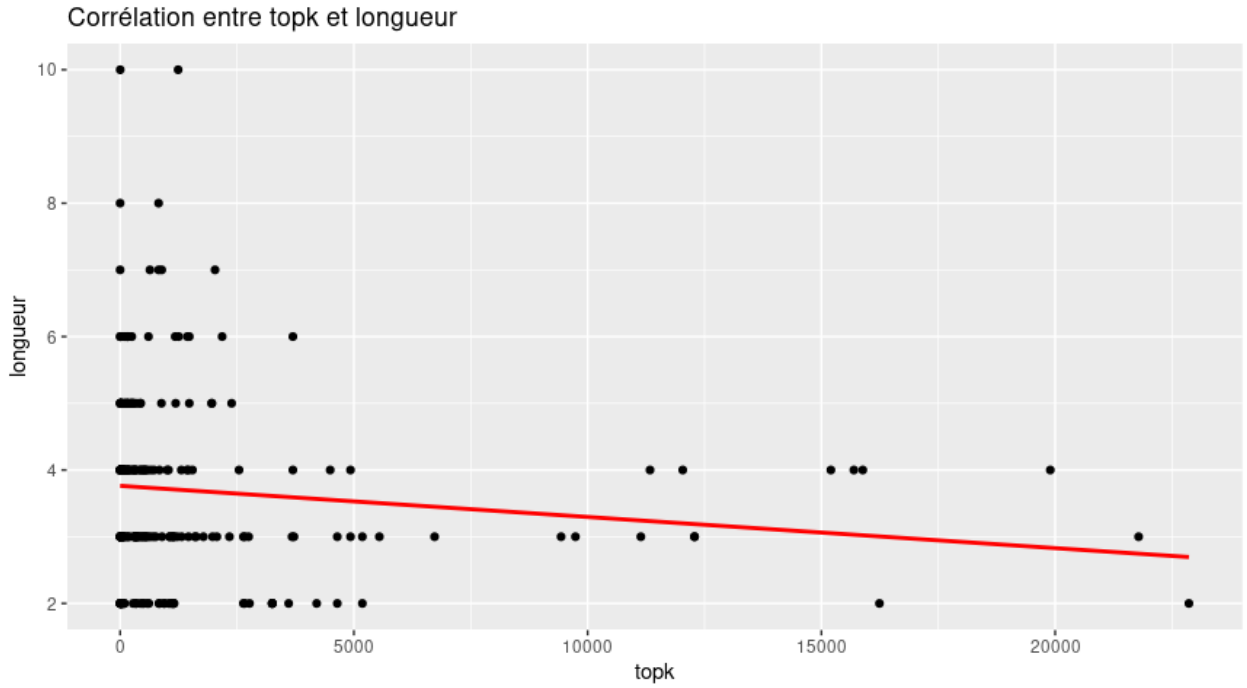

Figure 2 : Corrélation entre le rang auquel le mot ajouté/modifié/supprimé apparaît et la longueur de la requête originale

2. Rang auquel le mot ajouté/modifié/supprimé apparaît (variable numérique continue) et position du mot ajouté/supprimé/modifié dans la requête reformulée (variable catégorielle ordinale) :

    - Méthodologie : Le coefficient de corrélation de rang de Spearman est également utilisé pour évaluer la corrélation monotone entre ces deux variables.

- Justification : Comme la position est une variable catégorielle ordinale, il est plus approprié d'utiliser le coefficient de corrélation de Spearman pour évaluer la relation monotone entre ces variables.
- Résultats :

  Corrélation de Spearman : rho = 0.100, p-value = 0.121

  Il y a une corrélation positive non significative entre topk (rang auquel le mot ajouté/modifié/supprimé apparaît) et position.

  La figure 3 affiche un diagramme en boîte qui compare la distribution des scores topk (rang auquel le mot ajouté/modifié/supprimé apparaît) en fonction de la position des mots dans les requêtes, permettant d'analyser l'impact de la position sur les scores topk (rang auquel le mot ajouté/modifié/supprimé apparaît).

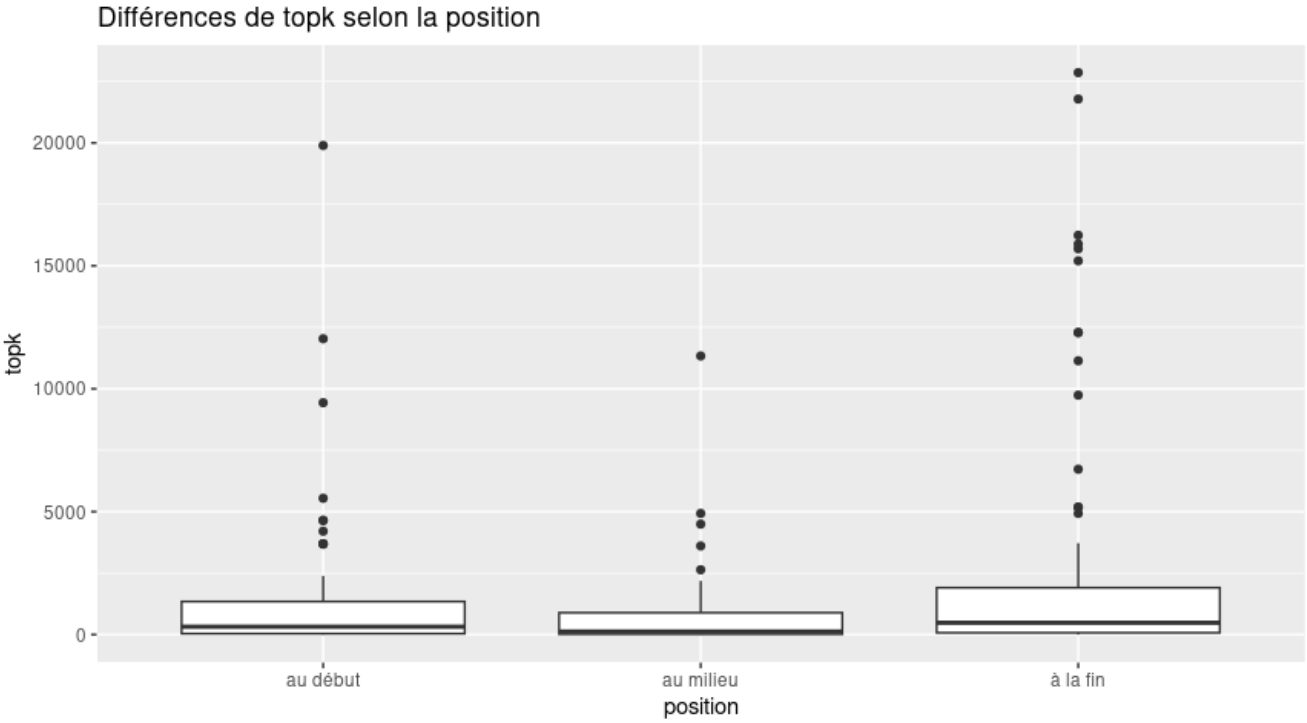

Figure 3 : Différences de topk selon la position

3. Rang auquel le mot ajouté/modifié/supprimé apparaît (variable numérique continue) et le mot appartenant à la même famille morphologique (variable catégorielle binaire) :
   - Méthodologie : Le test U de Mann-Whitney est utilisé pour comparer la distribution des valeurs de topk (rang auquel le mot ajouté/modifié/supprimé apparaît) entre les deux catégories de la variable même famille morphologique.
   - Justification : Étant donné que même famille morphologique est une variable catégorielle binaire et que les données ne suivent pas une distribution normale, le test U de Mann-Whitney est approprié pour évaluer s'il y a une différence significative dans les valeurs de topk (rang auquel le mot ajouté/modifié/supprimé apparaît) entre les deux catégories.
   - Résultats :

Test de Wilcoxon-Mann-Whitney : p-value = 0.927

Il n'y a pas de différence significative entre topk (rang auquel le mot ajouté/modifié/supprimé apparaît) et les mots appartenant à la même famille morphologique.

La figure 4 illustre un diagramme en boîte qui met en évidence la distribution des scores topk (rang auquel le mot ajouté/modifié/supprimé apparaît) en fonction de la similarité morphologique des mots dans les requêtes, permettant d'évaluer l'influence de la similarité morphologique sur les scores topk (rang auquel le mot ajouté/modifié/supprimé apparaît).

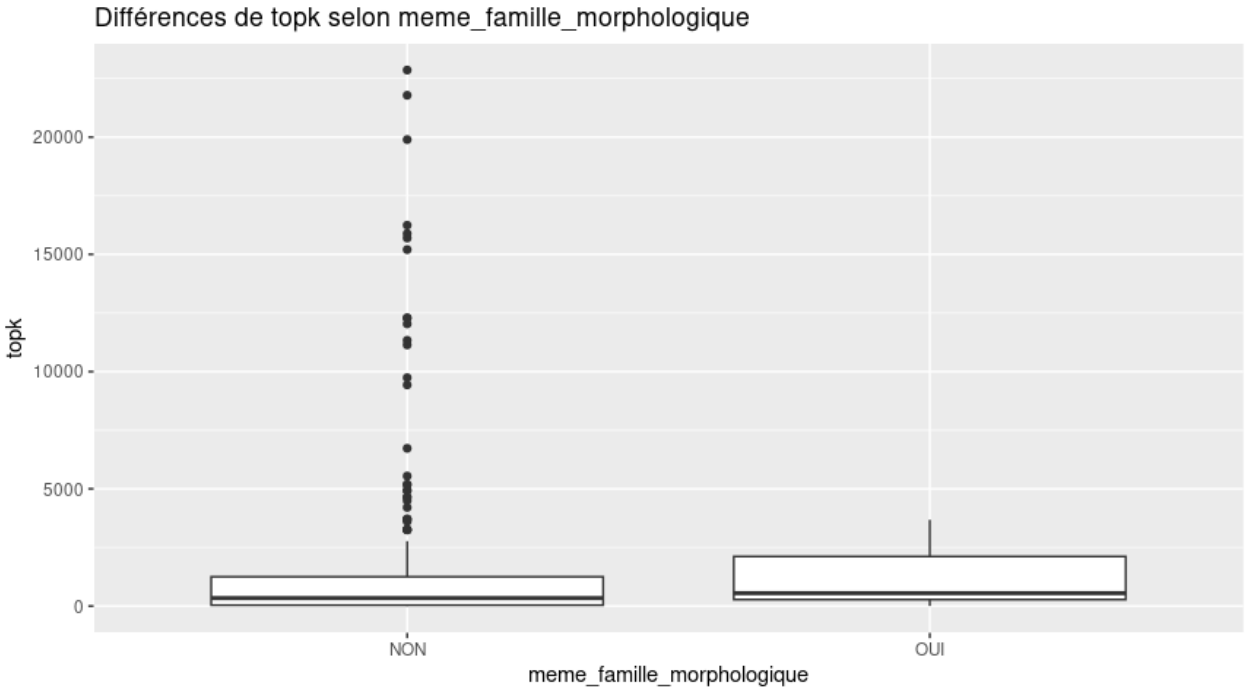

Figure 4 : Différences de topk selon la similarité morphologique

4. Rang auquel le mot ajouté/modifié/supprimé apparaît (variable numérique continue) et relation lexicale (variable catégorielle nominale) :

    - Méthodologie : Le test de Kruskal-Wallis est utilisé pour évaluer si les distributions des valeurs de topk (rang auquel le mot ajouté/modifié/supprimé apparaît) varient significativement entre les différentes catégories de la variable relation.
    - Justification : Le test de Kruskal-Wallis est un test non paramétrique qui ne nécessite pas de distribution normale. Il permet de déterminer s'il y a une corrélation significative entre topk (rang auquel le mot ajouté/modifié/supprimé apparaît) et les différentes catégories de relation.
    - Résultats :

        Test de Kruskal-Wallis : p-value = 0.652

        Il n'y a pas de différence significative entre les groupes de relation.

        La figure 5 présente un diagramme en boîte qui compare la distribution des scores topk (rang auquel le mot ajouté/modifié/supprimé apparaît) en fonction de la relation lexicale entre les mots dans les requêtes, permettant d'analyser l'impact de

la relation lexicale sur les scores topk (rang auquel le mot ajouté/modifié/supprimé apparaît).

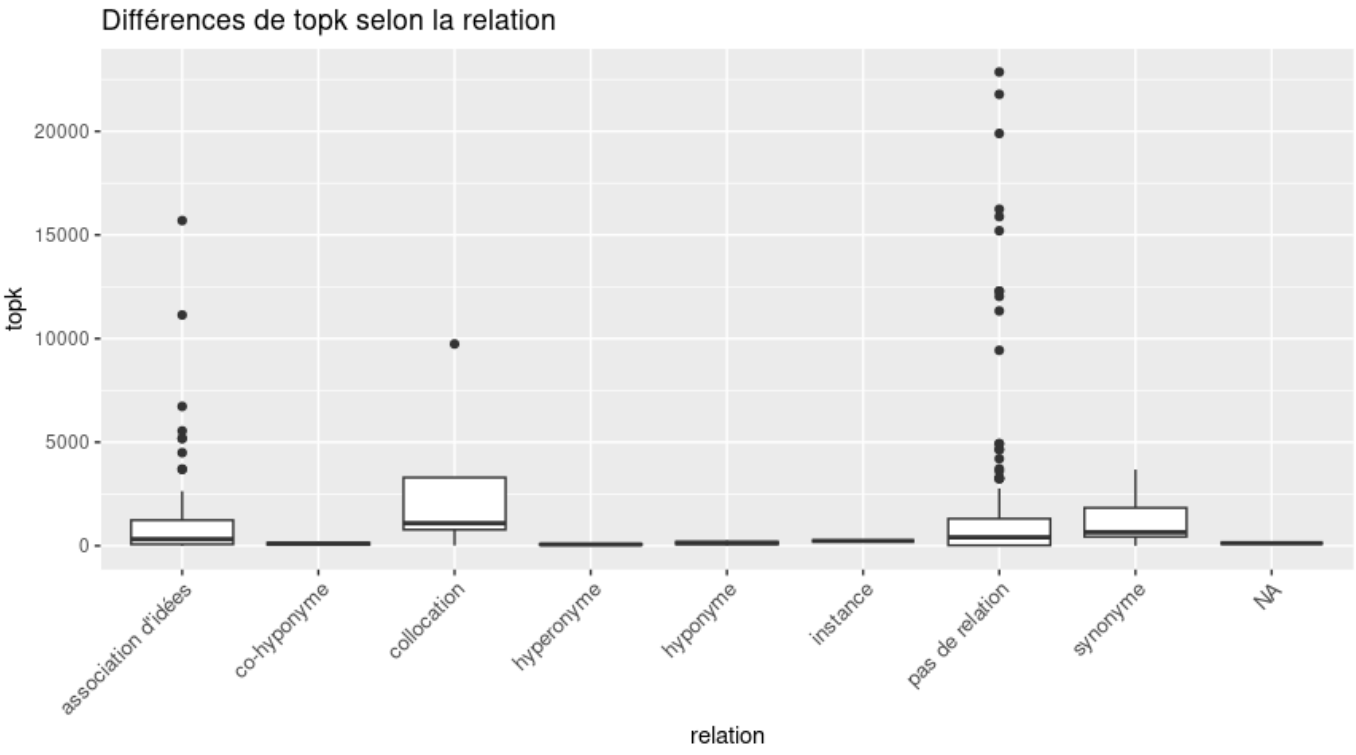

Figure 5 : Différences de topk selon la relation lexicale

5. Rang auquel le mot ajouté/modifié/supprimé apparaît (variable numérique continue) et changements formels (variable catégorielle nominale) :

    - Méthodologie : Encore une fois, le test de Mann-Whitney est utilisé pour évaluer s'il y a une différence significative dans les valeurs de topk (rang auquel le mot ajouté/modifié/supprimé apparaît) entre les deux catégories de la variable changements formels.
    - Justification : Étant donné que changements formels est une variable catégorielle nominale et que les données ne suivent pas une distribution normale, le test de Mann-Whitney est approprié pour évaluer s'il y a une différence significative dans les valeurs de topk (rang auquel le mot ajouté/modifié/supprimé apparaît) entre les deux catégories.
    - Résultats :

        Test de Kruskal-Wallis : p-value = 0.225

        Il n'y a pas de différence significative entre topk (rang auquel le mot ajouté/modifié/supprimé apparaît) et la présence des changements formels.

        La figure 6 présente un diagramme en boîte qui met en évidence la distribution des scores topk (rang auquel le mot ajouté/modifié/supprimé apparaît) en fonction de la présence ou de l'absence de changements formels dans les mots des requêtes, permettant d'évaluer l'impact des changements formels sur les scores topk (rang auquel le mot ajouté/modifié/supprimé apparaît).

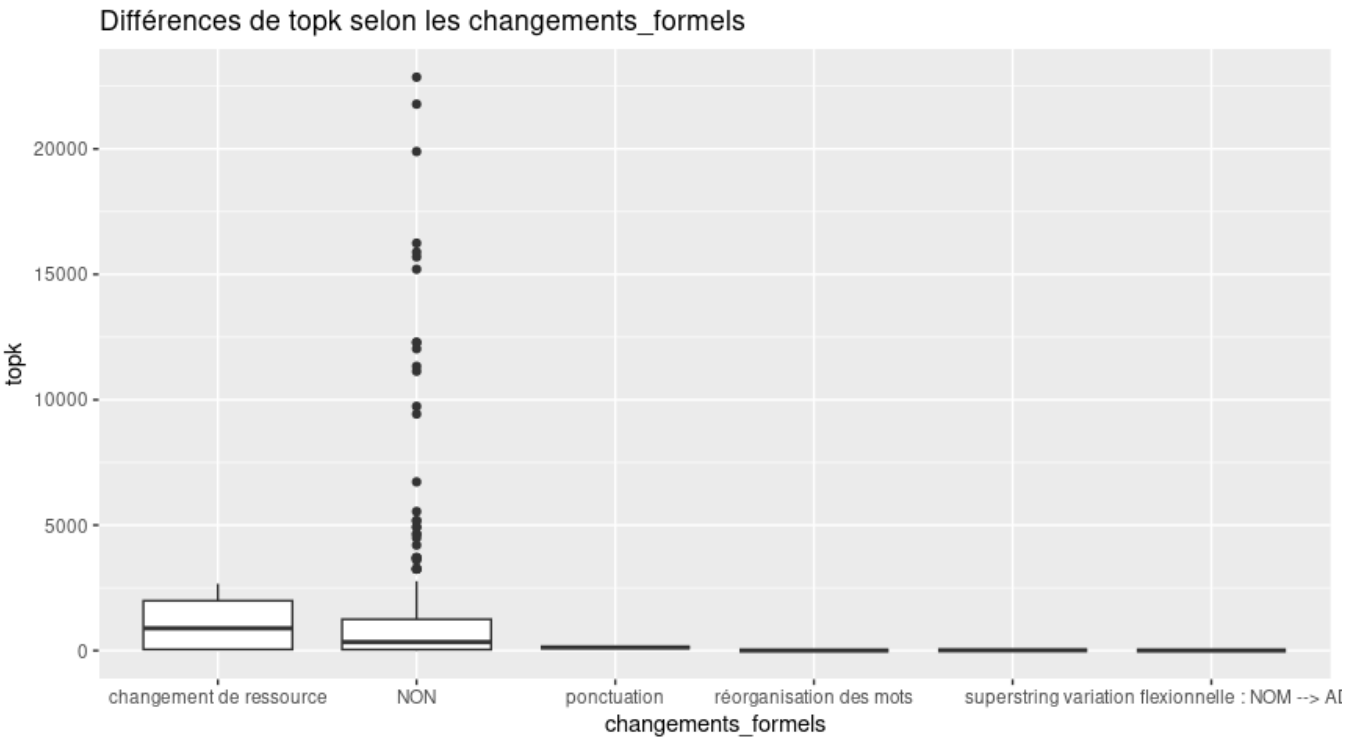

Figure 6 : Différences de topk selon les changements formels

5.2. Interprétation et analyse des résultats statistiques

Je vais maintenant présenter et analyser en détail les résultats obtenus à partir des tests statistiques réalisés. Cette étape d'interprétation est cruciale pour comprendre les relations et les différences entre les variables étudiées. En examinant les corrélations et les tests de différence effectués, je vais explorer les liens entre le score topk (rang auquel le mot ajouté/modifié/supprimé apparaît) et d'autres variables telles que la longueur des mots, la position, la similarité morphologique, la relation entre les mots et la présence de changements formels. Grâce à cette analyse approfondie, nous pourrons mieux appréhender les facteurs qui influencent les scores topk (rang auquel le mot ajouté/modifié/supprimé apparaît) et tirer des conclusions éclairées sur les caractéristiques des mots qui sont associées à des scores plus élevés ou plus bas.

1. Test de corrélation entre topk et longueur : La corrélation de Spearman entre topk et longueur est de -0.127, avec une p-value de 0.048. Cela indique une corrélation négative significative entre ces deux variables. En d'autres termes, lorsque la longueur de la requête originale augmente, le score topk diminue. Cela peut suggérer qu'il y a une tendance à associer des requêtes plus courtes à des scores topk plus élevés, tandis que des requêtes plus longues ont tendance à obtenir des scores topk plus bas.

   Conclusion : plus les requêtes originales sont longues, moins la reformulation d'un mot change de manière significative le sens de la requête reformulée.

2. Test de corrélation entre topk et position : La corrélation de Spearman entre topk et position est de 0.100, avec une p-value de 0.121. Cette corrélation positive n'est cependant pas significative. Cela signifie qu'il n'y a pas de relation linéaire forte entre topk et la position des mots. La position "au début", "au milieu" ou "à la fin" n'a pas d'impact significatif sur les scores topk.

   Conclusion : la position des mots ne semble pas avoir d'impact notable sur la mesure dans laquelle le sens des requêtes reformulées est modifié.

3. Test de différence entre topk et la même famille morphologique : Le test de Wilcoxon-Mann-Whitney montre une p-value de 0.927, ce qui indique qu'il n'y a pas de différence significative entre les mots appartenant à la même famille morphologique et leurs scores topk.

   Conclusion : la similarité morphologique des mots n'influence pas de manière significative le changement de sens des requêtes reformulées.

4. Test de différence entre topk et la relation lexicale entre les mots : Le test de Kruskal-Wallis montre une p-value de 0.652, ce qui suggère qu'il n'y a pas de différence significative entre les différents types de relations en ce qui concerne les scores topk.

   Conclusion : aucune des catégories de relation n'a d'effet significatif sur le changement de sens des requêtes reformulées.

5. Test de différence entre topk et la présence de changements formels : Le test de Kruskal-Wallis donne une p-value de 0.225, ce qui indique qu'il n'y a pas de différence significative entre les mots avec ou sans changements formels en ce qui concerne les scores topk.

   Conclusion : les changements formels ne semblent pas influencer de manière significative le sens des requêtes reformulées.

En conclusion, les résultats de l'analyse montrent une corrélation significative entre le rang auquel le mot ajouté/modifié/supprimé apparaît et la longueur des mots, avec une corrélation négative. En d’autres mots, plus la requête originale est longue, moins la modification d’un seul mot sera importante.

Cependant, aucune corrélation significative n'a été trouvée entre le rang auquel le mot ajouté/modifié/supprimé apparaît et la position des mots, la même famille morphologique, la relation entre les mots ou la présence de changements formels.

Cela suggère que d'autres facteurs peuvent jouer un rôle plus important dans le changement de sens lors des reformulations. Il est important de noter que ces résultats sont basés sur les données fournies et peuvent varier en fonction du contexte et des caractéristiques spécifiques de l'ensemble de données utilisé.

5.3. Analyse linguistique qualitative des prédictions de CamemBERT

Dans cette section, je vais me pencher sur une analyse linguistique qualitative des prédictions de CamemBERT dans la tâche en MASK. Plus précisément, j'examinerai les dix premières prédictions du modèle pour les mots ayant le score le plus bas, les mots ayant le score le plus élevé, ainsi que les mots non trouvés dans son vocabulaire.

5.3.1. Les mots ayant le score le plus bas

Commençons par les mots ayant le score le plus bas. Cette analyse qualitative a révélé des facteurs linguistiques communs qui peuvent être associés aux changements de sens dans les requêtes reformulées.

Les tableaux 1, 2 et 3 contiennent des exemples des trois requêtes avec le score le plus bas. Les mots entre les chevrons (ex. <anglais>) sont les mots masqués, c'est-à-dire les mots que CamemBERT a cherché à prédire.

Dans les exemples donnés, on observe que les prédictions de CamemBERT pour les mots trouvés et ayant un score de 0 sont souvent des co-hyponymes ou des variations morphologiques du mot original.

Requête masquée : plagiat traduction <anglais>
les dix premières prédictions :
anglais
francais
dictionnaire
anglaise
française
français
francaise
espagnol
automatique
gratuite

Par exemple, dans la requête "plagiat traduction <anglais>" (cf. tableau 1), les prédictions incluent des co-hyponymes tels que "français" et "espagnol", des variations morphologiques du mot "anglais" ("anglaise") et des variations morphologiques et orthographiques des hyponymes trouvés ("française", "francais"). Cette similarité sémantique entre les mots prédits et le mot original indique que le sens global de la requête n'a pas beaucoup changé.

Tableau 1 : plagiat traduction <anglais>

De même, dans la requête "méthodologie qui permet de transcrire un <text> avec langage familier a langage soutenu" (cf. tableau 2), la première prédiction est la variation orthographique du mot masqué : "texte". Les deux suivantes prédictions sont les synonymes : "document" et "message". Les sept autres prédictions sont des mots étroitement liés au mot original. On peut observer que les liens entre les prédictions et le mot masqué s'affaiblissent au fur et à mesure que le score top-k augmente. Encore une fois, cela suggère que le sens de la requête est resté cohérent malgré les modifications.

Dans la requête "expérience psychologie <effet> sur l'attention boutons lumières panneau" (cf. tableau 3), les prédictions incluent des mots tels que "effets", "test", "impact" et "action". Ces mots sont tous en relation avec le mot original "effet" et reflètent des variations sémantiques possibles autour de

ce concept. Cela indique que le sens de la requête a pu évoluer, mais reste centré sur le domaine de la psychologie et de l'attention.

En conclusion, ces exemples soulignent l'importance des relations lexicales, des variations morphologiques et des mots proches dans les prédictions de CamemBERT, et par conséquent montrent que leur présence dans les requêtes n'impacte pas significativement le changement de sens. Les variations morphologiques et les termes étroitement liés au mot original sont souvent privilégiés dans les prédictions de CamemBERT, et la logique derrière cela est que vous pourriez échanger l'un de ces mots avec le mot masqué et le sens de la requête resterait pratiquement le même, puisque le contexte lui-même n'est pas modifié.

Requête masquée : méthodologie qui permet de transcrire un <text> avec langage familier a langage soutenu
les dix premières prédictions :
texte
document
message
discours
énoncé
fichier
programme
article
anglais
langage

Tableau 2 : méthodologie qui permet de transcrire un <text> avec langage familier a langage soutenu

Requête masquée : expérience psychologie <effet> sur l'attention boutons lumières panneau
les dix premières prédictions :
effets
jeux
test
tests
travail
effet
impact
information
action
informations

Tableau 3 : expérience psychologie <effet> sur l'attention boutons lumières panneau

5.3.2. Les mots ayant le score le plus élevé

Les mots ayant les scores les plus élevés représentent les mots que CamemBERT a eu le plus de mal à trouver, car il ne les a pas considérés comme une suite logique de la requête originale. Selon mon hypothèse, ces mots devraient être ceux qui modifient le plus le sens des requêtes originales.

Les tableaux 4, 5 et 6 contiennent des exemples des trois requêtes avec les scores les plus élevés. Ici aussi, les mots situés entre les chevrons (par exemple <inventeur>) sont des mots masqués.

Dans la requête ""théorie de la modularité" <inventeur>" (cf. tableau 4), la grande majorité des prédictions incluent les changements formels, tels que l'utilisation de ponctuation (..., :, ?) et les changements de ressources (DOC, PDF). Je peux émettre l'hypothèse que cela est dû à la présence d'opérateurs dans la requête originale (guillemets autour "théorie de la modularité"). D'ailleurs, je

remarque que le mot masqué ("inventeur") est un mot méta. Cela va à l'encontre de mon hypothèse selon laquelle les mots méta devraient avoir un score plus bas avec CamemBERT, c'est-à-dire être prédits plus rapidement, car ils ne modifient pas le sens de la requête originale.

Requête masquée : "théorie de la modularité" <inventeur>	Score top-k : 22861
les dix premières prédictions :	
...	
DOC	
:	
»	
?	
PDF	
;	
"	
-	
(...)	

Tableau 4 : "théorie de la modularité" <inventeur>

De même, dans la requête "modularity hypothesis author <ergonomics>" (cf. tableau 5), ce qui domine les dix premières prédictions de CamemBERT sont les signes de ponctuation, mais cette fois nous avons aussi des mots pleins appartenant au domaine mathématique ("mathématique", "spatiale", "Statistique"). Je suppose que ce contexte mathématique est établi en raison de la présence du mot "modularité" dans la requête originale, puisque c'est le seul mot plein de la requête.

Requête masquée : auteur hypothèse de la modularité <ergonomie>	Score top-k : 21785
les dix premières prédictions :	
...	
[...]	
(...)	
mathématique	
:	
spatiale	
;	
des	
»	
Statistique	

Tableau 5 : auteur hypothèse de la modularité <ergonomie>

Pour la requête "big data <écologie>" (cf. tableau 6), la grande majorité des mots prédits par CamemBERT sont les chiffres, plus précisément les années (2018, 2016, 2017). Les autres sont des mots entiers appartenant au domaine de l'informatique et de la gestion d'entreprise ("software", "management", "business"), ce qui n'est pas surprenant étant donné que le concept de big data vient de l'informatique. J'ai également remarqué un changement de ressource ("wiki"). Au fond, c'est un résultat attendu puisque les mots "écologie" et "big data" sont assez éloignés sémantiquement, donc je peux conclure que le mot "écologie" modifie significativement le sens de la requête originale. Je note également que le participant a utilisé l'anglicisme "big data" au lieu du mot français "mégadonnées". Cela peut avoir rendu le modèle CamemBERT moins précis, étant donné qu'il a été construit spécifiquement pour le français.

Requête masquée : big data <écologie>	Score top-k : 16242
les dix premières prédictions :	
2018	
2016	
2017	
software	
management	
2015	
wiki	
france	
business	
2013	

Tableau 6 : big data <écologie>

Dans les exemples donnés, on observe que les points communs entre les dix premières prédictions des requêtes mentionnées sont principalement liés à la ponctuation, aux changements de ressources et aux méta-mots. Pour les exemples "big data <ecology>" et "modularity hypothesis author <ergonomics>", il n'est pas si surprenant que le modèle ait eu du mal à prédire les mots masqués, puisqu'il s'agit de mots assez éloignés sémantiquement des autres mots présents dans leurs requêtes respectives. En revanche, le mot "inventeur" dans la requête ""théorie de la modularité" <inventeur>" devrait être prédit plus rapidement, étant donné qu'il s'agit d'un méta-mot et qu'il est donc vide dans le contexte. Ceci montre que de nombreux facteurs entrent en jeu lors de la reformulation des requêtes, et que certains facteurs qui semblent insignifiants à première vue, comme la présence d'opérateurs, peuvent en fait influencer fortement la requête sous certaines conditions.

5.3.3. Les mots non trouvés par CamemBERT

Ici, j'aborderai les cas où CamemBERT n'a pas trouvé le mot masqué dans son vocabulaire. Je regarderai s'il a trouvé des mots provenant de la même famille morphologique, leur rang et quelles sont les caractéristiques des dix premières prédictions pour ces requêtes.

Les tableaux 7, 8 et 9 contiennent des exemples des trois requêtes avec les mots non trouvés par CamemBERT. Comme dans les deux exemples précédents, les mots entre les chevrons (par exemple <orhographique>) sont des mots masqués.

Pour le premier exemple, "algorithme de correction <orhographique>" (cf. tableau 7), en examinant la liste des prédictions de CamemBERT, j'ai trouvé des mots avec la même racine morphologique : "graphique", "ortho" et "orthographe". Le premier, "graphique", se trouve dans les dix premières prédictions (top-k 8) et partage la racine morphologique avec la deuxième partie du mot "orhographique". Le second, "ortho", a également un score top-k assez bas (63), ce qui signifie qu'il a été facilement prédit par CamemBERT, et qu'il s'agit simplement d'une abréviation du mot masqué "orhographique". Le troisième, "orthographe", n'est qu'une variation flexionnelle du mot masqué.

Par ailleurs, si l'on regarde les dix premières prédictions, on voit bien des signes de ponctuation et un changement de ressource ("DOC"), mais aussi des mots appartenant au même contexte que le mot "orhographique" ("grammatical", "graphique", "automatique").

Tout ceci pris en considération, je peux conclure que dans ce cas le mot "orhographique" n'a que très peu d'impact sur le changement de sens de la requête en question.

algorithme de correction <orhographique>		
les dix premières prédictions :	Les mots provenant de la même famille morphologique et leur rang	
...	graphique	8
automatique	ortho	63
[...]	orthographe	1486
de		
DOC		
|		
»		
grammatical		
graphique		
:		
...		

Tableau 7 : algorithme de correction <orhographique>

Concernant le deuxième exemple, "représentation mentale psychologie <cognitive>" (cf. tableau 8), on peut considérer que le mot masqué est retrouvé par CamemBERT et ce assez rapidement (top-k 48), mais dans le mauvais genre. En dehors du genre, il n'y a pas d'autre différence de sens entre ces deux mots. Cela est probablement dû à la collocation "psychologie cognitive", c'est-à-dire que les mots qui font partie d'une collocation sont plus facilement prévisibles.

En examinant les dix premières prédictions de CamemBERT, je peux constater que tous les mots appartiennent au contexte psychologique.

Ainsi, dans ce cas également, je conclus que l'ajout du mot "cognitive" n'a pas eu beaucoup d'impact sur le sens de la requête originale.

représentation mentale psychologie <cognitive>		
les dix premières prédictions :	Les mots provenant de la même famille morphologique et leur rang	
sociale	cognitif	48
mentale		
comportementale		
humaine		
clinique		
psychologie		
cerveau		
analytique		
graphique		
visuelle		
sociale		

Tableau 8 : représentation mentale psychologie <cognitive>

Dans le troisième cas (cf. tableau 9), nous avons également un mot qui fait partie d'une collocation (" persona ad hoc "), mais dans ce cas, même si CamemBERT a prédit des mots ayant la même racine morphologique (" Personne ", " Personnage "), ils n'ont pas été prédits aussi rapidement que dans les deux exemples précédents. De plus, bien qu'ils partagent de nombreuses similitudes, il ne

s'agit pas de synonymes complets, mais plutôt de quasi-synonymes, car ils ne sont pas interchangeables et sont utilisés dans des contextes différents.

Quant aux dix premières prédictions de CamemBERT, elles devinent le bon contexte, notamment les mots "page" et "site".

Ma conclusion, basée sur le rang des mots appartenant à la même famille morphologique et sur les dix premières prédictions, est que le mot "persona", contrairement aux mots des deux exemples précédents, a un impact sur le changement de sens de la requête, mais très légèrement.

<persona> ad hoc		
les dix premières prédictions :	Les mots provenant de la même famille morphologique et leur rang	
Page	Personne	333
Texte	Personnage	775
Message		
Formation		
Solution		
Site		
Catégorie		
Liste		
Article		
Support		
Page		

Tableau 9 : <persona> ad hoc

Dans le cas de mots non trouvés dans le vocabulaire de CamemBERT, en examinant les autres mots trouvés et leurs rangs, je peux déduire si le sens a été modifié ou non. De même, le fait qu'un mot n'ait pas été découvert par CamemBERT peut être trompeur et ne signifie pas nécessairement que le mot est rare. Parfois, il s'agit simplement d'une question de l’accord en genre, comme le montre le tableau 8, ou d'une autre variante du mot ayant exactement le même sens, comme le montre le tableau 7.

5.3.4. Conclusion

En conclusion, cette analyse linguistique qualitative des prédictions de CamemBERT dans la tâche de masquage a mis en évidence plusieurs observations intéressantes. Pour les mots ayant le score le plus bas, les prédictions du modèle étaient souvent des co-hyponymes, des variations morphologiques ou des mots étroitement liés au mot original de la requête. Cela indique que le sens global de la requête n'a pas beaucoup changé malgré les modifications. En revanche, pour les mots ayant le score le plus élevé, CamemBERT a eu du mal à prédire les mots masqués, ce qui suggère que ces mots modifient significativement le sens des requêtes originales. Les prédictions ont été dominées par des signes de ponctuation, des changements de ressources et des mots appartenant à des domaines sémantiques spécifiques. Enfin, lorsque CamemBERT n'a pas trouvé les mots masqués dans son vocabulaire, il a souvent prédit des mots de la même famille morphologique ou des mots appartenant au même contexte sémantique. Ces observations soulignent l'importance des relations lexicales, des variations morphologiques et des mots proches dans les prédictions de CamemBERT.

6. Conclusion

Dans ce mémoire, j'ai exploré l'impact de l'ajout, de la modification ou de la suppression d'un seul mot sur le sens d'une requête, ainsi que les facteurs linguistiques et contextuels qui contribuent à ces changements sémantiques.

Les résultats de l'analyse statistique ont révélé une corrélation significative entre le rang auquel le mot ajouté/modifié/supprimé apparaît dans le vocabulaire de CamemBERT et la longueur des mots, avec une corrélation négative. En d'autres termes, plus la requête originale est longue, la modification d'un seul mot a tendance à avoir moins d'impact sur le changement de sens.

Cependant, aucune corrélation significative n'a été trouvée entre le rang du mot et la position des mots, la même famille morphologique, la relation entre les mots ou la présence de changements formels. C'est-à-dire que ces facteurs ont très peu d'influence sur le changement de sens.

Cela suggère que d'autres facteurs non examinés dans cette étude peuvent jouer un rôle plus important dans les changements de sens lors des reformulations. Il est également important de souligner que ces résultats sont spécifiques aux données utilisées dans cette étude et peuvent varier selon le contexte et les caractéristiques propres à chaque ensemble de données.

L'analyse linguistique qualitative des prédictions de CamemBERT dans la tâche de masquage a également fourni des informations intéressantes.

Les prédictions du modèle pour les mots ayant le score bas, c'est-à-dire les mots qui ont été prédits le plus rapidement et qui devraient donc avoir très peu d'influence sur le changement de sens, étaient souvent des cohyponymes, des variations morphologiques ou des mots étroitement liés au mot de la requête d'origine.

D'autre part, pour les mots ayant obtenu les scores les plus élevés, c'est-à-dire les mots que CamemBERT a eu du mal à prédire et qui indiquent donc un changement significatif dans le sens des requêtes originales, les dix premières prédictions étaient principalement des signes de ponctuation, des changements de ressources et des mots appartenant à des domaines sémantiques spécifiques.

De plus, lorsque CamemBERT ne parvenait pas à trouver les mots masqués dans son vocabulaire, il prédisait souvent des mots de la même famille morphologique ou appartenant au même contexte sémantique que le mot masqué.

Ces observations soulignent l'importance des relations lexicales, des variations morphologiques et des mots similaires dans les prédictions de CamemBERT et, par conséquent, dans le changement de sens lui-même.

Pour répondre à mes questions de recherche, je constate que :

a) Bien que l'ajout, la modification ou la suppression d'un seul mot puisse avoir un impact sur le sens d'une requête, il semble que ce changement soit généralement limité. Un seul mot a plus souvent un impact sur le changement de sens de la requête lorsqu'il introduit un contexte très différent de celui établi dans la requête originale, comme le suggèrent les observations de l'analyse qualitative des prédictions de CamemBERT.

b) Comme le montrent les analyses linguistiques qualitatives et les tests statistiques, les facteurs linguistiques et contextuels tels que les relations lexicales, les variations morphologiques, les mots proches et la longueur de la requête originale jouent un rôle important dans les changements sémantiques résultant de l'ajout, de la modification ou de la suppression d'un mot. En d'autres termes, leur présence réduit l'impact sur le changement de sens.

La contribution de ma recherche réside dans la compréhension des changements sémantiques associés à la modification d'un seul mot dans les requêtes, et dans l'identification des facteurs linguistiques et contextuels qui y contribuent. Ces résultats peuvent être utilisés pour améliorer les moteurs de recherche, les systèmes de traduction automatique et d'autres applications de traitement du langage naturel, ainsi que par les utilisateurs pour s'adapter à la meilleure stratégie afin d'obtenir des résultats de meilleure qualité lors d'une recherche en ligne.

Cependant, il y a certaines limites à prendre en compte, telles que le fait que les résultats sont basés sur un jeu de données spécifique, ce qui peut limiter leur généralisation à d'autres contextes. En outre, l'analyse linguistique qualitative s'appuie sur les prédictions de CamemBERT comme point de référence pour un changement de sens, qui peut ne pas refléter la vérité, mais plutôt être lié au fonctionnement interne du modèle. Malgré ces limites, cette étude offre une base solide pour de futures recherches visant à approfondir notre compréhension de ces phénomènes.

Parmi les pistes de travail les plus prometteuses, il serait intéressant de demander à un large groupe de participants d'évaluer le changement de sens résultant de la modification d'un seul mot, puis de comparer leurs résultats avec ceux de CamemBERT. Cette approche nous permettrait d'explorer les différences entre la perception humaine et celle du modèle, offrant ainsi une meilleure compréhension du processus de compréhension du sens. En parallèle, une analyse quantitative et qualitative des facteurs linguistiques impliqués pourrait être réalisée afin d'approfondir notre compréhension des mécanismes sous-jacents aux changements de sens.

En conclusion, ma recherche a fourni des informations précieuses sur les changements sémantiques associés à la modification d'un seul mot dans les requêtes, et a identifié les facteurs linguistiques et contextuels pertinents. Ces connaissances peuvent contribuer au développement de systèmes de traitement automatique du langage naturel plus efficaces, et ouvrent des perspectives prometteuses pour les travaux futurs dans ce domaine en constante évolution.

7. Bibliographie

Adam, C., Fabre, C., & Tanguy, L. (2013). *Étude des relations sémantiques dans les reformulations de requêtes sous la loupe de l'analyse distributionnelle*.

Awadallah, A. H., Shi, X., Craswell, N., & Ramsey, B. (2013, octobre 1). *Beyond Clicks : Query Reformulation as a Predictor of Search Satisfaction*. ACM International Conference on Information and Knowledge Management (CIKM). https://www.microsoft.com/en-us/research/publication/beyond-clicks-query-reformulation-as-a-predictor-of-search-satisfaction/

Azzoug, W. (2014). *Contribution à la définition d'une approche d'indexation sémantique de documents textuels*. https://www.semanticscholar.org/paper/Contribution-%C3%A0-la-d%C3%A9finition-d%E2%80%99une-approche-de-Azzoug/c34934b94b5e906087dac728274dc3345c039453

Bagha, K. (2011). A Short Introduction to Semantics. *Journal of Language Teaching and Research*, *2*. https://doi.org/10.4304/jltr.2.6.1411-1419

Dosso, C., Moreno, J. G., Chevalier, A., & Tamine, L. (2021). *CoST : An annotated Data Collection for Complex Search*. 4455. https://doi.org/10.1145/3459637.3481998

Guelfand, G. (2013). Chapitre 6. La méthode des associations d'idées. In *Les études qualitatives* (p. 209-228). EMS Editions. https://www.cairn.info/les-etudes-qualitatives--9782847695519-p-209.htm

Hearst, M. A. (2009). *Search User Interfaces*. Cambridge University Press. https://doi.org/10.1017/CBO9781139644082

Huang, J., & Efthimiadis, E. N. (2009). Analyzing and evaluating query reformulation strategies in web search logs. *Proceedings of the 18th ACM conference on Information and knowledge management*, 77-86. https://doi.org/10.1145/1645953.1645966

Jurafsky, D., & Martin, J. H. (2023). Chapitre 6.1. Lexical Semantics. In *Speech and Language Processing*. https://web.stanford.edu/~jurafsky/slp3/

Lehmann, A., & Martin-Berthet, F. (2018). Chapitre 4. Les relations sémantiques. In *Lexicologie: Vol. 5e éd.* (p. 73-93). Armand Colin. https://www.cairn.info/lexicologie--9782200622374-p-73.htm

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., Clergerie, E. V. de L., Sagot, B., & Seddah, D. (2020). *Les modèles de langue contextuels Camembert pour le français : Impact de la taille et de l'hétérogénéité des données d'entrainement*. 54. https://hal.science/hal-02784755

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D., & Sagot, B. (2020). CamemBERT : A Tasty French Language Model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7203-7219. https://doi.org/10.18653/v1/2020.acl-main.645

Matthiessen, M. A. K. H., Christian M. I. M. (2013). The Architecture of Language. In *Halliday's Introduction to Functional Grammar* (4e éd.). Routledge.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed Representations of Words and Phrases and their Compositionality* (arXiv:1310.4546). arXiv. https://doi.org/10.48550/arXiv.1310.4546

Moreau, F. (2006). 2.2 Apport de connaissances morphologiques en RI. In *Revisiter le couplage traitement automatique des langues et recherche d'information*. Université Rennes 1. https://theses.hal.science/tel-00524514

Morris, J., & Hirst, G. (2004). Non-Classical Lexical Semantic Relations. *Proceedings of the Computational Lexical Semantics Workshop at HLT-NAACL 2004*, 46-51. https://aclanthology.org/W04-2607

Mothe, J., & Pai, S. (2017). *Mise en œuvre d'une base de données graphe pour l'analyse des logs de requêtes en recherche d'information.* https://doi.org/10.24348/CORIA.2017.1

Nettleton, D. (2014). Chapter e15—Analysis of Data on the Internet II – Search Experience Analysis. In D. Nettleton (Éd.), *Commercial Data Mining* (p. e15-e26). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-416602-8.00015-7

Pennington, J., Socher, R., & Manning, C. (2014). GloVe : Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543. https://doi.org/10.3115/v1/D14-1162

Russell, D. M. (2019). *Power Searching with Google Course*. http://dmrussell.net/PSWG6/PSWG6_1.5_Text_2019/PSWG6_1.5_Text_2019.html

Zargayouna, H., Roussey, C., & Chevallet, J. P. (2015). Recherche d'information sémantique : État des lieux. *Revue TAL*, *56*(3), 49.