Étude des variations linguistiques au sein des requêtes initiales lors d'une session de recherche sur le Web

Hugo Breton

Sous la direction de Ludovic Tanguy et Claire Ibarboure

Master 1 - Linguistique, Informatique, Technologies du Langage

Année 2025/2026

Remerciements :

Naturellement, ma gratitude se dirige en premier lieu vers mes directeurs de mémoire, Ludovic Tanguy et Claire Ibarboure, pour la bienveillance et la rigueur qu'ils ont eues à mon égard. Mes remerciements vont également au reste du corps enseignant du Master LITL pour leur accompagnement tout au long de cette année universitaire.

En second lieu, je souhaite remercier le reste de la promotion avec qui j'ai partagé cette année. J'espère avoir pu rendre un peu de la solidarité que j'ai pu recevoir de leur part. Je tiens particulièrement à exprimer ma gratitude à Elisa Tailhades et Adeline Gordien dont la présence a toujours été source de motivation et sans qui je ne serais pas parvenu jusqu'ici.

Enfin, je tiens à remercier ma famille et mes proches pour leur soutien. En particulier ma mère pour ses conseils toujours précieux, mon père pour le confort qu'il s'efforce de me garantir, ainsi que mes grands-parents pour leur soutien inconditionnel depuis le tout début.

Table des matières :

Introduction

Ce projet de recherche s'inscrit au sein du domaine de la Recherche d'Information (RI). Ce carrefour interdisciplinaire mêlant aussi bien linguistique, que psychologie cognitive ou informatique est en constante évolution. Celui-ci s'intéresse aux interactions entre un utilisateur doté d'un besoin d'information, et un système de recherche d'information (SRI) conçu pour répondre à ce besoin. Afin de communiquer avec ces systèmes, l'utilisateur a recours à des requêtes qu'il formule puis soumet à l'outil qu'il choisit. Une fois cela fait, le système, par un mécanisme de correspondance entre l'énoncé de la requête et une collection de documents indexés, renvoie l'utilisateur vers un certain nombre de ces derniers (Hiemstra, 2009). Ainsi, la formulation de ces requêtes constitue, pour l'utilisateur, l'unique moyen de répondre à ce besoin d'information. Ce mémoire s'intéresse particulièrement à la requête initiale formulée par l'utilisateur, à savoir celle qui initie l'interaction entre ce dernier et le système, pour un besoin d'information fixé.

On observe, pour un même besoin d'information, qu'un utilisateur peut retranscrire celui-ci par un grand nombre de moyens distincts. Ces différentes variantes ne sont que plus nombreuses et changeantes suite à l'avènement de la RI conversationnelle, propulsée par la popularisation massive et manifestement inévitable des intelligences artificielles conversationnelles. Les interactions découlant de ces dernières, bien que toujours motivées par un besoin d'information, ne sont en rien comparables à celles observables dans le cadre de l'utilisation d'un moteur de recherche, alimenté par un certain nombre de documents indexés. Ce projet de recherche s'inscrira alors dans une RI traditionnelle, préférant ainsi s'éloigner des nouvelles approches de recherche d'information. Il ne serait néanmoins pas impertinent que les méthodes d'analyse employées dans le cadre de cette étude soient pensées afin d'être potentiellement modifiées puis réutilisées pour mener une recherche similaire s'inscrivant en RI conversationnelle, le concept de requête initiale étant universel à chacune de ces approches.

Ainsi, nous exploitons le corpus PRIVaThe (Ibarboure, 2025) qui contient en son sein plus de 3000 requêtes, dont 300 requêtes initiales, soumises par 100 participants au cours de trois tâches de recherche différentes. Ainsi, les participants devaient répondre à un énoncé formulé en langage naturel, qui contenait lui-même un objectif de recherche correspondant à un besoin d'information. Notre hypothèse est que les différentes requêtes initiales présentes dans ce corpus vont se distinguer des autres productions, à savoir les diverses reformulations. En effet, en raison de son caractère racine, la première requête correspond à une tentative exploratoire de l'utilisateur quant à la résolution de son besoin d'information (Marchionini, 2006). Ce dernier n'étant pas encore pleinement stabilisé, nous pouvons imaginer que les requêtes initiales présenteront des différences formelles et lexicales par rapport aux multiples reformulations effectuées par les participants. Notre objectif sera donc d'automatiser un traitement systématique de chacune des requêtes du corpus PRIVaThe. Ce traitement nous permettra ainsi de mettre en avant, pour chacune d'entre elles, un ensemble de caractéristiques formelles et lexicales. Une fois cela fait, nous pourrons ainsi opérer une analyse structurelle et lexicale de l'ensemble des requêtes initiales présentes dans le jeu de données PRIVaThe. Une fois ces multiples caractéristiques décrites, nous pourrons comparer les requêtes initiales aux reformulations afin d'observer de potentielles différences entre ces deux éléments et alors vérifier ou non notre hypothèse.

Ce mémoire sera alors organisé en plusieurs parties distinctes. En premier lieu, nous présenterons les fondements théoriques nécessaires à la réalisation de ce projet de recherche. Nous évoquerons les notions de besoin d'information et de session de recherche avant de finalement nous attarder sur la requête initiale. En deuxième lieu, nous présenterons le jeu de données PRIVaThe, qui se trouvera au centre du traitement mais également de nos analyses. Nous effectuerons alors une description quantitative et qualitative de ce dernier, en continuité des travaux de Ibarboure (2025). En troisième lieu, nous proposerons un traitement bidimensionnel et automatisé des requêtes. Nous nous pencherons alors sur les caractéristiques structurelles que présentent ces dernières avant de nous focaliser sur l'organisation lexicale de chacune des requêtes. Enfin, nous mettrons à profit ce traitement afin de répondre à deux problématiques de recherche distinctes. Quelles sont les caractéristiques formelles et lexicales des requêtes initiales lorsqu'un utilisateur doit répondre à un besoin d'information contenu dans un énoncé en langage naturel ? Mais également, quelles continuités ainsi que quelles ruptures peut-on observer entre la requête initiale et l'ensemble des reformulations ?

1. Du besoin d'information à la formulation de la requête : État de l'art

Pour rappel, notre objet d'étude concerne les diverses variations affectant les requêtes initiales lors d'une session de recherche sur le Web. Avant de nous attarder plus en détail sur différents travaux partageant cet objet de recherche, il semble pertinent de décomposer le processus de recherche d'information. En premier lieu, nous observerons que l'existence d'un besoin d'information est un prérequis obligatoire pour donner naissance à une session de recherche, mais également que ce dernier peut prendre différentes formes. Puis, en second lieu, nous nous pencherons sur le fonctionnement des systèmes de recherche d'information, avant de nous intéresser en détail aux requêtes en découlant.

1.1. Le besoin d'information

Le domaine de la RI, se penchant entre autres sur la manière dont les humains interagissent avec l'information dans une situation de recherche, communique logiquement avec plusieurs disciplines, dont la psychologie cognitive. Cette dernière s'intéresse alors notamment au besoin d'information possédé par un utilisateur, qui constitue alors l'élément déclencheur de l'initiative de recherche. Définir clairement la notion de besoin d'information est alors crucial dans le cadre d'une étude qui vise à se focaliser sur différentes variations affectant les différents moyens d'expression d'un même besoin.

1.1.1. Définition globale

Wilson (1981) souligne alors la difficulté mais également l'importance de définir clairement cette notion. Selon lui, il était difficile de définir le concept de "besoin d'information" parce que la notion même "d'information" n'était pas clairement définie. En effet, en particulier lors de sessions de recherche sur le Web, les utilisateurs cherchent des informations de nature très différentes les unes des autres. Simplement différencier "données", "informations" et "savoirs" ne permet donc pas de rendre compte des diverses motivations distinguant un utilisateur d'un autre. De plus, la cohabitation de "faits", "conseils" et "opinions" au sein des systèmes de recherche complique encore plus la compréhension claire du terme "information" dans des écrits qui ne stipulent souvent pas dans quel sens le terme est employé.

Wilson segmente alors le processus de déplacement informationnel lors d'une opération de recherche. Ce cycle, initié par le besoin d'information d'un utilisateur, se répète suite à une ou plusieurs interactions avec un système de recherche ou un autre individu. Le besoin initie alors un comportement de recherche qui se réitère continuellement. Véritablement, un être humain est constamment motivé par ce besoin d'information. Il conclut en affirmant que le terme "besoin d'information" est ainsi inévitablement source de confusions et d'imprécisions, et propose le terme "recherche d'information pour satisfaire des besoins" en guise de substitution (Wilson, 1981).

Nombreuses et diverses sont alors les tentatives de définition de cette entité. Latour définit le besoin d'information comme le résultat d'une "prise de conscience par le sujet d'un manque de connaissances qui lui est nécessaire pour la résolution d'un problème ou pour l'atteinte d'un objectif visé et ce dans une situation donnée" (Latour, 2014, p. 2). Simonnot présente cette même notion comme "une sensation qui porterait l'individu à s'engager dans une activité de recherche d'information" (Simonnot, 2006, p. 41). Il apparaît donc délicat de définir clairement ce concept en particulier quand la nature même de l'information, ainsi que des moyens d'y accéder, évoluent constamment. Dans notre cas, le besoin d'information sera défini comme un objectif de recherche composé d'une ou plusieurs thématiques, et prenant la forme d'un énoncé en langage naturel. Dans ce cas de figure, le besoin d'information est alors matérialisé par une structure syntaxique et sémantique déjà existante, bien que cet énoncé puisse faire l'objet d'interprétations diverses de la part d'un utilisateur.

Il est également pertinent de rappeler qu'un utilisateur n'a à proprement parler pas besoin d'un besoin d'information précis pour entamer une session de recherche. Néanmoins, la formulation d'un besoin plus spécifique ou la possession de connaissances plus amples sur le ou les sujets associés lui permettront d'accéder à des résultats plus pertinents vis-à-vis de ce même besoin. Selon Hearst (2009), moins le besoin d'information et les concepts associés seront familiers à l'utilisateur, moins il sera aisé pour ce dernier de communiquer avec le système afin de trouver satisfaction.

1.1.2. Formes et expressions du besoin d'information

Une fois cette notion définie, il est important de mentionner que ce besoin d'information peut prendre différentes formes. En effet, un besoin d'information peut se traduire par différents types d'intentions de recherche ainsi que par différents comportements informationnels lorsqu'un utilisateur interagit avec un SRI.

Dans cette idée, Broder (2002) affirme que toutes les requêtes ne relèvent pas du même besoin cognitif. Il propose ainsi une typologie en prenant en compte le type de besoin des utilisateurs correspondant à l'intention ayant initié le processus de requête. Il distingue trois types de requêtes. Les requêtes navigationnelles **(1)**, celles où l'utilisateur poursuit l'objectif d'accéder à un site Internet spécifique dont il connaît ou suppose l'existence. Les requêtes informationnelles **(2)**, qui concernent les cas où l'utilisateur cherche à obtenir une ou plusieurs informations qu'il suppose également être disponibles. Enfin, les requêtes transactionnelles **(3***),* correspondent à un besoin de l'utilisateur de réaliser une action spécifique en accédant à un site qui propose ce service. Il parvient à la conclusion que près de 50% des requêtes formulées sur le Web prennent la forme de requêtes informationnelles.

(1) *wikipedia*
(2) *nom de la deuxième fille de hershel greene walking dead*
(3) *downloading files*

Dans cette perspective et en continuité, Rose & Levinson (2004) proposent une typologie visant à répondre à une catégorisation similaire, cette fois-ci en se focalisant sur l'élément déclencheur du besoin d'information. Sont ainsi distinguées les requêtes navigationnelles, informationnelles et de ressource. Les premières correspondent, comme pour Broder (2002), à des requêtes visant à accéder à un site spécifique dont l'utilisateur suppose ou connaît l'existence. L'utilisateur, ne connaissant pas l'URL exacte, ou bien souhaitant simplement gagner du temps, formule une requête renvoyant à l'élément choisi. Les secondes, à savoir les requêtes dites informationnelles, se rapprochent également en principe à celles définies par Broder. Rose & Levinson choisissent néanmoins de créer des sous-catégories permettant de différencier les requêtes visant à accéder directement à une information. On distingue alors les requêtes dirigées **(4)** qui visent à obtenir une information précise, de celles dites non-dirigées **(5)** où le besoin d'information n'est pas encore totalement déterminé. Sont également différenciées les requêtes informationnelles visant à accéder à un conseil, une localisation, ou à une liste. Enfin, les requêtes renvoyant à une intention par l'utilisateur d'obtenir l'accès à une ressource sont de ce fait nommées "requêtes ressources" **(6)**. Ces dernières sont également séparées en plusieurs sous-catégories en fonction de l'intention de l'utilisateur. Selon, par exemple, si celui-ci souhaite télécharger une ressource ou bien seulement la consulter.

(4) *what is a supercharger*
(5) *color blindness*
(6) *measure converter*

On observe que ces facteurs peuvent fondamentalement altérer la forme de la requête qui sera soumise par l'utilisateur, comme le démontrent les exemples **(1)** et **(2)** tirés du jeu de données utilisé dans cette étude. Comme mentionné précédemment, le besoin d'information sera, dans notre cas, induit par un énoncé en langage naturel comportant un objectif de recherche clairement délimité. Néanmoins, selon le traitement de cet énoncé par l'utilisateur, la requête formulée pourra être aussi bien dirigée que non dirigée. Il est donc intéressant d'observer, pour un même besoin, la diversité des mécanismes de traitement de l'information opérés par l'utilisateur. La requête initiale constituant la première mise en application concrète du besoin d'information, elle représente un point d'observation privilégié de sa traduction en recherche.

1.2. La recherche d'information

Une fois ce besoin d'information identifié, l'utilisateur doit maintenant choisir un outil lui permettant de répondre à celui-ci. Cet outil lui permettra de réaliser une session de recherche dans l'objectif d'assouvir son besoin d'information. Dans cette section nous présenterons le fonctionnements des systèmes de recherche d'information ainsi que le déroulé d'une session de recherche.

1.2.1. Processus de recherche

Une fois ce besoin d'information identifié, l'utilisateur doit maintenant choisir un outil lui permettant de répondre à celui-ci. Ces outils conçus pour répondre à des besoins d'informations ont évolué avec le temps. Autrefois réservés à des catégories spécifiques d'utilisateurs, ils sont maintenant pleinement intégrés dans le quotidien du grand public (Hearst, 2009). En réponse à cette démocratisation ainsi qu'aux avancées technologiques, les systèmes ont eux aussi évolué. Notamment en permettant une indexation automatique des documents ou encore en incluant des modèles d'analyse linguistiques plus performants afin de maximiser les correspondances entre les requêtes et les résultats en découlant.

Comme explicité dans la Figure 1, chacun de ces systèmes existants est alimenté par une banque de documents indexés dans laquelle il puise lors de chaque requête qui lui est soumise, en correspondance avec les informations comprises dans celle-ci (Hiemstra, 2009).

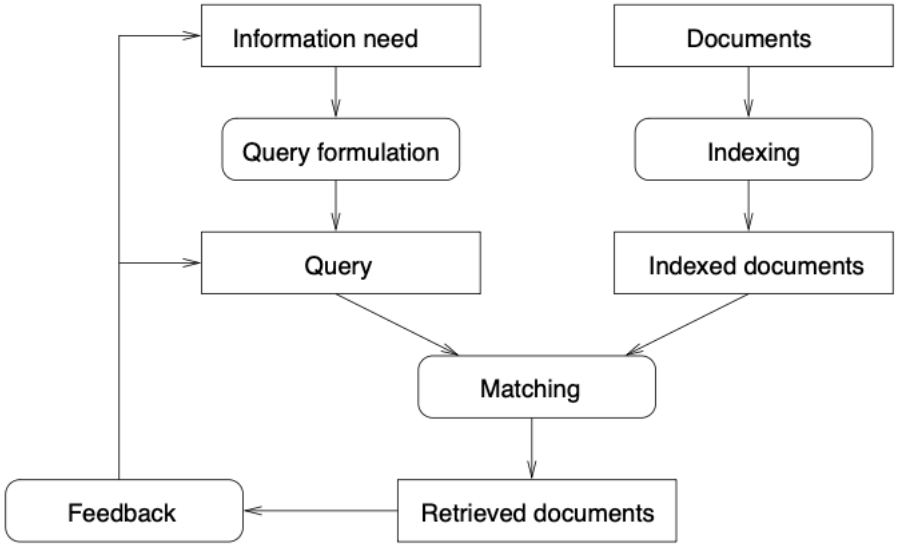

Figure 1 - Processus de recherche d'information, du point de vue du l'utilisateur, ainsi que du système (Hiemstra, 2009)

Ce phénomène de “matching“, c’est-à-dire la tentative du système de faire correspondre une partie de sa base de données avec la requête qui a été soumise, est au centre du processus de recherche. Si la liste de documents récupérés permet de répondre au besoin de recherche initial, alors la recherche d’information prendra logiquement fin. Dans le cas contraire, l’utilisateur devra faire appel à une révision de sa première requête en prenant en compte la requête initiale qu’il aura formulée ainsi que les résultats obtenus après le matching. Lorsque les résultats obtenus ne permettent pas de satisfaire son besoin d'information, l'utilisateur est alors amené à reformuler sa requête. Ce processus itératif, par lequel l'utilisateur ajuste successivement ses requêtes en fonction des résultats obtenus, constitue ce que Huang & Efthimiadis (2009) définissent comme la reformulation. Un utilisateur peut alors reformuler son besoin d’information jusqu’à obtenir satisfaction ou décider d’abandonner. Cette succession d’interactions entre l’utilisateur et le système forme alors une session de recherche.

1.2.2. Sessions de recherche complexes

La session de recherche, définie comme l'ensemble des interactions entre un utilisateur et un système autour d'un même besoin d'information, ne se limite cependant pas toujours à un enchaînement linéaire d'interactions visant un objectif unique. En effet, le processus de collecte d’information s’inscrit souvent dans le cadre de sessions dites complexes, impliquant plusieurs sous-tâches adjacentes et induisant des interactions de recherche successives entre l’utilisateur et le système (Dosso et al., 2021). La présence de plusieurs sous-tâches composant un même objectif de recherche va alors favoriser un nombre d’itérations plus importants du cycle décrit dans la Figure 1. Ainsi, à l’échelle de la session de recherche, l’utilisateur est constamment incité à réaliser un certain nombre de choix tels que revenir à une étape précédente ou bien simplement abandonner si les résultats obtenus ne lui donnent pas satisfaction (Sharit et al., 2008).

Dans le cadre de cette étude, les données sur lesquelles nous nous appuyons ont été collectées en contexte expérimental. La Figure 2 permet de visualiser concrètement le déroulé d'une session de recherche en contexte expérimental (Ibarboure, 2025). Celui-ci s'articule autour de deux phases distinctes. Une première phase de planification, durant laquelle l'utilisateur prend connaissance de l'énoncé correspondant à son objectif de recherche, avant de définir un objectif principal puis, le cas échéant, les sous-objectifs qui en découlent. S'ensuit une phase d'exécution au cours de laquelle l'utilisateur formule une requête qu'il soumet au moteur de recherche, qui lui renvoie en retour une page de résultats depuis laquelle il peut accéder aux pages web correspondantes. À l'issue de cette consultation, plusieurs issues sont alors envisageables. Une réponse satisfaisante met fin à la session ou au traitement du sous-objectif en cours, tandis qu'une insatisfaction entraîne soit une reformulation de la requête, soit un retour à la phase de planification. L'identification d'un sous-objectif annexe peut par ailleurs amener l'utilisateur à réorienter temporairement sa recherche. L'abandon constitue enfin une issue supplémentaire, intervenant lorsque l'utilisateur choisit de mettre fin à la session sans avoir obtenu satisfaction. Il est important de noter que ces différentes issues peuvent se manifester aussi bien après la consultation d'une page de résultats qu'après l'analyse d'une page web.

Dans cette même logique, et dans l’éventualité d’une session de recherche à caractère complexe, on peut ainsi distinguer différents comportements utilisateurs qui vont fortement influencer le processus de recherche d’information. Dans ce contexte, Ibarboure (2025) distingue plusieurs types de comportements selon les cas où l’utilisateur, lors d’une session de recherche complexe, segmente ou non la tâche en sous-objectifs distincts au moment de soumettre la requête. Ainsi, un utilisateur ayant un comportement dit logique traitera un seul sous-objectif à la fois, préférant segmenter la session de recherche en conséquence. Un comportement défini comme exploratoire correspondra à une approche où l’utilisateur segmente également la tâche en sous-objectifs. Cependant, il choisit de naviguer librement entre les deux sous-tâches, bien qu’il ne les traite qu’une seule à la fois. Au contraire des deux approches précédentes, un comportement dit frontal choisira de soumettre une ou plusieurs requêtes en lien avec l’ensemble des sous-objectifs. L’ensemble de ces éléments va naturellement influencer les propriétés formelles et logiques des requêtes initiales comme celles découlant d’une reformulation.

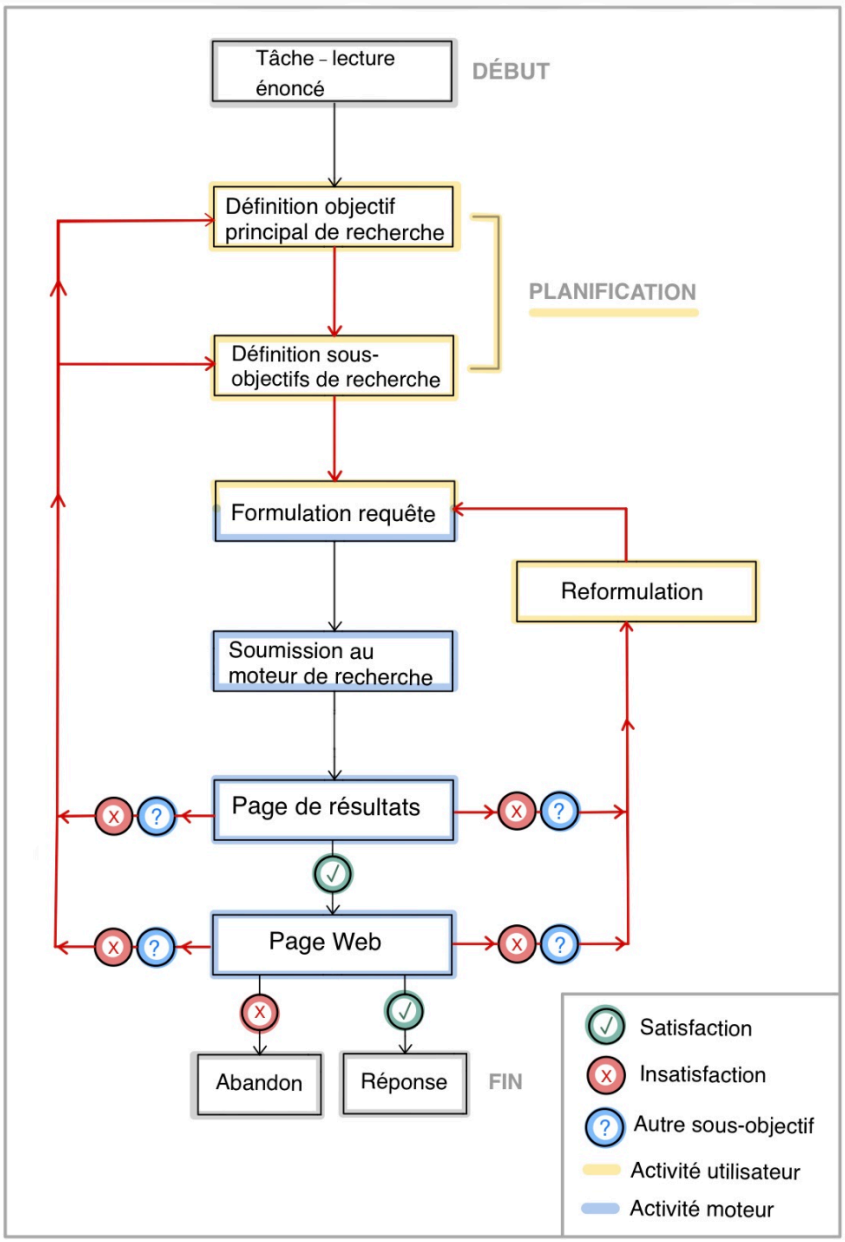

Figure 2 - Réalisation d'une session de recherche en contexte expérimental (Ibarboure, 2025)

1.3. La requête : Caractéristiques générales et spécificités de la requête initiale

La requête constitue une composante indissociable du processus de recherche formant une session de recherche. Elle est la matérialisation du traitement du besoin d'information par l'utilisateur qui soumet ensuite cette dernière au SRI. Dans cette section, nous aborderons d'abord la requête sous un angle plus global en mentionnant notamment les différentes formes qu'elle peut prendre. Ensuite, nous nous attarderons plus en détail sur ce qui caractérise une requête initiale, au cœur de notre objectif de recherche.

1.3.1. Types de requêtes

L'évolution des SRI s'est fortement accélérée conjointement à l'apparition et à la démocratisation du Web, cela dès les années 90. Ces outils, autrefois réservés à des usages professionnels et spécifiques, étaient désormais mis entre les mains du grand public, entraînant une variété et un nombre d'usages bien supérieurs à ce qu'il était.

Cette augmentation massive du nombre d'utilisateurs, entraînant inévitablement une augmentation similaire des données observables, a permis une étude approfondie des comportements utilisateurs lors de leurs sessions de recherche. Ainsi, les nombreuses évolutions ayant marqué l'histoire de ces systèmes ont eu une incidence indéniable sur les comportements des utilisateurs, comme l'affirme Hearst (2009). Selon elle, les requêtes précédant l'arrivée du Web étaient souvent longues et assemblées avec une grande attention, ce service étant généralement payant. Ces mêmes comportements utilisateurs devenant plus imprévisibles, elle distingue alors les requêtes qui contiennent des opérateurs booléens, à celles composées de commandes, et celles soumises en langage naturel (Hearst, 2009). Il est donc aisé de différencier, sur la forme, les requêtes formulées à l'aide de mots clés de celles qui se composent de langage naturel. Il est également envisageable pour un utilisateur de mélanger ces différents moyens d'expression.

De la même manière, Strohmaier et al. (2008) distinguent deux types de requêtes selon le caractère explicite ou implicite de ces dernières. Une requête explicite du type « *aller en Inde* » ne sera pas perçue par le système de la même manière qu'une requête implicite telle que « *Inde* » qui nécessitera sûrement une reformulation. Les requêtes explicites seront alors formulées en langage naturel alors que les requêtes dites implicites auront tendance à être formées à l'aide d'une concaténation de mots clés, en accord donc avec la typologie de Hearst (2009).

Ces différentes stratégies de recherche ne représentent souvent que des manifestations inconscientes de l'utilisateur d'optimiser la pertinence du résultat lors du matching. Cependant, comme cela a été démontré en *section 1.1.2*, la nature de la recherche, en accord avec la catégorisation de Broder (2002) ou bien de Rose & Levinson (2004), pourra notamment avoir une incidence sur la forme de la requête. Le Tableau 1 correspond à un ensemble d'extraits de sessions de recherche tirés du jeu de données PRIVaThe (*présenté plus en détail dans la section 2.1*). Les colonnes du Tableau 1 indiquent respectivement l'identifiant de la session de recherche, le contexte de la requête au sein de cette dernière, ainsi que l'énoncé exact de la requête en question.

	Contexte	Requête	Type de requête
(7)	Requête initiale	*quel est le nom de la deuxieme fille de hershel greene dans walking dead*	Requête informationnelle explicite formulée en langage naturel
(8)	Requête initiale	*wikipedia*	Requête navigationnelle implicite formulée en mots-clés
(9)	Requête initiale	*pete docter henri guybet*	Requête informationnelle implicite formulée en mots-clés
(10)	Reformulation 1	*"pete docter" + "henri guybet"*	Requête informationnelle implicite formulée en mots-clés et en opérateurs

Tableau 1 - Extraits de sessions de recherche tirées du corpus PRIVaThe

Comme nous pouvons le voir, la requête initiale de la session de recherche correspondant à l'exemple **(7)** est formulée en langage naturel et est à caractère explicite et informationnel. En opposition, pour un besoin d'information strictement similaire, la requête initiale représentée par l'exemple **(8)** affiche une stratégie de recherche diamétralement opposée. Il s'agit ici d'une requête formulée en mot-clé qui découle d'un besoin navigationnel. Son unique élément consiste alors en l'énoncé du site Internet qui vise à être accédé, néanmoins il s'agit d'un élément implicite qui nécessitera potentiellement une reformulation. L'utilisateur, afin d'accéder à une page Web plus adaptée à son objectif de recherche, a adjoint un élément qu'il lui semblait pertinent afin d'optimiser le matching. De la même manière, la requête initiale liée à l'exemple **(9)** est composée d'une concaténation de mots-clés implicites, cette fois dans une approche informationnelle. Néanmoins, voyant que les résultats du matching ne lui permettaient pas d'assouvir son besoin, l'utilisateur a choisi d'effectuer une reformulation de sa requête **(10)** en utilisant un opérateur, ici « + ». De plus, il met à profit sa connaissance du fonctionnement du système en segmentant les mots clés qu'il avait précédemment soumis en deux entités distinctes, cela grâce aux opérateurs « " ».

Ces différents exemples nous montrent que la requête, en qualité de témoin de l'intention et du besoin d'un utilisateur, peut prendre des formes radicalement différentes. Ainsi, les requêtes offrent alors, à l'utilisateur, l'opportunité de naviguer au sein d'un espace sémantique (Mitra, 2015). Comme mentionné précédemment, ces dernières étaient originellement longues et pleinement représentatives du besoin d'informations, avant d'évoluer suite à l'apparition du Web en se raccourcissant. Cette tendance semble de nouveau s'inverser comme le souligne Alaofi et al. (2022). Tendance dont nous pouvons théoriser la corrélation avec le fait que si un utilisateur possède des connaissances dans les domaines de l'informatique ou du Web, il aura alors tendance à formuler des requêtes plus longues et donc à la forme plus variable (Aula, 2003). La popularisation massive et progressivement universelle de ces outils aurait inévitablement accru la maîtrise globale de ces outils.

1.3.2. La requête initiale

En raison de sa proximité avec le besoin d'information initial, la première requête soumise par l'utilisateur en est logiquement sa traduction la plus littérale et spontanée. Bien que ce constat soit partagé par l'ensemble de la littérature, et que l'étude de la variation ait constitué un des sujets centraux en RI, la question des variations présentes au sein de requêtes initiales a été relativement peu exploitée. On peut alors s'interroger sur l'ampleur des variations qui peuvent affecter un ensemble de requêtes initiales pour un même besoin donné. On définira la formulation en requête comme la première itération de cette interaction entre l'utilisateur et le système, qui prend donc la forme d'une requête initiale. Bien qu'issues d'un même besoin d'information, les requêtes initiales peuvent prendre des formes particulièrement diverses selon l'interprétation qu'en fait l'utilisateur. La traduction en recherche d'un objectif pourtant identique peut ainsi varier sensiblement en fonction des mécanismes individuels de sélection, de compréhension et de reformulation mobilisés lors de cette première interaction avec le système (Abu Onq, 2023).

En effet, l'utilisateur commence souvent le processus de recherche avec une représentation incomplète de la stratégie qu'il va mettre en place lors de la session de recherche. Ainsi, la première interaction avec le système correspond donc à une entrée exploratoire dans l'espace informationnel, plutôt qu'à une expression stable et complète du besoin (Marchionini, 2006). Dans cette optique, Aula (2003) démontre que la requête initiale est généralement courte et peu descriptive. Les résultats récoltés sur 32 utilisateurs devant répondre à 20 tâches de recherche (soit 640 requêtes initiales au total) montrent que les utilisateurs initient les sessions à l'aide de formulations minimales, prenant souvent la forme de concaténation de mots-clés jugés représentatifs du domaine de recherche. En effet, elle observe que les requêtes ont en moyenne une taille de 3,0 mots, et attribue ce résultat au fait que la requête initiale intervient dans une phase encore exploratoire du processus de recherche, en accord avec les observations de Marchionini (2006).

Pour répondre aux difficultés rencontrées lors de la formulation en requête, l'utilisateur met en place un certain nombre de stratégies. Cependant, ces stratégies varient fortement d'un utilisateur à un autre. C'est ce que démontrent Abu Onq et al. (2025) notamment par le biais de l'application d'une typologie qui catégorise ces stratégies. Ils affirment que la formulation en requête d'un besoin d'information constitue une activité complexe pour l'utilisateur, qui demande de manière inhérente un effort cognitif important. Récoltées dans un cadre expérimental, plus de 10 000 requêtes provenant de 263 utilisateurs différents ont été collectées dans le cadre de la constitution du corpus UQV100. Ce corpus est alors composé de 100 tâches de recherche distinctes nommées "backstories". Les utilisateurs devaient répondre à des objectifs de recherche divers classés selon le niveau de complexité de la tâche. Les objectifs de recherche plus simples étaient différenciés de nature plus complexes, comportant par exemple plusieurs sous-objectifs que les utilisateurs devaient traiter.

A ensuite été appliquée, sur ces mêmes données, une typologie différenciant stratégies de modification et stratégies d'enrichissement. Les stratégies de modification correspondent à des changements opérés par rapport à l'énoncé correspondant au besoin d'information. Cela concerne par exemple la modification d'un terme en abréviation ou bien encore l'utilisation d'un synonyme. Les stratégies d'enrichissement concernent les différents ajouts réalisés par l'utilisateur lors de la formulation en requête. Ces termes permettent à l'utilisateur d'augmenter la pertinence de sa requête et ainsi maximiser les chances d'un résultat satisfaisant. Abu Onq et al. (2025) proposent ainsi la typologie de stratégies de recherche suivante :

- **Stratégies de modification** :
  - Abréviation : L'utilisateur choisit ici de transformer par un phénomène d'abréviation un élément déjà présent dans l'objectif de recherche.
  - Transformation : Transformation de nature morphologique d'un élément présent dans l'énoncé correspondant à l'objectif de recherche. Cela peut prendre la forme d'une opération flexionnelle, dérivationnelle ou compositionnelle.
  - Faute d'orthographe : Opération involontaire qui concerne une erreur typographique comme la substitution d'un caractère par un autre.
  - Opération sémantique : Modification d'un terme présent dans l'énoncé de l'objectif de recherche à l'aide d'un phénomène de synonymie ou de généralisation. Cette stratégie peut également prendre la forme de l'utilisation d'un terme plus spécifique.
  - Antonymie : Correspond à une modification d'un terme présent dans l'énoncé par son antonyme.
  - Opérateur : Modification par le remplacement d'un terme par un opérateur correspondant à un caractère spécial. Par exemple, le terme « *et* » pourra être remplacé par le caractère « + ».

Stratégie	Exemple
Abréviation	*doctor → dr*
Transformation	*history → historical*
Faute d'orthographe	*syndrome → syndrone*
Opération sémantique	*get → obtain*
Antonymie	*south → north*
Opérateur	*south or north → south/north*

Tableau 2 - Exemples de manifestations des stratégies de modifications (Abu Onq et al., 2025)

- **Stratégies d'enrichissement** :

  - Modificateur : Un terme, prenant généralement la forme d'un adjectif, ajouté afin de spécifier le sens d'un élément présent dans l'énoncé de l'objectif de recherche.
  - Ressource : Un terme rajouté afin d'accéder à un site Internet déterminé, ces ajouts prennent alors la forme de références spécifiques telles que « *Wikipédia* » ou « *Amazon* ».
  - Type d'information : Un terme propre à la requête qui vise à indiquer au système la forme de l'information qui souhaite être obtenue, comme c'est le cas de termes tels que « *définition* » ou « *photo* ».
  - Entité : Ajout d'un terme par l'utilisateur que ce dernier trouve pertinent quant au domaine lié à l'objectif de recherche. Cet ajout peut alors prendre des formes extrêmement diverses en fonction du besoin d'information.
  - Impératif : Utilisation d'une forme verbale à l'impératif afin de s'adresser au système de manière directe. Tendance liée au développement de la RI conversationnelle et à la popularisation de formes impératives telles que « *donne moi* » ou « *liste* » dans les requêtes.
  - Hors-sujet : Un ajout autre qui n'est sous aucun aspect en lien avec le domaine propre à l'objectif de recherche, prend souvent la forme d'ajouts tels que « *etc* ».

Stratégie	Exemple(s)
Modificateur	*"current"* dans *"current price"*
Ressource	*"amazon" ; "wiki"*
Type d'information	*"list" ; "review" ; "music"*
Entité	*"dates" ; "english" ; "cafeteria"*
Impératif	*"give me" ; "show"*
Hors-sujet	*"etc"*

Tableau 3 - Exemples de manifestations des stratégies d'enrichissement (Abu Onq et al., 2025)

Les résultats obtenus démontrent que près de 70 % des termes composant les requêtes n'apparaissaient pas originellement au sein des énoncés. De plus, les objectifs de recherche considérés comme plus complexes apparaissent comme plus propices à ces ajouts ainsi qu'à d'autres variations linguistiques en raison de requêtes plus longues. De plus, on peut clairement observer, grâce à la mise en place de la présente typologie, que certaines stratégies sont nettement plus représentées que d'autres. Le Tableau 4 explicite ainsi l'ensemble des résultats obtenus en indiquant respectivement : la stratégie en question, le nombre de sujets utilisant la stratégie, le nombre total d'usages, la fréquence moyenne d'utilisation pour un même utilisateur, le pourcentage de sujets ayant utilisé la stratégie et le nombre de tâches dans lesquelles la stratégie est présente. Les stratégies précédées par un "M" correspondent à des stratégies de modification, celles précédées par un "E" appartiennent à la catégorie des stratégies d'enrichissement. À noter que les stratégies sont ordonnées selon la fréquence moyenne d'utilisation d'une même stratégie par utilisateur.

Category	Unique Workers Using Category	Total Times Used	Avg. Category Contribution per Worker	% Unique Workers (263 in total)	Number of Backstories
M-Antonym	25	27	1.08	10%	8
M-Operator	11	13	1.18	4%	11
E-Imperative	9	12	1.33	3%	9
M-Misspelling	21	31	1.48	8%	23
M-Abbreviation	87	162	1.86	33%	23
E-Resource	66	136	2.06	25%	50
E-Off-topic	109	278	2.55	41%	76
E-Entity	159	427	2.69	60%	66
M-Transformation	218	725	3.33	83%	55
E-Modifier	149	568	3.81	57%	82
E-Information-type	197	829	4.21	75%	83
M-Semantic	228	1,536	7.74	87%	95

Tableau 4 - Résultats obtenus après l'annotation des stratégies sur l'ensemble des requêtes initiales (Abu Onq et al., 2025)

Ainsi, la stratégie d'opération sémantique est la plus fréquente avec une utilisation dans 95 tâches différentes. 87 % des sujets l'ont employée pour un total de 1 556 fois. La stratégie d'enrichissement de type d'information est également particulièrement fréquente, en témoignent les 75 % d'utilisateurs qui y ont fait appel pour un total de 829 utilisations. Au contraire, d'autres stratégies apparaissent comme bien moins intuitives, comme c'est le cas de l'ajout d'opérateurs ou de formes à l'impératif. Une application de cette typologie sur un segment du corpus PRIVaThe est présentée en *section 2.2*. Les variations présentes dans les requêtes, et ici plus spécifiquement les requêtes initiales, peuvent alors être représentées dans des structures bien précises (Abu Onq et al., 2025). La prise en compte de ces structures par les SRI pourrait alors fortement bénéficier au bon fonctionnement de ces derniers.

2. Présentation des données et perspectives de recherche

Ce chapitre vise à présenter le jeu de données utilisé dans le cadre de cette étude ainsi que les différentes problématiques et hypothèses de recherche qui nous guideront.

2.1. Constitution du jeu de données PRIVaThe

Le jeu de données PRIVaThe (Parcours de Recherche d'Information avec Variations Thématiques), constitué par Ibarboure (2025), sera le corpus sur lequel nous fonderons nos analyses. Les données qui le constituent ont été rassemblées afin de réaliser une étude liée aux variations thématiques caractérisant les comportements utilisateurs lors d'une session de recherche complexe. Cette ressource contient ainsi 3 162 requêtes formulées par 100 participants dans un cadre expérimental. Cet ensemble de requêtes provient de deux tâches de recherche distinctes, chacune composée de deux thématiques différentes, et conçues pour amener l'utilisateur vers une session de recherche complexe. De plus, nous exploiterons aussi les données d'une troisième tâche de recherche, cette fois-ci moins complexe et conçue afin de servir d'entraînement avant la participation au processus de collecte. Cette tâche a donné lieu à la soumission de 203 requêtes, toujours pour 100 participants, portant le nombre total de requêtes observables à 3 365. Ce jeu de données nous donne alors accès à 300 sessions de recherche distinctes, correspondant logiquement au même nombre de requêtes initiales, ces dernières réparties dans des objectifs de recherche de complexité variable.

Les sujets devaient donc, dans un temps limité de 20 minutes par tâche, réaliser trois sessions de recherche afin de répondre à ces trois objectifs de recherche. Le seul outil dont ils avaient à disposition était le moteur de recherche Google. À noter que l'énoncé correspondant au besoin d'information ne comportait qu'une seule tâche à la fois et était fourni sous la forme d'un papier physique. Il était donc impossible d'utiliser un raccourci informatique afin de "copier-coller" l'énoncé directement vers le moteur de recherche. À noter que certains participants ont verbalisé les différentes actions qu'ils ont pus effectuées, mais ces dernières n'ont pas été exploitées.

- **Tâche d'entraînement (TWD) :**

  - "Quel est le nom de la deuxième fille de Hershel Greene dans la série Walking Dead ?"

La tâche d'entraînement se distingue des deux suivantes car elle ne vise aucunement à répondre aux mêmes critères de constitution. Elle n'est composée que d'une seule thématique et peut, en théorie, être résolue en une seule et unique requête. Par conséquent, elle n'est pas concernée par la limite de temps précédemment mentionnée.

Concernant les deux tâches de recherche principales. Elles ont été imaginées afin d'être composées de deux sous-objectifs de recherche distincts. Distinction qui repose sur une approche thématique. De plus, elles sont chacune conçues pour inciter l'utilisateur à faire plusieurs requêtes, pour trouver les informations ou pour les vérifier. Une durée maximale de 20 minutes est imposée pour chacune des deux tâches. Cela également dans le but de limiter les temps de lecture pour favoriser la formulation de requêtes. De plus, ces tâches ont la particularité d'exister sous deux formes différentes. Il existe deux variantes de ces deux tâches qui comportent les mêmes informations mais agencées différemment. Concrètement, les deux thématiques sont inversées selon la variante qui a été attribuée à l'utilisateur. L'objectif est d'observer si certains éléments de l'énoncé vont avoir tendance à être traités en priorité, malgré leurs positions dans celui-ci. À noter que la répartition des deux variantes est totalement symétrique : 50 utilisateurs se sont vus attribuer la variante 1 et 50 autres la variante 2. Les différentes variantes des deux tâches principales sont ainsi présentées ci-dessous, les thématiques de recherche correspondent aux éléments soulignés.

- **Tâche 1 (TMR) - Variante 1 :**

- "D'après la mythologie grecque, le père d'Héraclès s'est transformé trois fois en animal afin de séduire des femmes mortelles (humaines). Ces transformations ont été le sujet de nombreuses peintures de la Renaissance italienne. Pour compléter cette tâche vous devez remplir le tableau suivant en précisant pour les trois cas de séduction, l'animal utilisé pour la transformation, la femme séduite, le titre de la peinture représentant la séduction et le peintre du tableau. Vous n'êtes pas obligé d'avoir le même peintre pour les trois réponses, mais vous devez bien avoir trois animaux différents et trois femmes différentes en tout"

- **Tâche 1 (TMR) - Variante 2 :**

- "La Renaissance italienne s'est beaucoup intéressée à la mythologie grecque dans ses peintures, notamment aux transformations animales que pouvaient prendre certains personnages. Le père d'Héraclès s'est transformé trois fois en animal afin de séduire des femmes mortelles (humaines). Pour compléter cette tâche vous devez remplir le tableau suivant en précisant pour les trois cas de séduction, l'animal utilisé pour la transformation, la femme séduite, le titre de la peinture représentant la séduction et le peintre du tableau. Vous n'êtes pas obligé d'avoir le même peintre pour les trois réponses, mais vous devez bien avoir trois animaux différents et trois femmes différentes en tout »

- **Tâche 2 (TSF) - Variante 1 :**

- *"Quel est le lien entre le long métrage d'animation réalisé par Pete Docter où Henry Guybet prête sa voix et la version 1.2 du système d'exploitation représenté par une spirale, associé à celui avec le pingouin ?"*

- **Tâche 2 (TSF) - Variante 2 :**

- *Quel est le lien entre la version 1.2 du système d'exploitation représenté par une spirale, associé à celui avec le pingouin et le long métrage d'animation réalisé par Pete Docter où Henri Guybet prête sa voix ?"*

La variété des tâches proposées, en termes de taille et de complexité, permettra d'étudier l'influence qu'exercent ces facteurs sur les productions des utilisateurs. De plus, bien que ce mémoire s'intéresse aux variations affectant les requêtes initiales, ce jeu de données nous permet aussi d'avoir accès à l'entièreté de la session et donc d'avoir un point de vue plus global sur le profil de l'utilisateur. Le Tableau 5 correspond à un exemple de session de recherche d'un utilisateur interagissant avec la variante 1 de la tâche 2. La première colonne correspond à l'identifiant de l'utilisateur. À noter que l'acronyme TWD correspond à la tâche d'entraînement, l'acronyme TMR fait référence à la tâche 1, enfin une session correspondant à la tâche 2 sera indiquée par les lettres TSF. La seconde colonne indique le contexte de la requête au sein de la session de recherche et la dernière correspond à la production de l'utilisateur. On s'aperçoit alors qu'un même utilisateur peut produire des requêtes très différentes au cours de sa session et que, plus globalement, les comportements utilisateurs sont marqués par une forte variabilité.

Identifiant	Contexte	Requête
NV6358_TSF	Requête initiale	*long métrage d'animation réalisé par pete docter*
NV6358_TSF	Reformulation 1	*long métrage d'animation réalisé par pete docter henry guybet*
NV6358_TSF	Reformulation 2	*système d'exploitation représenté par une spirale*
NV6358_TSF	Reformulation 3	*système d'exploitation liste*
NV6358_TSF	Reformulation 4	*système d'exploitation associé à linux*
NV6358_TSF	Reformulation 5	*debian*
NV6358_TSF	Reformulation 6	*debian version 1.2*

Tableau 5 - Exemple d'une session de recherche correspondant à une production découlant de la Tâche 2 (Variante 1)

2.2. Description quantitative et qualitative des données

Cette section vise à offrir une première description en surface des données composant le corpus PRIVaThe. En premier lieu, nous nous intéresserons aux sessions de recherche contenues dans ce dernier en observant les productions complètes des participants. En second lieu, nous nous attarderons sur les 300 requêtes initiales observables, qui se trouveront au cœur de notre analyse.

2.2.1. Du point de vue des sessions de recherche

On remarque que les deux tâches principales amènent les participants à se lancer dans des sessions de taille variable, en témoignent les Figures 3 et 4. Ces histogrammes, tout deux tirés de Ibarboure (2025), représentent le nombre de requêtes contenues dans les différentes sessions de recherche effectuées au cours des tâches de recherche TMR et TSF. Pour la tâche TMR, les sessions comptent entre 5 et 35 requêtes, avec une distribution relativement homogène centrée autour de 5 à 25 requêtes. On ne remarque aucune session de moins de 5 requêtes. Quant à la tâche TSF, la distribution est plus étalée, allant de 0 à 50 requêtes, avec un pic marqué autour de 10-15 requêtes. On observe ici des sessions de moins de 5 requêtes. Pour rappel, une session s'achève lorsque le participant a récolté toutes les informations qui lui sont nécessaires à la résolution de la tâche de recherche, ou bien qu'il est à court de temps. Le faible nombre de requêtes observables au sein de la tâche d'entraînement, qui pour rappel n'en contient que 203, nous indique que les sessions de recherche sont significativement plus courtes. Cela pour cause de la simplicité de la tâche qui ne contient notamment qu'une seule et unique thématique.

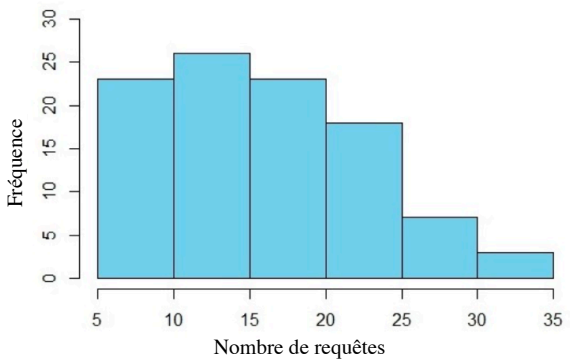

Figure 3 - Histogramme de longueur des sessions en nombre de requêtes pour la tâche TMR (Ibarboure, 2025)

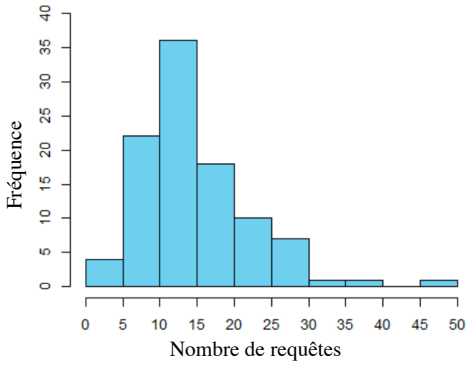

Figure 4 - Histogramme de longueur des sessions en nombre de requêtes pour la tâche TSF (Ibarboure, 2025)

Concernant la taille des requêtes, on observe une moyenne de 4,3 tokens pour la tâche TMR contre 4,8 tokens pour la tâche TSF. On observe qu'ici, la tâche d'entraînement a tendance à produire des requêtes plus longues avec une moyenne de 5,4 tokens par requête, malgré le fait qu'elle comporte des sessions de recherche plus courtes. On remarque également que ce facteur de longueur est cohérent d'une tâche à l'autre. Un utilisateur formulant des requêtes longues pour l'une aura tendance à faire de même pour l'autre (Ibarboure, 2025).

D'un point de vue qualitatif, les requêtes sont porteuses d'un grand nombre de variations. En accord avec la typologie de Hearst (2009) présentée précédemment, certaines requêtes sont formulées en langage naturel, que ce soit sous forme interrogative **(11)** ou déclarative **(12)**. D'autres le sont sous forme de mots-clés comme c'est le cas avec l'exemple **(13)**. On remarque que le nombre de requêtes composées par mot-clés semble être important. Enfin, certaines contiennent également des opérateurs **(14)**, qui visent à interagir plus efficacement avec le moteur de recherche. On remarque que les participants tendent très souvent à utiliser des noms propres afin de composer leurs requêtes. D'ailleurs, les opérateurs utilisés dans la requête correspondant à l'exemple **(14)** visent précisément à chercher une correspondance exacte avec deux noms propres présents dans l'énoncé de la tâche TSF.

(11) *en quel animal s'est transformé zeus ?*
(12) *systeme d'exploitation représenté par une spirale*
(13) *walking dead personnage*
(14) \"*pete docter*\" + \"*henri guybet*\"

De plus, on observe que, en lien avec la typologie de Broder (2002), l'immense majorité des requêtes sont à caractères informationnelles, visant ainsi à obtenir une information directement par le biais du processus de matching. Ce caractère informationnel peut tout aussi bien transparaître par le biais de requêtes en langage naturel **(15)** que par le moyen de mots-clés **(16)**. On observe également certaines requêtes à caractère navigationnelles, toutes cherchant à accéder à des encyclopédies collaboratives en ligne **(17)**. Nous pouvons également observer que bien que la majorité des requêtes soit formulée en français, certains participants ont utilisé la langue anglaise **(18)**. Cela par le biais de la traduction de termes présents dans l'énoncé ou bien par l'ajout de mots-outils anglophones au sein des requêtes.

(15) *femmes séduites par zeus en se transforman en animal*
(16) *zeus trasformation séduction*
(17) *pixar wiki fandom*
(18) *transformation of heracles in paintings*

2.2.2. Du point de vue des requêtes initiales

Après avoir décrit les sessions de recherche dans leur ensemble, nous nous intéressons à présent plus spécifiquement aux requêtes initiales. Ces dernières constituent en effet le point d'entrée de chaque session et représentent en cette qualité l'objet central de ce projet de recherche. Comme mentionné précédemment, elles ne représentent qu'une proportion minoritaire des requêtes du corpus, avec 300 requêtes initiales sur l'ensemble des sessions. Cependant, elles témoignent d'une variété de formulations comparable à celle observée dans l'ensemble du corpus.

Comme cela a été fait pour l'ensemble des requêtes du corpus, nous nous intéressons ici à la longueur des requêtes initiales, mesurée en nombre de tokens, afin de la comparer pour chacune des trois tâches. Pour la tâche d'entraînement, la longueur moyenne des requêtes initiales est de 5,9 tokens. Cette valeur est légèrement supérieure à la longueur moyenne observée pour l'ensemble des requêtes de la session, qui s'élève à 5,43 tokens. Pour la tâche TMR, la longueur moyenne des requêtes initiales est de 3,7 tokens. Contrairement aux deux autres tâches, cette valeur est inférieure à la moyenne de l'ensemble des requêtes, qui est de 4,33 tokens. Enfin, pour la tâche TSF, la longueur moyenne des requêtes initiales atteint 6,31 tokens, soit une valeur nettement supérieure à la moyenne de 4,8 tokens observée pour l'ensemble des requêtes de cette tâche. On remarque alors que la longueur moyenne de la requête initiale diffère systématiquement de la longueur moyenne de l'ensemble des requêtes, et ce pour chacune des trois tâches.

À l'image de ce qui a été observé pour l'ensemble des requêtes, les requêtes initiales présentent une grande richesse de variations. On retrouve ainsi des requêtes formulées en mots-clés ou bien en langage naturel. Rien ne semble, sur la forme, distinguer ces requêtes initiales du reste du jeu de données. Les mêmes observations concernant la présence de l'anglais ou la prédominance de requêtes informationnelles peuvent être effectuées, confirmant ainsi que les requêtes initiales présentent une diversité de variations comparable à celle de l'ensemble du corpus. Cette diversité reflète alors des stratégies de recherche très différentes d'un utilisateur à l'autre, comme en témoigne le Tableau 6. Ce dernier contient quelques requêtes initiales formulées dans le cadre de la tâche TMR par différents participants. Ces exemples illustrent que dès la requête initiale, avant même toute interaction avec le système, des traitements radicalement variables de l'énoncé sont susceptibles d'être produits. On remarque également que certaines observations réalisées précédemment sont ici exacerbées. C'est notamment le cas de l'importante présence de noms propres qui sont d'ailleurs systématiques dans les exemples ci-dessous.

	Identifiant	Requête
(19)	**NV3996_TMR**	*tableau représentant le père d'Héraclès séduisant des femmes*
(20)	**NV4650_TMR**	*père d'héracles*
(21)	**NV5860_TMR**	*transformation of heracles in paintings*
(22)	**NV6187_TMR**	*père d'héracles 3 animaux*

Tableau 6 - Exemple de requêtes initiales découlant de la Tâche TMR (Variante 2)

Le Tableau 22 disponible en *Annexe 1* représente la fréquence d'apparition des lemmes employés par les participants afin de répondre à la tâche TWD. On s'aperçoit que les quatre lemmes les plus employés sont des noms propres, tous découlant de l'énoncé. Chacun d'entre eux a plus de 80 occurrences sur les 100 requêtes étudiées. De telles tendances sont également observables au sein des Tableau 21 et 23, qui correspondent respectivement aux tâches TMR et TSF. On remarque que la tâche d'entraînement entraîne la production de requêtes moins variées d'un point de vue lexical. Ceci semble être dû au fait que l'énoncé de la tâche contient bien moins de tokens que les deux autres. En effet, on s'aperçoit que la majorité des tokens présents dans les requêtes découlent de lemmes déjà observables dans l'énoncé. Cela semble indiquer que les participants se reposent particulièrement sur l'énoncé au début de la session. Enfin, bien qu'une vérification statistique précise soit irréalisable sur un échantillon aussi peu dense, la distribution des fréquences présente une tendance proche de celle décrite par la loi de Zipf. Une poignée de lemmes très fréquents concentrent systématiquement une part importante des occurrences, tandis que la plupart des autres lemmes apparaissent de manière beaucoup plus sporadique.

2.3. Présentation de la démarche suivie

Il apparaît rapidement qu'une grande variété d'opérations peuvent être réalisées lors du premier traitement du besoin d'information lors d'une session de recherche sur le Web. Comme le démontre le Tableau 6, exposant différentes requêtes initiales découlant de la variante 2 de la tâche 1, des stratégies radicalement différentes sont mises en œuvre selon les utilisateurs. Ce projet de recherche va ainsi s'articuler autour de deux problématiques distinctes.

Dans un premier temps, notre objectif sera d'observer les stratégies mises en place par les participants lors d'une première interaction avec un énoncé en langage naturel, représentant lui-même un besoin d'information donné. Cette première interaction donnera ainsi lieu à une requête initiale formulée sur un moteur de recherche. Le corpus PRIVaThe nous offre alors la possibilité d'observer un total de 300 requêtes initiales découlant de trois tâches de recherche distinctes. Cette approche descriptive des données sera rendue possible par l'automatisation d'un système de traitement prenant la forme d'un script en langage Python. Ce script, appliqué à l'ensemble du jeu de données, permettra alors de récolter, pour chaque requête, un certain nombre d'informations la concernant. Ce traitement rendra possible une analyse bidimentionelle. En premier lieu, nous nous pencherons sur les caractéristiques structurelles des requêtes initiales. C'est-à-dire leur taille, ou bien encore la manière dont elles sont formulées, en référence à la distinction entre composition en mots-clés et formulation en langage naturel pensée par Hearst (2009). Nous nous pencherons ici également sur les potentielles opérations de permutations effectuées par le participant lors de ses interactions avec l'énoncé. C'est-à-dire des changements affectant l'ordre canonique des tokens de l'énoncé tel qu'ils sont ordonnés dans celui-ci. En second lieu, le traitement automatique visera aussi à rendre compte de l'agencement lexical propre à chacune des requêtes. Pour cela, chaque token de chaque requête sera étiqueté selon son statut lexical, en lien avec l'énoncé de la tâche de recherche correspondant. Nous mettrons alors au point une taxinomie des différentes stratégies lexicales que le participant peut opérer. L'ensemble des résultats de ce traitement sera disponible au format JSON afin de permettre leur analyse. Le format de données choisi est décrit plus en détail en *Annexe 2*.

Dans un second temps, nous répondrons à une seconde problématique de recherche. Celle-ci permet de mettre à profit le traitement global effectué sur l'ensemble des données en essayant d'observer les différences entre requêtes initiales et reformulations présentes au sein du jeu de données PRIVaThe. Comme nous l'avons vu, la requête initiale se distingue, par son statut initiateur, des autres requêtes qui la suivent. Les informations structurelles et lexicales contenues dans les sorties obtenues grâce aux traitements nous permettront alors d'observer si des différences concrètes distinguent requêtes initiales et reformulations dans un contexte expérimental. L'ensemble des résultats obtenus rendront ainsi possible une observation globale des comportements utilisateurs caractérisant une première interaction avec un moteur de recherche dans le cadre de la résolution d'un besoin d'information prenant la forme d'un énoncé en langage naturel.

3. Procédure de traitement des requêtes

La présente section vise à expliciter les différentes opérations de traitement opérées sur les données du corpus PRIVaThe. Ces traitements visent à rendre possible l'analyse de ces données selon deux axes complémentaires. Dans un premier temps, les traitements effectués à l'échelle des unités lexicales seront décrits. Puis, dans un second temps, les traitements portant sur les propriétés structurelles des requêtes seront présentés. Enfin, nous proposerons une évaluation de ces différents moyens de traitement afin d'en mesurer la fiabilité.

3.1. Traitement lexical des requêtes

Cette section traite du traitement lexical soumis à chacune des requêtes. En effet, chaque requête est composée d'un certain nombre d'éléments lexicaux qui la constituent. Ces éléments témoignent de plusieurs types d'opérations que réalise le participant lors de la formulation en requêtes. Chacun des tokens composant ces dernières a alors été soumis à un traitement qui vise à décomposer les requêtes et donc les choix opérés par l'utilisateur lors de l'interaction avec l'énoncé de la tâche.

3.1.1. Taxinomie des stratégies lexicales

La Figure 5 correspond à la taxinomie de traitement qui a été appliquée à chaque requête du corpus PRIVaThe. Ainsi, chaque token de la requête est soit présent ou absent de l'énoncé. Si il est déjà présent dans celui-ci, alors le token correspond à une stratégie de recopie. Si il est absent de ce dernier, alors il correspond soit à une stratégie de modification d'un token déjà existant, ou alors à une stratégie d'ajout, c'est-à-dire à un élément qui n'est lié à aucun token de l'énoncé. Ce cycle se répète pour chaque token présent dans la requête, jusqu'à ce qu'il n'en reste plus.

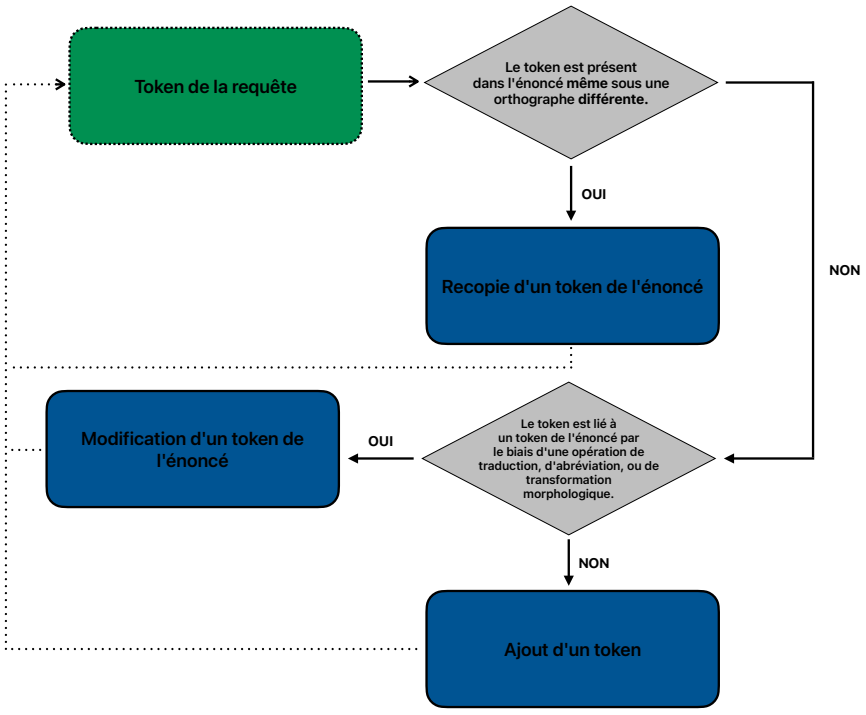

Figure 5 - Taxinomie de traitement des tokens de la requête d'un point de vue lexical

Ainsi, on distingue trois types d'opérations que le participant peut effectuer : la recopie d'un token de l'énoncé, la modification d'un token présent dans l'énoncé, ainsi que l'ajout d'un nouvel élément. Le système de traitement doit donc pouvoir être en mesure, sans aucune annotation réalisée en amont, d'attribuer à chacun des tokens de la requête une de ces catégories. On distingue en conséquence trois étapes distinctes de traitement.

3.1.2. Extraction des stratégies de recopie

En premier lieu, un utilisateur peut recopier un élément de l'énoncé afin de l'incorporer sans sa requête. Ces stratégies de recopie concernent donc des tokens qui apparaissent aussi bien dans l'énoncé de la tâche qui a été donnée au participant, que dans la requête de ce dernier. Il est important de spécifier que nous ne considérons pas les variations typographiques comme des stratégies de modification, à l'inverse de Abu Onq et al. (2025). De ce fait, deux tokens qui ne sont pas orthographiés de la même manière mais qui correspondent à la même unité lexicale seront considérés comme des recopies. Cela signifie que le système de traitement doit prendre en compte les potentiels variations orthographiques qu'il pourrait rencontrer.

Ce choix a été opéré pour plusieurs raisons. Premièrement, on observe une quantité considérable de fautes d'orthographe dans le jeu de données PRIVaThe, bien plus que dans les données récoltées par Abu Onq et al. (2025). Ainsi, la prise en compte de ces variations typographiques comme des stratégies manifestes de modification aurait dilué les modifications dans une masse quasi exclusive de fautes d'orthographe. Deuxièmement, ces variations ne sont en rien des manifestations d'une stratégie de la part du participant. Elles ne témoignent que d'une certaine hâte de finir la tâche à temps ou bien d'un geste involontaire. Le Tableau 7 rassemble quelques exemples de requêtes comportant des variations typographiques.

	Identifiant	Requête
(23)	**NV3343_TWD**	*WALKING DEAD*
(24)	**NV4005_TWD**	*hershel greene wlking dead*
(25)	**NV4509_TSF**	*logn métrage de pete docter*
(26)	**NV9135_TWD**	*deuxième fille e HershelGreene*

Tableau 7 - Exemples de requêtes comportant des variations typographiques

Comme on peut le voir, ces variations typographiques sont de différentes natures. Dans l'exemple **(23)**, le participant a composé sa requête à partir de deux tokens présents dans l'énoncé. Seulement ces deux-mêmes éléments sont ici réorthographiés intégralement en majuscules. Dans les exemples **(24)** et **(25)**, les majuscules propres à chacun des noms propres ont ici été omises. De plus, ces deux requêtes comportent respectivement une suppression ("*wlking*") puis une inversion de caractère ("*logn*"). Enfin, l'exemple **(26)** contient également une omission de caractère mais aussi une concaténation de deux tokens distincts. Pour résumer, les variations typographiques sont omniprésentes et prennent des formes très variables. Les prendre en compte au cours du traitement des données devient alors indispensable, en particulier dans le cadre de l'extraction des stratégies de recopie.

Le traitement s'effectue alors token par token, dans l'ordre d'apparition dans la requête. Pour chaque token, on cherche à identifier s'il correspond à un token de l'énoncé, et lequel. L'énoncé de la tâche qui va être comparé à chacune des requêtes est contenu au sein d'une variable qui contient la chaîne de caractères correspondante. À noter que dans les cas où la tâche possède deux variantes, deux variables différentes sont initialisées. Ainsi le traitement s'effectue en plusieurs étapes successives. En premier lieu, chaque token de la requête mais aussi de l'énoncé est normalisé. Tous les caractères sont alors mis en majuscules, les apostrophes sont uniformisées et l'intégralité des diacritiques sont supprimés. Le système est alors en capacité d'établir des correspondances entre deux tokens purement similaires mais également de traiter des éléments tels que ceux présents dans l'exemple **(23)**. Pour se faire, le token normalisé de la requête est comparé successivement à chaque token normalisé dans la liste de l'énoncé. Si une correspondance stricte est trouvée, le token est étiqueté en tant que recopie. De plus, le token de l'énoncé avec lequel il correspond est enregistré comme correspondance.

En second lieu, si aucune correspondance exacte pour un token de la requête n'a été trouvée, une mesure de similarité est calculée entre le token de la requête et chaque token de l'énoncé qui n'a pas encore trouvé de correspondance. Cette mesure, automatisée grâce à la bibliothèque Python *difflib*, correspond à une distance d'édition relative au nombre de caractères contenus dans le token. Si la similarité atteint ou dépasse le seuil fixé de 0,75, alors le token de l'énoncé sera considéré comme une recopie, tout en enregistrant en correspondance le token de l'énoncé associé. Pour chaque token recopié, le système fournit alors comme information : la chaîne de caractères associée au token, la position de ce dernier au sein de la requête, l'indication que le token correspond à une recopie, ainsi que la correspondance trouvée. Un exemple de sortie d'un token recopié est montré ci-dessous. Pour rappel, une description de la structure JSON est disponible en *Annexe 2*.


```
{
  "token": "wlking",
  "position": 3,
  "statut": "recopie",
  "correspondance": "Walking"
},
```


Extrait d'une sortie obtenue lors du traitement du token "wlking" tiré de d'une requête découlant de la tâche TWD

En plus du détail pour chaque token, le système inscrit aussi dans le JSON, pour chaque requête, le nombre total de recopies ainsi que le taux de recopie. Le premier est obtenu en dénombrant le nombre total de tokens correspondant à une opération de recopie contenus dans la requête. Par exemple, la requête correspondant à l'exemple **(24)** aura un nombre de recopies de 4. Le taux de recopies s'obtient quant à lui en divisant le nombre total de recopies par la taille de la requête. Par exemple, si une requête contient 5 tokens et que 3 d'entre eux sont des recopies, alors le taux de recopie sera égal à 0,60. Cette valeur est alors systématiquement comprise entre 0 et 1 et correspond à la proportion de la requête qui équivaut à une recopie. Si la requête ne contient aucune recopie alors cette valeur sera de 0.

3.1.3. Extraction des stratégies de modification

En second lieu, l'utilisateur peut choisir d'interagir avec l'énoncé différemment qu'en recopiant simplement un ou plusieurs segments de celui-ci. Il a également la possibilité de modifier un de ces derniers à l'aide de plusieurs types d'opérations. L'utilisateur s'appuie donc encore sur l'énoncé mais opère une modification sur un de ces éléments. Similairement à la typologie conçue par Abu Onq et al. (2025), nous avons distingué plusieurs types de modifications, en nous éloignant néanmoins de celle-ci.

- **Abréviation**

Premièrement, l'utilisateur peut choisir de réaliser une abréviation d'un des tokens de l'énoncé. Cette catégorie a été directement reprise depuis la typologie proposée par Abu Onq et al. (2025). La plupart du temps, l'utilisateur modifie un adjectif, numéral, ordinal en une abréviation associée. L'utilisateur peut également convertir un adjectif en chiffre. Ces opérations ne témoignent pas véritablement d'une intention d'optimiser le processus de matching mais plutôt de gagner du temps en choisissant une option moins chronophage. À noter que tous les énoncés ne contiennent pas nécessairement de tokens pouvant être affectés par une modification de ce type.

- **Transformation morphologique**

Deuxièmement, l'utilisateur peut effectuer une opération morphologique sur un token de l'énoncé. Il peut alors effectuer une flexion morphologique sur un élément de l'énoncé en modifiant son genre ou son nombre. Il peut également transformer un token de l'énoncé par dérivation morphologique en faisant varier sa catégorie grammaticale. Ce type de modification est également cité au sein des stratégies de modification pensées par Abu Onq et al. (2025)

- **Traduction**

Enfin, l'utilisateur peut également traduire un token de l'énoncé dans une autre langue. Les tâches de recherche étant formulées en français, on aperçoit quelques cas de traductions vers l'anglais. Il est important de préciser que seuls les mots pleins sont ici pris en compte. Sont considérés comme mots pleins tous les noms propres et communs, les verbes, les adjectifs, ainsi que les adverbes. En effet, il est impossible de relier avec certitude un mot outil de la requête à un token de l'énoncé. Afin d'éviter de faire des inférences et de potentiellement confondre modification et ajout, ces cas n'ont pas été considérés comme des traductions.

Le Tableau 8 correspond à un ensemble de stratégies de modifications observées dans le jeu de données. On peut alors observer des cas d'abréviations **(27)(28)** qui ont été effectuées sur des adjectifs présents dans les énoncés des tâches TWD et TMR. L'exemple **(29)** représente quant à lui un cas de flexion morphologique suite à un changement de nombre du token "système". L'opération de modification suivante **(30)** fait apparaître un cas de dérivation morphologique. Ici, le nom "animal" a été converti en l'adjectif "*animal*" afin d'être utilisé dans un autre contexte tout en conservant son sens initial. Enfin, l'exemple **(31)** représente une opération de traduction ayant affecté le nom "fille". L'enjeu est donc pour le système de traitement de savoir différencier ces différentes opérations tout en les distinguant elles-mêmes de stratégies d'ajouts.

	Modification effectuée	Type de modification
(27)	deuxième → "*2e*"	Abréviation
(28)	trois → "*3*"	Abréviation
(29)	système → "systèmes"	Transformation morphologique
(30)	animal → "*animale*"	Transformation morphologique
(31)	fille → "*daughter*"	Traduction

Tableau 8 - Exemples de stratégies de modifications observées dans le corpus PRIVaThe

Chronologiquement, l'identification des stratégies de modification s'effectue après que la recherche de recopie exacte s'est révélée infructueuse, mais avant que la recherche de recopie approximative ne soit calculée. En effet, si la correspondance approximative était vérifiée avant les modifications, certains tokens qui sont en réalité des transformations morphologiques ou des abréviations pourraient être capturés par le seuil de similarité et étiquetés à tort comme recopies.

Les stratégies de modifications regroupant trois sous-types sont détectées à l'aide de trois fonctions distinctes apparaissant dans l'ordre suivant :

1. La bibliothèque python *num2words* a été utilisée afin de détecter les cas d'**abréviations**. Lorsqu'un token de la requête est identifié comme un chiffre ou un ordinal abrégé, son suffixe ordinal est retiré et le chiffre restant est converti en deux formes : sa forme cardinale (par exemple "*deux*") et sa forme ordinale (par exemple "*deuxième*"). Ces deux formes sont ensuite comparées, après normalisation, à chaque token de l'énoncé. En cas de correspondance, le token de la requête est étiqueté en tant qu'abréviation et l'élément correspondant de l'énoncé est enregistré comme correspondance.

2. Afin de détecter et d'extraire les **transformations morphologiques**, l'analyseur morpho-syntaxique *spaCy* a été importé dans le script Python. Pour chaque token de la requête identifié comme mot plein, son lemme est extrait et comparé au lemme de chaque token de l'énoncé. En cas de correspondance, le token de la requête est étiqueté et la correspondance est enregistrée. Si la correspondance concerne un lemme présent sous plusieurs formes distinctes dans l'énoncé, une similarité sémantique est calculée entre les différents tokens afin d'éviter que le système n'établisse automatiquement une correspondance avec la première occurrence. Lorsque plusieurs tokens de l'énoncé partagent le même lemme, celui présentant la similarité sémantique la plus élevée est retenu comme correspondance.

3. La fonction *GoogleTranslator* importée de la bibliothèque *deep_translator* est employée afin d'extraire les opérations de **traduction**. Pour chaque token de la requête identifié comme mot plein, une vérification préalable s'assure qu'il n'est pas déjà très proche graphiquement d'un token de l'énoncé grâce à un seuil de similarité. Cela permet de ne pas capter de potentiels faux cas de recopies. Si cette vérification est passée, le token est traduit de l'anglais vers le français. La traduction obtenue est ensuite comparée, après normalisation, à chaque token de l'énoncé. Si une correspondance est trouvée, comme pour les autres modifications, le token est étiqueté en conséquence. L'exemple ci-dessous représente le cas du traitement d'un token traduit. On s'aperçoit que le seuil de similarité présent dans la normalisation a permis au système de passer outre la variation typographique observable. Le système renvoie alors les mêmes informations que pour les recopies, à l'exception qu'il spécifie ici le type de modification qu'il a détecté.


```
{
  "token": "dughter",
  "position": 5,
  "statut": "modification : traduction",
  "correspondance": "fille"
}
```


Extrait d'une sortie obtenue lors du traitement du token "dughter" tiré de d'une requête découlant de la tâche TWD

En plus du détail de chaque token, le système renvoie également le nombre total de modifications pouvant être observées dans la requête ainsi que le taux de modification. La logique du calcul du taux de modification est similaire à celle du taux de recopie. Si une requête de 4 tokens comporte 1 modification, peu importe laquelle, alors le taux de modification sera égal à 0,25.

3.1.4. Extraction des stratégies d'ajout

Enfin, l'utilisateur peut choisir de complètement se détacher de l'énoncé de la tâche en ajoutant des éléments qui n'y sont pas présents. Nommées "Stratégies d'enrichissement" par Abu Onq et al. (2025), elles prennent donc des formes fortement imprévisibles. Cependant, contrairement à ces derniers, nous ne constituons pas une typologie d'ajouts. En effet, on remarque rapidement que ces derniers sont relativement rares à l'échelle du corpus. Également, nous avons considéré comme ajouts certaines opérations que Abu Onq et al. (2025) prenaient en compte comme des modifications. En effet, les opérations d'antonymie ou autres modifications sémantiques ont été considérées comme des ajouts. Par exemple, le token "*famille*" ne sera pas considéré comme une modification de "fille" mais comme un ajout à part entière. De même pour les opérateurs tels que "+" qui seront systématiquement traités comme des ajouts. Le Tableau 9 regroupe plusieurs requêtes qui contiennent des opérations d'ajouts, mises ici en surbrillance en rouge.

	Identifiant	Requête
(32)	NV1162_TSF	\"*pete docter*\" ***+*** \"*henri guybet*\"
(33)	NV3576_TWD	*hershel greene* ***daughters***
(34)	NV2024_TMR	***comment zeus*** *séduit* ***hera ?***
(35)	NV1434_TSF	***debian linux APT + monstre & co***

Tableau 9 - Exemples de requêtes comportant des stratégies d'ajouts

Comme ces différents exemples le démontrent, ces ajouts prennent des formes extrêmement variées. Par exemple, la requête **(32)** contient plusieurs opérateurs que le participant adjoint à des tokens provenant de l'énoncé. L'exemple **(33)** comporte quant à lui un ajout assez particulier. En effet, le token "*daughters*" résulte de deux opérations de modifications distinctes. À savoir une traduction et une transformation morphologique du mot "fille". Il a été choisi que ces éléments soient considérés comme des ajouts, ceci pour faciliter le traitement lexical mais également car l'élément en question se distance trop de l'énoncé. L'exemple **(34)** contient également plusieurs ajouts. On remarque que l'utilisateur a ici inséré des éléments qu'il avait rencontrés à force d'interagir avec le système tels que "zeus" ou "hera" qui sont tous deux des référents de la tâche TMR. En effet, il s'agit de la troisième requête que l'utilisateur a formulée pour cette tâche. On observe également ce genre d'ajouts dans certaines requêtes initiales. Le participant possède déjà des connaissances sur le domaine de recherche et est alors en mesure d'alimenter ses requêtes à l'aide d'informations inférées. On voit que dans cette même requête, le participant a également ajouté le pronom interrogatif "comment" ainsi que le signe de ponctuation "?". Enfin, la requête **(35)** est entièrement composée d'ajouts de différentes natures.

Afin d'extraire tous ses éléments qui se distinguent tous les uns des autres, la condition qui repère les ajouts se trouve tout à la fin du traitement. Comme cela est explicité dans la *Figure 5*, si un token n'est ni une recopie, ni une modification, alors il s'agit obligatoirement d'un ajout. Ainsi, si un token ne trouve aucune correspondance lors des différentes étapes de traitement décrites précédemment, alors il sera automatiquement étiqueté comme ajout.

À noter qu'un token ajouté n'aura par conséquence aucune correspondance avec un élément de l'énoncé. Il renverra ainsi une valeur nulle à la balise "correspondance" comme cela est observable en *Annexe 2*. Le système renvoie néanmoins, pour chaque requête, le nombre total d'ajouts mais également le taux d'ajouts. Ce dernier est obtenu de la même manière que les précédents. À savoir en divisant le nombre d'ajouts par le nombre total de tokens contenus dans la requête.

3.2. Traitement structurel des requêtes

Cette section vise à présenter le traitement structurel soumis à chaque requête. Contrairement au traitement lexical, qui s'intéresse aux tokens individuellement, le traitement structurel appréhende la requête dans sa globalité. Cela en caractérisant sa taille et sa structure, l'ordre des éléments qu'elle contient, ainsi que la distance qui la sépare de l'énoncé.

3.2.1. Extraction de la taille et du type de requête

On cherche maintenant, pour chaque requête, à déterminer deux informations supplémentaires et interdépendantes. La première est la taille de la requête qui sera mesurée en tokens. La seconde est la catégorie auquel appartient la requête. En accord avec la typologie de Hearst (2009), on cherche à déterminer si la requête est formulée en langage naturel ou bien à l'aide d'une concaténation de mots-clés.

Afin d'extraire la taille des requêtes, le système procède simplement à une addition du nombre de recopies, modifications et ajouts présents dans la requête. Cette valeur est donc égale au nombre de tokens dénombrables au sein de la production du participant. On prend donc naturellement en compte tous les opérateurs ou signes de ponctuation. Cette valeur a déjà été mentionnée précédemment car c'est cette dernière qui permet de calculer les taux de proportion des différentes stratégies lexicales. Dans chaque fichier JSON, la balise "Nombre de tokens" indique donc la valeur correspondante.

Déterminer si une requête est formulée en mots-clés ou en langage naturel est légèrement plus complexe et nécessite un traitement à part entière. D'abord, il est important de délimiter ces deux éléments. Le Tableau 10 permet d'illustrer cette distinction. Lorsque le participant formule sa requête en mots-clés, il structure cette dernière en sélectionnant un ensemble de termes qui lui paraissent pertinents, tel que c'est le cas dans les exemples **(36)** et **(37)**. Au contraire, lorsqu'il la formule en langage naturel, il structure clairement cette dernière comme un énoncé, avec une syntaxe plus proche de la langue naturelle que d'une simple juxtaposition de termes. C'est notamment le cas au sein des exemples **(38)** et **(39)**.

	Identifiant	Requête	Type de requête
(36)	NV1721_TMR	*zeus wiki*	Mots-clés
(37)	V9291_TMR	*père d'heraclès*	Mots-clés
(38)	NV6369_TWD	*beth greene qui c'est*	Langage naturel
(39)	NV3996_TWD	*Quel est le nom de la deuxième fille de Hershel Greene dans Walking Dead*	Langage naturel

Tableau 10 - Exemples de requêtes illustrant la frontière entre construction en mots-clés et formulation en langage naturel

Dans le cadre du traitement, le système distingue ces deux moyens d'expressions en dénombrant le nombre de mots-outils présents dans la requête. Les mots-outils correspondent à des déterminants, des prépositions, des conjonctions ou des pronoms. Si la requête contient deux ou plus de ces mots-outils, alors cette dernière sera considérée comme formulée en langage naturel. Autrement, la requête sera étiquetée comme composée en mots-clés. La présence de ces mots-outils est vérifié par l'étiqueteur morpho-syntaxique spaCy, qui ne se concentre uniquement sur les éléments soumis en français. Cette frontière de deux mots-outils a été fixée arbitrairement en observant les données. Il est difficile de considérer la requête correspondant à l'exemple **(37)** comme formulée en langage naturel, bien qu'elle comporte la préposition "*d'*". Au contraire, la requête de l'exemple **(38)** contient deux mots-outils, à savoir les pronoms "*qui*" et "*c'*". Cette dernière est largement plus en accord avec la description de ce type d'expression donnée par Hearst (2009).

On peut ainsi apercevoir que, comme cela est mentionné dans l'étude initiale, cette distinction prend concrètement la forme d'un continuum. En effet, les requêtes **(38)** et **(39)** diffèrent très largement, et ce malgré leur catégorisation similaire. La première résulte de l'ajout du groupe de tokens "*qui c'est*" afin d'interroger directement le moteur de recherche. Dans la seconde, le participant a tout simplement recopié une grande majorité de l'énoncé, lui-même formulé en langage naturel. Le processus de traitement tel que décris précédemment n'a donc pas vocation à représenter la complexité causée par ce continuum. Il permet néanmoins d'établir une distinction cohérente et reproductible afin de différencier ces deux comportements, cela dans l'objectif de faciliter l'analyse des résultats obtenus.

3.2.2. Stratégies de suppression

Lorsqu'il formule sa requête, l'utilisateur interagit avec l'énoncé en puisant dans celui-ci, à des degrés divers. Comme nous avons pu le voir, ce dernier recopie et modifie à sa guise des éléments de la tâche afin de répondre à cette dernière. Si on considère l'énoncé comme une série d'éléments, en l'occurrence de tokens plus ou moins porteurs de sens, alors on s'aperçoit que souvent, une grande partie d'entre eux disparaît. L'utilisateur choisit ainsi de les supprimer car il les considère comme impertinents ou secondaires à ce moment de la session de recherche. Notre objectif sera ici, pour chaque requête, de dénombrer le nombre de tokens de l'énoncé n'apparaissant pas dans la requête, puis de convertir cette mesure en un pourcentage proportionnel à la taille de l'énoncé. En effet, obtenir l'information du nombre de tokens supprimés n'est pas réellement pertinent car, comme cela a été montré, la taille des énoncés varie significativement. Obtenir un pourcentage de suppression qui prend en compte la taille de l'énoncé apparaît alors comme plus exploitable.

Le Tableau 11 contient dans sa troisième colonne trois requêtes formulées dans le cadre de la tâche TWD. La quatrième colonne indique le pourcentage de suppression associé.

	Identifiant	**Requête**	**Pourcentage de suppression**
	Énoncé	*Quel est le nom de la deuxième fille de Hershel Greene dans la série Walking Dead ?*	0 %
(40)	NV9444_TWD	*quelle est le nom de la deuxiéme fille de Hershel Greene dans Walking Dead*	18 %
(41)	NV5267_TWD	*nom deuxième fille de hershel greene walking dead*	53 %
(42)	V9858_TWD	*wikipedia*	100 %

Tableau 11 - Exemples de requêtes découlant de la tâche TWD ainsi que le pourcentage de suppression associé

La requête correspondant à l'exemple **(40)** obtient un indice de 18 %. En effet, on observe qu'une proportion similaire de l'énoncé a disparu de cette dernière. Les tokens "la", "série" et "?" ont tous les trois été supprimés. On peut alors supposer que l'utilisateur part du principe que le système associera de lui-même la construction "Walking Dead" à la série du même nom, mais également que le pronom interrogatif "quel" suffit pour exprimer une interrogation. La requête **(41)** est quant à elle composée à l'aide de mots-clés. Ainsi, 9 tokens de l'énoncé ont ici disparu, soit près de 50% de la surface totale de l'énoncé. L'exemple **(42)** représente un cas de figure où la requête ne contient qu'un seul et unique token, en l'occurrence un ajout. Dans ce cas de figure, le taux de suppression sera égal à 100% car l'intégralité des constituants de l'énoncé ont été supprimés.

Afin de traiter automatiquement chaque requête, la liste complète des tokens de l'énoncé est d'abord constituée. Parallèlement, l'ensemble des correspondances enregistrées lors du traitement lexical est récupéré. C'est-à-dire tous les tokens de l'énoncé ayant servi de correspondance à un token de la requête, que ce soit par recopie ou par modification. Dans ce cadre, peu importe si l'élément présent dans l'énoncé correspond à une recopie ou à une modification, dès lors qu'une correspondance a été établie, le token de l'énoncé concerné ne sera pas considéré comme supprimé. Une copie de la liste des tokens de l'énoncé est ensuite créée. Pour chaque correspondance identifiée, le token de l'énoncé concerné est retiré un à un de cette copie. Ce retrait progressif garantit qu'un même token de l'énoncé ne peut être retiré qu'une seule fois. De plus, ce processus de retrait permet également de prendre en compte les cas où deux tokens identiques sont présents dans l'énoncé. Les tokens qui subsistent dans la copie après ce retrait sont ceux pour lesquels aucune correspondance n'a été trouvée dans la requête. Leur nombre constitue le nombre de suppressions. Ce nombre est ensuite rapporté au nombre total de tokens de l'énoncé et multiplié par cent pour obtenir un pourcentage, qui sera immédiatement arrondi à l'entier le plus proche.

3.2.3. Stratégies de permutation

La dernière étape du traitement concerne les stratégies de permutations. En effet, nous avons déjà observé qu'un utilisateur pouvait supprimer un certain nombre d'éléments originaires de l'énoncé, ce dernier étant une suite d'éléments ordonnés. Néanmoins, le participant peut également, lorsqu'il formule sa requête, modifier l'ordre d'apparition de ces derniers. Les stratégies de permutations correspondent ainsi aux cas où l'utilisateur a altéré la structure de l'énoncé en modifiant l'ordre dans lequel les tokens étaient énoncés.

Afin de détecter ainsi que de mesurer ses phénomènes, le Tau de Kendall a été employé. Concrètement, l'idée est de comparer si l'ordre des tokens communs à l'énoncé et à la requête apparaît dans le même ordre. Chacune des paires de mots de l'énoncé et de la requête est alors comparée afin de déterminer si l'ordre apparaît comme concordant ou discordant. Le résultat est alors égal à (C - D) / (C + D) où C correspond au nombre de paires concordantes et D au nombre de paires discordantes. Il est important de préciser que, dans le cadre de ce traitement, les mots outils n'ont pas été considérés. Seuls les noms propres et communs, les verbes et les adjectifs sont alors considérés. Cela pour faciliter le traitement, mais également pour observer si certains éléments porteurs de sens ont tendance à apparaître dans une position modifiée lors de la formulation en requête.

Dans une perspective de traitement, la liste ordonnée des mots pleins de l'énoncé est d'abord constituée. Pour chaque token de la requête ayant une correspondance, la position de cette correspondance dans cette liste est enregistrée. On obtient ainsi une séquence de positions reflétant l'ordre dans lequel les correspondances apparaissent dans la requête. Cette séquence est comparée à une séquence de référence, qui représente l'ordre original des mots pleins dans l'énoncé, via le Tau de Kendall. Le résultat brut du tau est ensuite transformé selon la formule suivante : (1 - tau) / 2, afin d'obtenir une valeur comprise entre 0 et 1, où 0 indique une conservation parfaite de l'ordre et 1 une inversion totale.

À noter que si le calcul ne peut pas être effectué, la valeur renvoyée est nulle (*Null*). C'est notamment le cas quand la requête ne contient pas assez d'éléments communs avec l'énoncé. Le Tableau 12 contient un ensemble de requêtes découlant de la tâche TWD. Pour chacune d'entre elles, le Tau de Kendall est indiqué à droite. Les termes de l'énoncé en question qui sont pris en compte dans le cadre du calcul de l'indice sont indiqués en surbrillance. La requête **(43)** affiche un Tau de Kendall égal à 0,2 pour 5 tokens pris en compte lors du traitement. En effet, on remarque qu'une version modifiée du token "fille" modifie l'ordre initial des composantes de l'énoncé. Dans l'exemple **(44)**, on retrouve une requête au sein de laquelle on peut observer une permutation des groupes de noms propres "Hershel Greene" et "walking Dead" entraînant alors un indice de 0,67. Enfin, l'exemple **(45)** contient une requête qui n'a pas pu être traitée. La requête ne comportant pas assez d'éléments recopiés ou modifiés, en l'occurrence seulement une seule recopie en la présence du token "*fille*".

	Identifiant	**Requête**	**Indice de Kendall**
	Énoncé	*Quel **est** le **nom** de la **deuxième fille** de **Hershel Greene** dans la **série Walking Dead** ?*	0
(43)	NV6649_TWD	Hershel greene filles walking dead	0,20
(44)	NV6191_TWD	walking dead hershel greene	0,67
(45)	NV2629_TWD	fille cadette définition	Null

Tableau 12 - Exemples de requêtes découlant de la tâche TWD ainsi que le Tau de Kendall associé

3.3. Évaluation et limites du traitement

Afin de mesurer l'efficacité du traitement automatique des requêtes présenté ci-dessus, une évaluation des différentes performances du système a été opérée. L'objectif est de confronter une partie des sorties obtenues en aval du traitement à une correction manuelle servant de référence. Cette évaluation permet ainsi de porter un regard critique sur les performances du système en mettant en évidence ses réussites ainsi que ses potentielles lacunes. Une première section sera consacrée à l'évaluation des différentes métriques calculées automatiquement. Une seconde portera sur les potentielles pistes d'optimisation quant au fonctionnement du traitement.

3.3.1. Évaluation de la fiabilité du système de traitement

Afin d'évaluer l'efficacité des différents niveaux de traitements effectués, il a été choisi d'opérer une vérification manuelle des résultats obtenus lors du traitement de 100 requêtes choisies aléatoirement. Le système de traitement ayant été conçu sur la base des 100 requêtes initiales présentes au sein de la tâche TWD, il a été choisi d'évaluer le système sur un ensemble de requêtes tirées des tâches TMR et TSF. Ainsi, un générateur de nombres aléatoires a fourni 50 valeurs aléatoires entre 1 et 1645 (correspondant aux requêtes de la tâche TMR) puis 50 autres valeurs entre 1 et 1519 (correspondant aux requêtes de la tâche TSF). Les requêtes correspondantes ont été récupérées puis ont permis de réaliser une évaluation du système en deux étapes, en relation avec les deux dimensions de traitement présentées précédemment. La dimension lexicale a été évaluée grâce à la construction d'une matrice de confusion qui évalue le bon traitement de chaque token selon la taxinomie de stratégies lexicales. Les valeurs relatives à la dimension structurelle ont été évaluées en regardant le nombre de requêtes affichant une valeur correcte afin d'obtenir une série de scores d'exactitude, encore une fois sur la base de la correction des 100 requêtes de référence.

Le Tableau 13 présente la matrice de confusion synthétisant la comparaison entre le traitement obtenu par le biais du système et la correction manuelle apportée sur les 100 requêtes sélectionnées. Pour chacun des 484 tokens présents dans l'ensemble de requêtes, on vérifie alors si l'étiquette qui lui a été attribuée par le système correspond ou non à un traitement valide. La matrice oppose alors en colonnes les étiquettes de références obtenues après un traitement à la main, et en lignes les étiquettes produites par le système.

Système / Gold	Recopie	Modification	Ajout	Total réel
Recopie	**299**	0	**5**	304
Modification	0	**9**	0	9
Ajout	**3**	0	**168**	171
Total prédit	302	9	173	484

Tableau 13 - Matrice de confusion utilisée pour comparer les stratégies lexicales produites par le système avec celles de référence

On s'aperçoit alors que les stratégies lexicales sont presque systématiquement bien catégorisées par le système en témoigne la proportion de vrais positifs, ici mis en surbrillance en vert. On remarque que les seules erreurs de traitement concernent les catégories Recopie et Ajout. En effet, 5 ajouts sont considérés comme des recopies par le système. Inversement, 3 recopies sont injustement catégorisées comme ajouts. Toutes les modifications ont quant à elles bien été étiquetées. À noter que les 100 requêtes ne comprenaient que 9 cas de modifications. Cependant, pour ces derniers, on vérifiait également si le type de modification correspondait bien à la sortie attendue. En l'occurrence, le système n'a pas confondu les différentes catégories de modifications présentées précédemment. Pour la dimension lexicale, on obtient ainsi une f-mesure égale à 0,983. Cela démontre que, dans plus de 98 % des cas, le système est en capacité de correctement attribuer la bonne étiquette aux tokens de la requête.

$$Fmesure = \frac{2 \times VP}{2 \times VP + FP + FN} = \frac{2 \times 476}{2 \times 476 + 8 + 8} = \frac{952}{968} = 0{,}983$$

D'un point de vue structurel, l'efficacité du système a été évaluée en dénombrant, pour les 100 requêtes corrigées à la main, le nombre de celles que le système avait correctement traitées. On obtient alors deux scores d'exactitude. Le premier correspondant au nombre de tokens par requête et le second au type de requête selon la distinction langage naturel/mots-clés. Évaluer le reste des variables n'apparaissait pas comme essentiel car les stratégies de permutation ainsi que de suppression dépendent entièrement de la catégorisation des tokens ainsi que de la segmentation de la requête. Le Tableau 14 représente alors les scores d'exactitude obtenus.

Métrique évaluée	Score d'exactitude
Nombre de tokens	98/100
Type de requête	96/100

Tableau 14 - Scores d'exactitude obtenus lors de l'évaluation des variables *Nombres de token* et *Type de requête*

Les scores montrent que dans 98 % des cas, le système est en capacité de dénombrer le nombre de tokens présents dans les requêtes. De plus, dans 96 % des cas, le traitement a correctement caractérisé les requêtes. On observe que deux requêtes composées en mots-clés ont été catégorisées comme langage naturel par le système. Et que inversement, deux requêtes formulées en langage se sont vues considérées comme composées en mots-clés. Globalement, cette évaluation montre que le système est globalement très fiable. Néanmoins, certaines erreurs dont les causes sont identifiables empêchent un traitement qu'on pourrait considérer comme idéal. Les erreurs en question sont présentées dans la section suivante.

3.3.2. Limites et perspectives d'optimisation

La comparaison entre les performances du système et celles de la version corrigée, réalisée sur 100 requêtes sélectionnées aléatoirement, nous permet d'identifier les éléments qui réduisent l'efficacité du traitement. En effet, on observe quatre facteurs qui conduisent le traitement à commettre des erreurs.

- **Présence d'une langue étrangère** (Impacte : *Type de requête*)

Bien que le système soit explicitement conçu pour repérer des cas de traductions. La présence de langues étrangères au français peut causer des erreurs. Prenons l'exemple **(46)** qui correspond à une requête soumise dans le cadre de la participation à la tâche TMR. Cette dernière est entièrement formulée en anglais. Pour rappel, la détection des stratégies de modification ne prend en compte que les mots pleins. Ainsi, l'analyse de la requête à un niveau lexical ne sera pas compromise. Néanmoins, le système a considéré la requête suivante comme une concaténation de mots-clés. Or, on repère au sein de cette dernière deux mots-outils, à savoir "the" et "of". Ces mots n'appartenant pas au lexique français, le système n'a pas pu les prendre en considération. Dans une perspective d'optimisation du système, il faudrait alors faire en sorte que le système puisse également traduire les mots-outils lorsque ce dernier se penche sur la distinction Mots-clés / Langage naturel. De plus, les 100 requêtes sélectionnées aléatoirement comprenaient une requête en néerlandais **(47)**. Il semble que le participant V2060, également dans le cadre de la tâche TMR, ait subitement en milieu de session choisi de formuler deux requêtes en langue néerlandaise. La requête en question a ainsi été injustement considérée comme composée en mots-clés. En l'occurrence, la présence surprenante d'une autre langue étrangère que l'anglais n'a pas causé d'autres erreurs de traitement. Néanmoins, si le système venait à être réutilisé sur d'autres ensembles de données, il serait judicieux que ce dernier puisse aussi prendre en compte d'autres langues étrangères.

(46) *the rape of europa*
(47) *Aegina wacht op de komst van Zeus*

- **Normalisation trop agressive** (Impacte : *Stratégies lexicales*)

Ensuite, la normalisation qui est effectuée lors de la détection des recopies, et qui pour rappel affecte autant les tokens de l'énoncé que ceux de la requête, apparaît dans certains cas comme trop agressif. Un exemple spécifique démontre cela, celui du token "à". Pour rappel, les diacritiques de tous les caractères sont supprimées afin de trouver des correspondances dans les cas où le participant aurait par exemple omis un accent lors de la recopie d'un token. Or, il y a une différence fondamentale entre le verbe "a" et la préposition "à". Pourtant, au yeux du système, ces deux éléments sont alors similaires. Par exemple, le système a considéré le token "a" présent dans l'exemple **(48)** comme une recopie du token "à" figurant dans l'énoncé de la tâche TMR. Ce genre de cas crée des faux positifs qui nuisent à l'étiquetage des tokens de la requête. Globalement, le Tableau 13 démontre que ces cas de figures ont un impact moindre sur le traitement. Néanmoins, il serait bénéfique de faire en sorte que certains cas particuliers propres aux spécificités de la langue française ne soient pas pris en compte.

(48) *a quoi sert un système d'exploitation*

- **Portée insuffisante du seuil de similarité** (Impacte : *Stratégies lexicales*, *Type de requête*, *Stratégies de permutation*, *Stratégies de suppression*)

Il arrive que le seuil de similarité qui repose, pour rappel, sur le calcul d'une distance d'édition reportée sur le nombre de caractères présents dans le token, ne suffise pas à aller au-delà de certaines variations typographiques. Par exemple, dans l'exemple (49), on observe que le participant a soumis une requête comportant le token "*ma*". Or, si on regarde le contexte autour du token, on se rend compte que le participant voulait formuler le token "la", figurant dans l'énoncé de la tâche TWD. Ici, le seuil de similarité n'a pas pu établir de correspondance entre ces deux éléments et a étiqueté "*ma*" en tant qu'ajout. Équilibrer le seuil de similarité devient alors délicat. Une valeur trop élevée pourrait créer des faux négatifs, et au contraire, une valeur trop faible risque d'entraîner l'apparition de faux positifs. Le choix d'établir cette valeur autour de 0,75 a permis de trouver un compromis optimal. Néanmoins, certains rares cas de variations typographiques trop aberrantes nuisent à l'efficacité globale du système. Ainsi, il serait peut-être judicieux de mettre à profit un vrai correcteur orthographique plutôt qu'une mesure de similarité au caractère.

(49) *nom de la deuxieme fille de Hershel Greene dans ma serie Walking dead*

- **Suppression des caractères d'espaces** (Impacte : *Stratégies lexicales*, *Nombre de tokens*, *Stratégies de permutation*, *Stratégies de suppression*)

Enfin, on observe que, dans certains rares cas, le participant a omis certains caractères d'espaces. Vu que la requête est traitée token par token, si deux tokens sont mal segmentés par le participant, alors le traitement s'en trouvera fortement affecté. Par exemple, la requête **(50)** contient le token "*Guybetprete*" qui découle d'une concaténation entre "Guybet" et "prête", deux éléments constitutifs de l'énoncé TSF. Or, le système va alors considérer cet élément comme un ajout au lieu de deux recopies distinctes. Cela va alors impacter le nombre de tokens qui est attribué à la requête mais également un grand nombre d'autres variables mesurées par le système. En l'occurrence, lors de l'évaluation de la fiabilité du traitement, ces cas exceptionnels ont réduit de 2% le score d'exactitude lié à la variable "Nombre de tokens". Idéalement, il faudrait alors trouver un moyen de contourner ces exceptions typographiques.

(50) *quel est le lien entre la version 1.2 du systeme d'exploitation représenté par une spriale, associé à celui avec le pingouin et le ling métrage d'animation réalisé par Peter Docter où Henri Guybetprete sa voix*

Pour conclure, l'évaluation de la fiabilité du système de traitement des requêtes révèle que les erreurs produites demeurent marginales par rapport au volume total de données traitées, ce qui permet de considérer les résultats obtenus comme suffisamment robustes pour la suite de l'analyse. Néanmoins, certaines erreurs redondantes que commet le système pourraient être corrigées dans le futur afin d'accroître sa fiabilité.

4. Analyses et résultats

Le traitement de cet ensemble de données nous permet alors de procéder à un certain nombre d'analyses qui ont pour but de répondre à plusieurs objectifs. En premier lieu, nous mettrons à profit le traitement structurel qui a été opéré sur les requêtes. Nous nous attarderons alors sur les caractéristiques structurelles propres aux requêtes initiales. En second lieu, nous examinerons les propriétés lexicales des premières requêtes. Nous nous appuierons alors sur le traitement taxinomique appliqué à l'ensemble des requêtes initiales afin d'en examiner la composition lexicale. Enfin, nous tirerons parti du traitement de l'ensemble du jeu de données PRIVaThe afin de comparer les propriétés structurelles et lexicales des requêtes initiales par rapport aux reformulations. Ces différentes sections seront suivies d'une discussion visant à interpréter les résultats obtenus.

4.1. Variations lors de la formulation en requête : Aspect structurel

Dans un premier temps, le traitement effectué nous permet de nous pencher sur les propriétés structurelles des requêtes initiales du corpus PRIVaThe. Nous nous attarderons alors sur leur taille, la manière dont elles sont structurées, ainsi que sur les potentielles stratégies de permutation observables en leur sein.

4.1.1. Taille des requêtes

On peut d'abord supposer que l'énoncé de la tâche va avoir une incidence sur la taille des requêtes initiales. En effet, la production du participant découle de son interaction avec ses différents énoncés. Comme le rappelle le Tableau 15, ces énoncés possèdent des caractéristiques variables. Les énoncés des variantes 1 et 2 de la tâche TMR sont significativement plus longs que ceux des tâches TWD et TSF. À noter qu'il n'y a aucune différence de taille entre les deux variantes de la tâche TSF. Également, l'énoncé propre à la tâche de recherche TWD se distingue des autres car il ne contient qu'une seule thématique de recherche. On peut alors émettre l'hypothèse que plus l'énoncé de la tâche va posséder une densité linguistique importante, plus la requête qui en découle sera longue. Afin de tester cette hypothèse, le Tableau 15 indique dans ses deux dernières colonnes la taille moyenne puis médiane des requêtes selon la taille de l'énoncé.

Tâche	Nombre de tokens	Nombre de thématiques	Longueur moyenne en tokens	Longueur médiane en tokens
TWD	17	1	5,93	5
TMR (Variante 1)	108	2	3,72	3
TMR (Variante 2)	113	2	3,64	3
TSF	39	2	6,31	5

Tableau 15 - Nombre de tokens et de thématiques présentes dans les énoncés lié aux tâches ainsi que la longueur moyenne et médiane respective des requêtes initiales associées

On s'aperçoit alors que les résultats varient fortement entre les requêtes initiales de la tâche TMR et les deux autres tâches de recherche. Les participants à la tâche TWD produisent en moyenne des requêtes initiales de 5,93 tokens contre 6,31 tokens pour ceux de la tâche TSF. En opposition, les requêtes initiales découlant de la tâche TMR, peu importe la variante, sont en moyenne significativement plus courtes avec une moyenne de 3,72 pour les requêtes de la variante 1. Les longueurs médianes observées confirment cette tendance. Les requêtes initiales des deux tâches TWD et TSF obtiennent une longueur médiane de 5 tokens chacune. Les deux variantes de la tâche TMR produisent, quant à elles, des requêtes initiales d'une longueur médiane de 3 tokens. Ainsi, notre hypothèse n'est pas vérifiée. La taille de l'énoncé a bien une incidence sur la taille des requêtes initiales produites mais dans le sens inverse de ce qui était attendu. Plus l'énoncé est long, plus les requêtes initiales tendent à être courtes. Un coefficient de corrélation de Spearman (rho = 0,29) vient vérifier cette corrélation avec une p-value significative affichée de 2.265e-07. En addition, la Figure 6 représente la distribution de la taille des requêtes initiales selon les différentes tâches de recherche. Ces dernières sont ordonnées dans un ordre croissant selon leur taille respective.

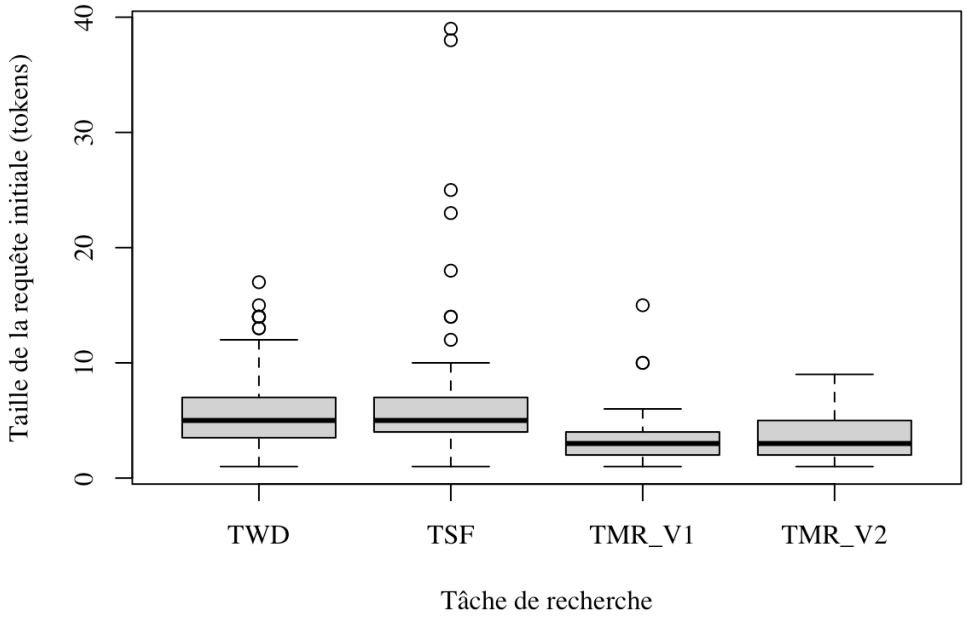

Figure 6 - Distribution de la taille des requêtes initiales (en tokens) selon la tâche de recherche

On observe ainsi, en accord avec les résultats obtenus, que les tâches TWD et TSF présentent des distributions plus étendues et des médianes plus élevées que les deux variantes de la tâche TMR. On note également la présence de valeurs aberrantes positives plus nombreuses et plus élevées pour les tâches TWD et TSF, témoignant d'une plus grande variabilité individuelle dans la formulation des requêtes initiales. Ce résultat peut paraître contre-intuitif, mais lorsque l'on regarde les données, on s'aperçoit rapidement que la taille imposante de l'énoncé de la tâche TMR a tendance à dissuader les utilisateurs de copier tel quel un segment de ce dernier. Là où les utilisateurs ont beaucoup plus tendance en TWD ou TSF à recopier littéralement une partie, voire l'intégralité de la consigne. En ce sens, la Figure 7 nous montre le nombre moyen de tokens correspondant à des recopies d'un token de l'énoncé lors de la première formulation en requête. Un nombre plus important de ces derniers sous-entend alors que la requête correspond à une recopie globale plus fidèle à l'énoncé initial.

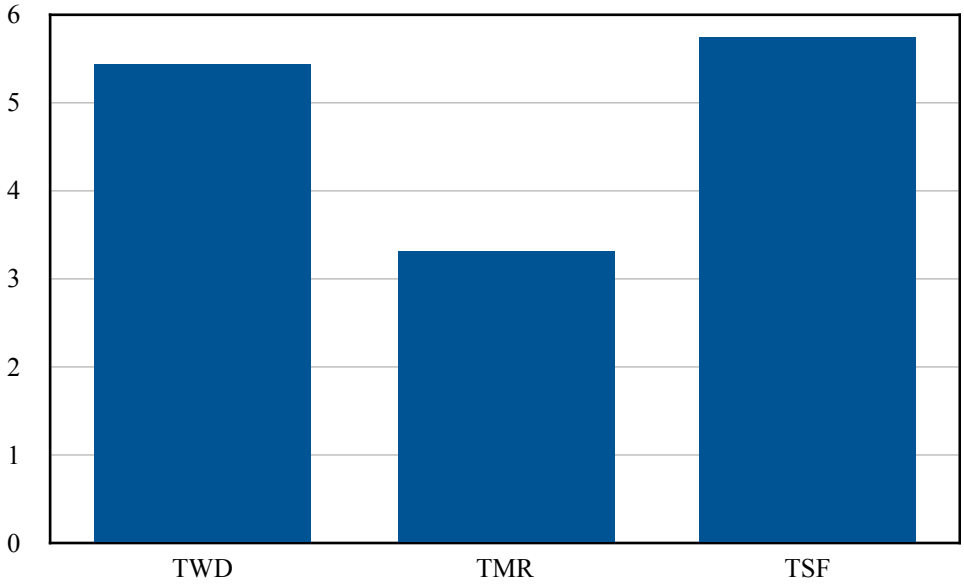

Figure 7 - Nombre moyen de tokens correspondant à une opération de recopie au sein des requêtes initiales selon la tâche

On peut alors remarquer que les requêtes initiales découlant de la tâche TMR contiennent en moyenne près de 3 tokens recopiés contre plus de 5 pour les deux autres tâches. Ce résultat corrobore celui obtenu précédemment comme quoi une taille plus importante de l'énoncé dissuaderait les participants à recopier celui-ci. On remarque également que le nombre de thématiques contenus dans les énoncés n'impacte aucune de ces variables.

4.1.2. Types des requêtes

Dans cette section, nous nous intéresserons à la distinction Langage naturel / Mots-clés telle que peut impacter les requêtes initiales. Pour rappel, une requête formulée en langage naturel est considérée comme telle lorsque elle contient deux mots-outils ou plus. Il s'agit de cas où le participant structure sa requête en suivant les règles syntaxiques qui s'appliquent à une conversation courante. Ces dernières s'opposent ici, selon la typologie de Hearst (2009), aux requêtes composées en mots-clés. Ces requêtes en mots-clés consistent en une concaténation d'éléments que le participant juge pertinents pour retranscrire son besoin informationnel. Toute requête qui contient un seul mot-outil ou moins sera étiquetée comme composée en mots-clés. Naturellement, ces requêtes vont conduire le participant à supprimer une plus grande partie de l'énoncé. Ainsi, on observe une corrélation significative entre le pourcentage de suppressions et le type de requête (p-value égale à 5.027e-10 selon le test de Wilcoxon).

Dans un premier temps, nous nous intéresserons à la stratégie de formulation globalement privilégiée par les participants. Lors de la présentation du corpus, nous avions noté un nombre important de requêtes formulées en mots-clés. Cette observation se confirme à l'analyse des requêtes initiales, où la formulation par mots-clés s'avère nettement majoritaire parmi les requêtes des participants. La Figure 8 représente la répartition de ces deux types de requêtes selon les différentes tâches de recherche. Les résultats montrent que sur les 300 requêtes étudiées, 240 sont formulées en mots-clés. Cette tendance se vérifie sur l'ensemble des trois tâches. Globalement, près de 75 % du temps, les participants décident de s'éloigner de la formulation en langage naturel. Bien que leur besoin d'information soit dicté par un énoncé lui-même en langage naturel, ils choisissent de se distancer de ce mode d'énonciation.

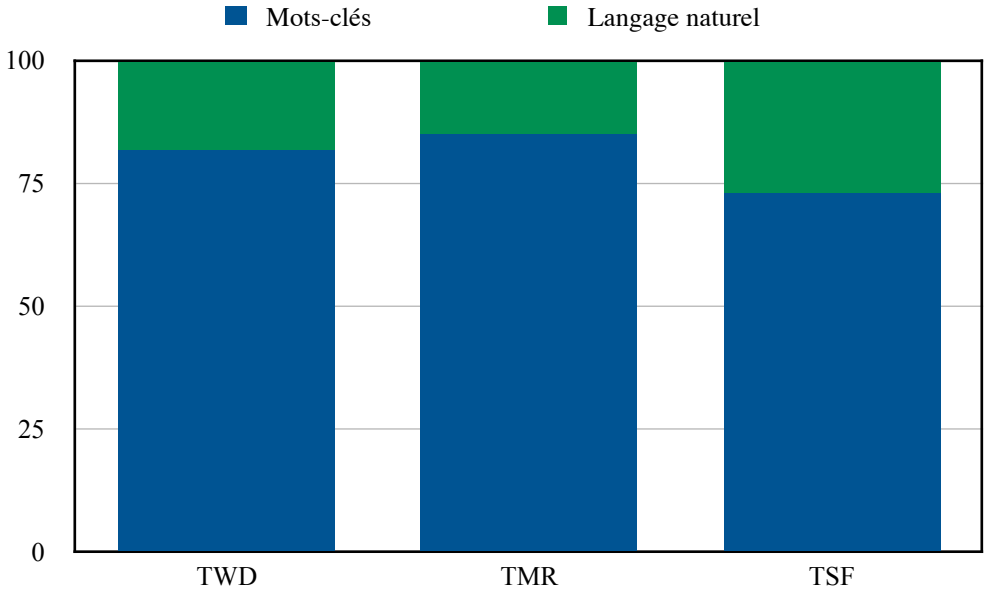

Figure 8 - Répartition de la fréquence d'apparition des types de requêtes lors de la requête initiale selon la tâche

Ce facteur va alors inévitablement impacter la taille des requêtes formulées par les participants. Logiquement, une requête en langage naturel aura tendance à être plus imposante de taille qu'une requête en mots-clés. En effet, on observe que les requêtes en langage naturel font en moyenne 11 tokens, contre seulement 4 tokens de long pour les requêtes en mots-clés. On peut alors supposer que les participants adoptant une formulation par langage naturel ont tendance à largement recopier l'énoncé de la tâche. Cette hypothèse est largement vérifiée car on observe que le taux de recopies associé aux requêtes en langage naturel est en moyenne de 0,93. Cela signifie que les requêtes formulées ainsi contiennent en moyenne 93 % de tokens recopiés. Cependant, cette valeur s'élève également à 0,89 % pour les requêtes en mots-clés. Ainsi, malgré des tailles de requêtes sensiblement différentes selon la stratégie adoptée, les participants recopient l'énoncé dans des proportions comparables, qu'ils formulent en langage naturel ou en mots-clés. Le fait qu'il y ait autant de tokens recopiés dans les deux cas indique que les participants s'appuient sur l'énoncé de manière similaire, indépendamment de la stratégie de formulation choisie. Seulement, une rare proportion d'entre eux va également choisir d'incorporer un certain nombre de mots-outils, suggérant que la formulation en langage naturel découle davantage de la structure même de l'énoncé que d'un choix délibéré du participant.

Néanmoins, on observe que certains participants ont sélectionné quelques mots-clés figurant dans l'énoncé avant d'adjoindre ces derniers à une formulation en langage naturel. Ces rares cas sont observables au sein de la tâche TMR car cette dernière n'est pas formulée sous forme de question. Au total, sur les 60 requêtes formulées en langage naturel, 19 d'entre elles sont formulées sous forme de question. Seulement 4 d'entre elles découlent de la tâche TMR et correspondent ainsi à des reformulations de l'énoncé. Par exemple, la requête **(51)** comporte une requête initiale qui a pour objectif d'obtenir du contexte sur un des référents cités au sein de l'énoncé, ici le "père d'Héraclès" en l'occurrence. Cette requête témoigne d'une stratégie d'exploration de l'énoncé, le participant cherchant à résoudre une ambiguïté référentielle avant d'aborder la tâche dans sa globalité, ce qui le conduit naturellement à formuler sa requête en langage naturel.

(51) *qui est le père d'heraclès*

La Figure 8 nous indique que la formulation en mots-clés est fortement majoritaire selon les tâches. Néanmoins, il reste intéressant de regarder si un même utilisateur aura tendance à adopter le même mode de formulation tout au long de l'expérience. Le Tableau 16 correspond ainsi au nombre de participants ayant adopté une stratégie de formulation uniforme ou non à travers les trois tâches. Une stratégie de formulation par mots-clés ou langage naturel signifie que l'utilisateur a formulé l'intégralité de ces trois requêtes selon la dite stratégie. Une stratégie de formulation combinée fait référence aux cas où l'utilisateur a alterné entre ces deux types de formulation selon la tâche.

Stratégie de formulation	Nombre de participants
Mots-clés	63
Combiné	31
Langage naturel	6

Tableau 16 - Distribution des profils de formulation des requêtes initiales par participant

Les résultats montrent une nette prédominance de la stratégie par mots-clés. 63 % des participants ont exclusivement formulé leurs requêtes en mots-clés. En opposition, seulement 6 d'entre eux ont opté pour une formulation en langage naturel de manière systématique. 31 participants ont alterné entre les deux. À noter que sur l'ensemble de ces derniers, 20 d'entre eux ont formulé deux requêtes en mots-clés pour une seule en langage naturel. Pour conclure, ces résultats nous montrent que la nature des énoncés n'a pas conduit les participants à adopter une stratégie de formulation similaire dans leurs requêtes initiales, les mots-clés s'imposant comme mode de formulation dominant.

4.1.3. Stratégies de permutation

Cette section est consacrée à l'étude des phénomènes de permutations observables au sein des requêtes initiales contenues au sein du corpus PRIVaThe. Pour rappel, on définit une opération de permutation comme la modification par le participant de l'ordre canonique des éléments composant l'énoncé de la tâche de recherche. Pour cela, on regarde lors du traitement si les mots-pleins de l'énoncé sont ordonnés de la même manière dans la requête. Sont alors pris en compte toutes les correspondances entre la tâche de recherche et la requête, c'est-à-dire tous les tokens recopiés ou modifiés. Ensuite, on mesure ce phénomène de permutation grâce au Tau de Kendall qui fournira alors une distance entre l'ordonnancement des termes de l'énoncé et ceux de la requête traitée. Si cet indice est égal à 0, alors le participant n'aura interchangé aucun mot plein de l'énoncé. Au contraire, si le résultat est égal à 1, cela signifie que la requête comporte un ordre de tokens complètement modifié.

Ibarboure (2025) traitait déjà du sujet des permutations au sein des requêtes initiales du jeu de données PRIVaThe. Globalement, on observait que l'ordre d'énonciation des thématiques contenues dans les énoncés TMR et TSF impactait directement la formulation de la requête. Ainsi, les participants avaient davantage tendance à reprendre les différentes thématiques dans l'ordre où elles leur étaient présentées, l'ordre de présentation influençant directement leur ordre d'apparition dans la requête. Nous ne nous pencherons alors pas sur l'aspect thématique de ces requêtes initiales. La mesure attribuée à chaque requête nous permet néanmoins d'opérer une étude se situant au niveau du terme. Le Tableau 17 représente ainsi l'ordre des tokens pris en compte lors du traitement des requêtes quant au calcul du Tau de Kendall. Chacun des tokens se voit alors attribué un chiffre qui correspond à une position spécifique au sein de l'énoncé, puis ensuite dans la requête.

TWD	nom **[1]** deuxième **[2]** fille **[3]** Hershel **[4]** Greene **[5]** série **[6]** Walking **[7]** Dead **[8]**
TMR_V1	mythologie **[1]** grecque **[2]** père **[3]** Héraclès **[4]** transformé **[5]** fois **[6]** animal **[7]** séduire **[8]** femmes **[9]** mortelles **[10]** humaines **[11]** transformation **[12]** sujet **[13]** nombreuses **[14]** peintures **[15]** Renaissance **[16]** italienne **[17]**
TMR_V2	Renaissance **[1]** italienne **[2]** intéressée **[3]** mythologie **[4]** grecque **[5]** peintures **[6]** transformations **[7]** animales **[8]** pouvaient **[9]** prendre **[10]** personnages **[11]** père **[12]** Héraclès **[13]** transformée **[14]** fois **[15]** animal **[16]** séduire **[17]** femmes **[18]** mortelles **[19]** humaines **[20]**
TSF_V1	lien **[1]** long **[2]** métrage **[3]** animation **[4]** réalisé **[5]** Pete **[6]** Docter **[7]** Henry **[8]** Guybet **[9]** prête **[10]** voix **[11]** version **[12]** système **[13]** exploitation **[14]** représenté **[15]** spirale **[16]** associé **[17]** pingouin **[18]**
TSF_V2	lien **[1]** version **[2]** système **[3]** exploitation **[4]** représenté **[5]** spirale **[6]** associé **[7]** pingouin **[8]** long **[9]** métrage **[10]** animation **[11]** réalisé **[12]** Pete **[13]** Docter **[14]** Henri **[15]** Guybet **[16]** prête **[17]** voix **[18]**

Tableau 17 - Ordre des tokens pris en compte lors du calcul du Tau de Kendall pour chacune des tâches

Avant tout, on observe un Tau de Kendall moyen égal à 0,16 sur l'ensemble des 300 requêtes initiales. Seulement 62 d'entre elles possèdent un indice supérieur à 0. Ces stratégies de permutation apparaissent alors relativement rares à l'échelle du corpus. De plus, on observe une forte disparité selon la tâche. En effet, 33 de ces requêtes initiales découlent de la tâche TWD, 21 proviennent de la tâche TMR, et seulement 8 sont liées à la tâche TSF. C'est donc ces 62 requêtes qui vont alors attirer notre attention.

À noter qu'on observe également une corrélation significative entre le fait de permuter des tokens et de formuler une requête en mots-clés (p-value de 4.051e-05 suite à un test non-paramétrique de Wilcoxon). Le fait de formuler sa requête en mots-clés augmente la chance de permuter et de réaliser des opérations de permutations. Cela peut être expliqué par le fait que les requêtes en langage naturel consistent surtout en des recopies strictes de segments de l'énoncé, comme observé précédemment. Cependant, on remarque que les requêtes interrogatives de la tâche TMR, citées précédemment, contiennent des opérations de permutations. En effet, les participants restructurent ici l'agencement syntaxique de l'énoncé, conduisant à une modification des termes de l'énoncé. C'est le cas dans l'exemple **(52)** qui correspond à une requête appartenant à la variante 1 de la tâche TMR. Cette dernière obtient un Tau de Kendall de 0,83. En effet, on s'aperçoit que le participant restructure totalement l'énoncé afin de former une phrase interrogative, ce qui le conduit à opérer une inversion presque totale des éléments provenant de l'énoncé.

(52) *en quel animal***[7]** *s'est transformé***[5]** *le père***[3]** *d'heracles***[4]**

La deuxième cause observable qui conduit les participants à permuter est celle qui était attendue. On observe que chacune des 62 requêtes comportant des permutations place en tête un terme qui n'occupe pas la première position dans l'énoncé, témoignant d'une tendance des participants à réordonner les tokens en fonction d'un traitement informationnel opéré par ces derniers. Les participants identifient alors un token ou un groupe de tokens qu'ils jugent représentatif du besoin d'information et vont ensuite réorganiser les éléments de l'énoncé en fonction de cela. Les trois exemples ci-dessous illustrent ce phénomène.

(53) *hershel***[4]** *green***[5]** *daughter***[3]**
(54) *renaissanced***[16]** *italienne***[17]** *peinture***[15]** *père***[3]** *d'héraclès***[4]**
(55) *henri***[15]** *guybet***[16]** *et pete***[13]** *docter***[14]**

Ainsi, l'exemple **(53)** tiré de la tâche TWD présente une requête affichant un indice d'altération égal à 0,67. On observe qu'ici le participant a choisi de permuter le groupe de tokens "*Hershel Greene*" afin de le placer avant la traduction du mot "fille". La requête met alors en avant ce groupe d'éléments en question. Néanmoins, on peut également attribuer cette permutation à la volonté de reproduire une syntaxe anglophone en suivant la structure génitive de l'anglais, telle que dans le syntagme "Hershel Greene's daughter". L'exemple **(54)** contient une requête découlant de la variante 1 de la tâche TMR. Cette dernière possède un Tau de Kendall de 0,67 également. Ici, on remarque que le participant a choisi de permuter les deux thématiques, un cas de figure minoritaire selon ibarboure (2025). Le participant a alors très clairement jugé, malgré l'ordre de présentation des éléments dans l'énoncé, que la thématique "Renaissance italienne" était plus représentative du besoin d'information que "Mythologie grecque". Il opère alors plusieurs permutations en plaçant d'abord le syntagme nominal "renaissance italienne" avant d'ajouter "peinture" puis de changer de thématique en faisant référence au "père d'Héraclès". Enfin, le dernier exemple **(55)** correspond à une production liée à la variante 2 de la tâche TSF ayant obtenu un indice de permutation de 0,67 une nouvelle fois. On observe que le participant a interchangé les deux groupes de noms propres "Pete Docter" et "Henri Guybet". On peut théoriser que le participant a jugé que le groupe que l'information "Henri Guybet prête sa voix" est plus représentatif du besoin d'information que le groupe "réalisé par Pete Docter". On peut également produire l'hypothèse que le participant a simplement adjoint du groupe de mots "*henri guybet*" en premier car il venait de lire ce dernier.

Ces observations qualitatives nous permettent de tirer plusieurs conclusions. En premier lieu, la majorité des éléments concernés par des opérations de permutations apparaissent comme étant des noms propres **(56)**. On peut relier cela au fait qu'ils sont très fréquents au sein des requêtes mais on remarque néanmoins que 44 des 62 requêtes comportant un indice supérieur à 1 débutent par un nom propre qui n'occupait pas cette position dans l'énoncé. Les noms propres sont alors considérés comme des éléments clés par les participants qui auront alors plus tendance à les manipuler. En second lieu, on observe que les participants ont plus tendance à déplacer des groupes de tokens plutôt que des éléments individuels **(57)**. 43 requêtes comportent alors la permutation d'un de ces groupes de mots, qui sont d'ailleurs souvent des noms propres. Ces groupes correspondent majoritairement à des entités nommées ou à des syntagmes nominaux figés, tels que "walking dead", "hershel greene" ou "renaissance italienne", que les participants semblent traiter comme des unités indivisibles lors de la formulation de leur requête.

(56) *pete***[6]** *docteur***[7]** *long***[2]** *metrage***[3]** *animation***[4]**
(57) *peinture***[6]** *mythologie***[4]** *grecque***[5]** *ET renaissance***[1]** *italienne***[2]**

4.2. Variations lors de la formulation en requête : Aspect lexical

Dans un second temps, le traitement lexical décrit dans la *Section 3.1* nous permet de mener une étude sur l'organisation interne des requêtes initiales. En effet, nous avons différenciés trois types d'opérations qui peuvent alors catégoriser chacun de ces tokens de la requête. Ces opérations correspondent à des stratégies mises en place par l'utilisateur afin de répondre à son besoin d'information, ici induit par un énoncé en langage naturel. Il peut alors recopier un des éléments constitutifs de cet énoncé. Il peut également en modifier un de ces derniers en le traduisant, le transformant en une abréviation, ou bien en opérant sur celui-ci une conversion morphologique. Enfin, il peut tout aussi bien ajouter un élément qui ne figure pas dans l'énoncé. Ainsi, notre objectif est de quantifier ces phénomènes mais aussi d'en décrire les manifestations concrètes au sein du jeu de données. Afin de les dénombrer, le système produit, pour chacun d'entre eux, deux mesures distinctes pour chaque requête. La première consiste en une addition de chacune des étiquettes attribuées aux tokens de la requête. Si une requête comporte 10 tokens dont 5 recopies et 5 modifications, alors son nombre d'ajouts et de modifications sera logiquement égal à 5. La seconde consiste à diviser le nombre attribué à chaque opération par la taille totale de la requête. Par exemple, si une requête contient 5 tokens et que 3 d'entre eux sont des ajouts, alors le taux d'ajouts sera égal à 0,60.

4.2.1. Stratégies de recopie

En premier lieu, l'utilisateur peut recopier une partie ou l'intégralité de l'énoncé afin de le retranscrire dans ses requêtes. Il est important de préciser que les participants, dans le cadre de la constitution du jeu de données PRIVaThe, ne pouvaient pas copier à l'aide d'un raccourci-clavier une partie de l'énoncé afin de la coller directement au sein du moteur de recherche. Les énoncés étant fournis sous une forme physique. Toutes les actions de recopie effectuées sur des éléments spécifiques de l'énoncé reflètent ainsi une volonté précise du participant.

Comme observé précédemment, le nombre de recopies apparaît comme très important au sein des requêtes initiales. Les tables de fréquences des lemmes figurant dans les requêtes initiales (disponibles en *Annexe 1*) montrent que la plupart des tokens constitutifs des productions des apprenants sont déjà présents dans l'énoncé. Ces observations nous montrent déjà que les opérations de recopies sont particulièrement fréquentes au sein des requêtes initiales des participants. De plus, d'autres indicateurs nous montrent que les recopies sont omniprésentes. En effet, nous avons découvert que le taux de recopies est sensiblement similaire entre les requêtes formulées en langage naturel et celles composées en mots-clés. Cela signifie que même si le participant choisit de formuler des requêtes plus longues, ses requêtes seront encore principalement composées de recopies. On n'observe en ce sens aucune corrélation entre la taille de la requête et le taux de recopies selon le coefficient de corrélation de Spearman (p-value égale à 0.8606).

Ainsi, les résultats de l'analyse appuient fortement ces observations. En effet, sur les 1592 tokens contenus dans les 300 requêtes initiales, 1475 ont été étiquetés comme recopies. On observe alors un taux de recopie global d'environ 0,90. Cela signifie que en moyenne, 90 % des tokens composant une requête initiale sont directement issus de l'énoncé de la tâche. Un intervalle de confiance à 95 % calculé sur le taux de recopies confirme cette tendance, avec des bornes comprises entre 0,87 et 0,92, attestant que la recopie constitue bien la stratégie dominante dans la formulation des requêtes initiales du corpus PRIVaThe. De plus, on observe que le taux de recopies est sensiblement similaire selon la tâche de recherche. La tâche TWD affiche un taux de recopies moyen de 0,89. Taux très similaires aux tâches TMR et TSF, qui présentent respectivement un taux de recopies moyen de 0,90 et 0,88. Cela renforce l'idée que la recopie est une stratégie omniprésente et constante, indépendante de l'ensemble des variables recueillies.

Maintenant que ce constat est établi, nous pouvons nous pencher sur les différentes formes que prennent ces recopies. On remarque que certains tokens sont plus propices à être recopiés que d'autres. Dans ce sens, le Tableau 18 présente la fréquence d'apparition des catégories syntaxiques au sein des tokens étiquetés comme recopies au sein des requêtes initiales des trois tâches de recherche.

Catégorie syntaxique	Fréquence absolue d'apparition (sur 1459)
Noms propres	589
Noms communs	412
Préposition	204
Adjectifs	83
Articles	70
Autres	58
Verbes	43

Tableau 18 - Fréquence d'apparition des catégories syntaxiques appartenant aux tokens étiquetés comme recopies

En lien avec ce qui a pu être observé auparavant, les noms propres sont particulièrement fréquents au sein des requêtes initiales. Ce sont les tokens qui sont les plus souvent recopiés avec 589 occurrences. On observe alors que les participants recopient énormément de noms propres quitte à composer leurs requêtes uniquement en recopiant ces derniers. Sur les 300 requêtes initiales, 47 sont composées uniquement de noms propres recopiés. On remarque que, plus l'énoncé de la tâche contient elle-même de noms propres, plus ce nombre augmente. La tâche TWD très fournie en noms propres (4 sur 17 tokens) produit à elle seule 20 de ces requêtes, nécessairement courtes de taille. L'exemple **(58)** en présente l'une d'entre elles. Néanmoins, même quand la tâche de recherche comporte moins de noms propres, les participants ont tendance à se raccrocher à ceux-ci. On observe par exemple qu'au sein de la tâche TMR, dont l'énoncé ne contient que deux noms propres, à savoir "Héraclès" et "Renaissance", que 11 des requêtes formulées ne contiennent qu'une recopie du token "Héraclès". C'est-à-dire que plus d'un participant sur dix a intuitivement supprimé l'intégralité de l'énoncé afin d'uniquement conserver un des deux seuls noms propres.

(58) *walking dead hershell greene*

On observe plus globalement que la plupart des requêtes sont composées uniquement d'associations de noms propres et communs recopiés. C'est par exemple le cas au sein des requêtes correspondant aux exemples **(59)** et **(60)**. On relève ensuite 204 occurrences de prépositions. On remarque que dans 82 de ces occurrences, la préposition "d'" qui est difficilement dissociable de certains noms propres ou communs. On observe alors beaucoup de requêtes comportant "système d'exploitation" ou "père d'Héraclès". Le reste des catégories présentes sont employées afin de composer les requêtes en langage naturel présentées dans la *Section 4.1.2*. À noter que la catégorie *Autre* contient tous les pronoms, les signes de ponctuation, les auxiliaires ainsi que les conjonctions. Encore une fois, ces derniers ne sont employés qu'au sein de requêtes en langage naturel et à des fréquences extrêmement faibles. Il est intéressant de noter que le signe de ponctuation "?" n'apparaît qu'à 2 reprises au sein des 300 requêtes, alors que le participant interroge systématiquement le moteur de recherche. Ce dernier est alors perçu comme un outil à qui l'on soumet des termes plutôt qu'un système à qui l'on adresse une question. Les tokens les plus fréquemment supprimés appartiennent ainsi à ses catégories.

(59) *transformation père héraclès*
(60) *film animation voix guybet*

Ainsi, les participants établissent une hiérarchie entre les différents composants de l'énoncé. Plus un élément est jugé représentatif du besoin d'information, plus ce dernier aura tendance à apparaître. Comme nous l'avons vu, il arrive que certains participants n'établissent pas de hiérarchie entre les éléments de l'énoncé, les conduisant alors à formuler leur requête en langage naturel. On peut alors se pencher plus en détail sur leurs requêtes en langage naturel du point de vue des stratégies de recopies. On observe que sur les 300 requêtes, 17 seulement affichent un taux de recopies de 1 et plus de 9 tokens. Ces requêtes recopient alors l'énoncé dans une proportion significative. Pour rappel, les participants ne pouvaient pas copier-coller la tâche directement sur le moteur de recherche. Par exemple, la requête **(61)** nous montre un utilisateur qui a recopié l'intégralité de la tâche TWD, à l'exception du groupe de tokens "la série" ainsi que du signe de ponctuation "?". Enfin, il est intéressant d'observer que sur les 300 requêtes initiales, seulement une seule consiste en une recopie exacte de l'énoncé. L'exemple **(62)** correspond ainsi à une recopie littérale (à une majuscule près) de la variante 2 de la tâche TSF. Le participant, visiblement non inquiété par la limite de temps liée à la réalisation de la tâche, a alors adopté une stratégie littéralement opposée à celle d'une simple concaténation de noms propres ou communs.

(61) *Quel est le nom de la deuxième fille de Hershel Greene dans walking dead*
(62) *quel est le lien entre la version 1.2 du système d'exploitation représenté par une spirale associé à celui avec le pingouin et le long métrage d'animation réalisé par Pete Docter où Henri Guybet prête sa voix ?*

4.2.2. Stratégies de modifications

En second lieu, l'utilisateur peut choisir d'interagir avec l'énoncé différemment qu'en recopiant simplement un ou plusieurs segments de celui-ci. Il a également la possibilité de modifier un de ces derniers. Comme cela a été explicité, nous distinguons trois types de modifications. Les traductions, c'est-à-dire les cas où le participant a traduit un mot plein de l'énoncé en anglais dans sa requête. Les abréviations qui concernent les cas où le participant à abréger un des termes de la tâche de recherche. Ainsi que les transformations morphologiques qui correspondent à de potentielles opérations de flexion ou de dérivation pouvant affecter des tokens de l'énoncé.

Comme cela a été observé, le nombre de recopies est particulièrement important au sein des requêtes initiales. En moyenne, ces mêmes requêtes sont composées à 90 % de recopies. On peut alors légitimement se demander quel poids pèsent les modifications au sein des 10 % restants. On s'aperçoit alors que les modifications occupent une place minime au sein des requêtes initiales avec un taux de modification global égal à 0,017 avec un intervalle de confiance à 95 % compris entre 0,010 et 0,024. Les modifications sont alors très largement marginales dans les requêtes initiales. En effet, on observe uniquement 25 tokens résultant de modifications au sein des 300 requêtes initiales traitées.

Avant de regarder en détail chacun de ces cas, on peut observer une disparité entre les tâches. On observe une corrélation significative confirmée par un test de Kruskal-Wallis qui affiche une p-value de 0.0141. En effet, on observe que sur ces 25 modifications, 13 découlent de la tâche TWD, 10 de la tâche TMR, et seulement 2 de la tâche TSF. Ce résultat peut en partie être expliqué par le fait qu'aucun token de l'énoncé de la tâche de recherche TSF ne peut être exposé à une opération d'abréviation, réduisant alors le nombre de modifications possibles. De plus, le test de Wilcoxon ne révèle pas de différence significative (p-value égale à 0.1713) du taux de modifications selon le type de requête, ce qui indique que les participants opèrent des modifications à une fréquence similaire qu'ils formulent en langage naturel ou en mots-clés. On peut donc s'attendre à une grande diversité de productions.

Au regard des productions des participants, nous observons que ces trois types de modifications sont utilisés dans des intentions clairement différentes. Le Tableau 10 présente alors les différentes propriétés distinctives des trois types de modifications suite à l'observation de leur manifestation concrète. Comme nous le verrons, chacune des opérations de traduction concerne des cas de conversions sémantiques littérales du français vers l'anglais. Ainsi, le sens du token traduit est systématiquement entièrement conservé. Néanmoins, il est difficile d'attribuer cette opération de modification à une quelconque volonté d'optimiser le processus de matching. Les participants traduisant des éléments de la tâche de recherche semblent le faire instinctivement, en lien avec une supposée compétence de bilinguisme. Néanmoins, ceci peut être relativisé par le fait qu'on observe une corrélation significative entre le fait d'opérer des modifications et de permuter des tokens de l'énoncé (p-value de 4.221e-05 au coefficient de corrélation de Spearman). Ce résultat est à manier avec précaution car le nombre de valeurs non-définies est très élevé au sein de la variable Taux de modifications. Quoi qu'il en soit, on observe que les participants qui modifient, réorganisent presque systématiquement l'énoncé, même dans des cas de traduction.

Type de modification	Modifie le sens du terme	Témoigne d'une volonté d'optimiser la recherche
Traduction	NON	NON
Abréviation	NON	NON / OUI
Transformation morphologique	OUI	OUI

Tableau 19 - Propriétés distinctives des types de modifications telles qu'observées dans les productions des participants

De la même manière, les cas d'abréviation d'un token de l'énoncé ne modifient pas le sens de celui-ci. Néanmoins, ici, on peut attribuer cette action à une potentielle volonté d'optimiser la vitesse de formulation de la requête, cela du point de vue du déroulé de l'expérience plutôt que de l'obtention de l'information en elle-même. En effet, taper au clavier "2e" apparaît comme bien moins chronophage que d'écrire manuellement "deuxième". Le participant, dans l'optique, inconsciemment ou non, de gagner du temps, chercherait alors à trouver une orthographe alternative qui sous-entendrait moins d'actions mécaniques à réaliser. Cette hypothèse n'est malheureusement pas vérifiable. Le participant peut en effet être enclin à réduire l'effort de formulation de sa requête, à l'instar de ce qui est observé dans les requêtes en mots-clés, le système étant théoriquement capable d'identifier le référent associé au token modifié. Enfin, les opérations de transformations morphologiques témoignent systématiquement d'une volonté d'altérer la valeur sémantique du token modifié, afin d'optimiser le processus de recherche et directement l'obtention de l'information.

- **Stratégies de traduction**

On observe au sein des 300 requêtes initiales, 9 cas de traductions qui sont répartis dans 8 requêtes distinctes. Les traductions ne concernent alors que des noms communs, plus spécifiquement les tokens "fille", "nom" et "peinture". 4 de ces traductions sont observables dans les exemples ci-dessous. Nous pouvons observer que chacune de ces requêtes est composée en mots-clés. Bien qu'il soit plus prudent d'attribuer cela à une coïncidence étant donné le faible nombre d'occurrences, on observe que ce résultat diffère fortement au sein des autres types de modifications. Les annotations liées à l'ordonnancement initial des tokens tel que présenté dans le *Tableau 17* ont été ajoutées afin de rendre compte des phénomènes de permutation. Comme cela a pu être observé, on observe que ces conversions d'une langue à l'autre peuvent conduire à une réorganisation des éléments comme avec "*daughter's name*" présent dans l'exemple **(64)**. Dans cette même requête, le participant intervertit également les groupes de tokens "Walking Dead" et "Hershel Greene", témoignant d'une tendance à la réorganisation de l'énoncé.

(63) *hershel***[4]** *green***[5]** *daughter***[3]**
(64) *Walking***[7]** *dead***[8]** *Hershel***[4]** *Greene***[5]** *daughter's***[3]** *name***[1]**
(65) *zeus animal***[7]** *seduction transformation***[12]** *painting***[15]**

- **Stratégies d'abréviations**

Seulement quatre occurrences d'observations sont observables dans les requêtes initiales du corpus PRIVaThe. Ces dernières concernent les tokens "deuxièmes" et "trois" qui sont les seuls tokens figurant dans les énoncés pouvant être affectés par ce genre de conversion. À noter que la tâche TSF ne peut ainsi pas être concernée par ce type de modification. Ainsi, trois participants ont converti le token "trois" en le chiffre "3" comme cela peut être observé dans l'exemple **(68)**. Un seul a opéré la transformation de "deuxième" en "2e". On observe que dans ce cas-là **(67)**, le participant a repris l'ordre de l'énoncé afin de formuler une requête en langage naturel. Cette requête est néanmoins associée à un pourcentage de suppression de 41 %. Le participant a donc globalement transformé l'énoncé en supprimant les éléments qu'il jugeait moins pertinents et en réalisant l'abréviation observée. Dans ce cas, on peut théoriser que le participant voulait effectivement atteindre une certaine efficacité de recherche. On remarque que globalement, depuis le début, les résultats montrent que les participants recherchent un équilibre entre vitesse de formulation et représentativité du besoin d'information. Or, une abréviation permet de conserver entièrement le sens d'un élément de l'énoncé, tout en réduisant le temps nécessaire à la formulation de la requête.

(67) *nom***[1]** *de la 2e***[2]** *fille***[3]** *de Hershel***[4]** *Greene***[5]** *the walking***[7]** *dead***[8]**
(68) *Père***[12]** *d'Héraclès***[13]** *3 transformation en animal***[8]**

• **Stratégies de transformations morphologiques**

On repère 13 occurrences de transformations morphologiques au sein des requêtes initiales, ce qui en fait alors le type de modification le plus fréquent dans ce contexte initial. 12 d'entre eux concernent des cas de flexions morphologiques où le participant modifie alors le nombre ou bien le temps d'un des tokens de l'énoncé. C'est notamment le cas dans l'exemple **(69)** dans lequel figure une flexion en nombre du mot "peintre". Le participant devant trouver le nom de trois peintres distincts dans le cadre de la tâche de recherche TMR, il décide d'utiliser le pluriel. L'ajout de ce simple caractère lui permet en principe de gagner du temps car le token "*peintres*" est équivalent à "les peintres". Le participant semble alors encore à la recherche de ce compromis entre économie de formulation et représentativité. Une requête similaire est observable dans l'exemple **(70)** qui comporte une conversion à la troisième personne du présent de l'indicatif du verbe "séduire". On note que ces deux requêtes sont formulées en langage naturel. Le dernier exemple **(71)** nous montre un cas de dérivation morphologique qui affecte le nom "animal" qui se voit converti en l'adjectif "animale" que le participant a ici adjoint au nom commun "transformation".

(69) *peintres inspirés par la mythologie grecque dans la renaissance italienne*
(70) *peinture héracles qui séduit une femme en animal*
(71) *transformersmation animale mythologie grecque*

Ainsi, les stratégies de modification apparaissent comme un outil parmi tant d'autres afin de communiquer avec le système dans un contexte de recherche informationnelle. Leur faible fréquence d'apparition contraste fortement avec l'importance des recopies, qui demeurent la stratégie privilégiée des participants lors de la formulation de leurs requêtes. On observe alors que ces opérations de modification ne témoignent pas systématiquement d'une intention marquée du participant de guider le moteur de recherche. Les modifications apparaissent ainsi moins comme une stratégie concrète de reformulation que comme un ajustement ponctuel venant compléter d'autres opérations plus fréquentes, telles que la recopie ou la permutation des éléments de la tâche de recherche.

4.2.3. Stratégies d'ajout

Enfin, l'utilisateur peut tout simplement ajouter un nouvel élément. C'est-à-dire un token qui n'apparaît pas dans l'énoncé, même sous une forme modifiée. Le système récupère ainsi tout ce qui n'a pas été classifié en tant que recopie ou modification, y compris pour rappel les mots-outils formulés dans une autre langue que le français. On peut ainsi s'attendre à une certaine variabilité au sein de cette catégorie. Cette dernière est reprise des travaux d'Abu Onq et al. (2025), néanmoins nous n'avons pas distingué plusieurs types d'ajouts lors du traitement. L'objectif étant d'observer, après le traitement des données, quelles formes prennent ces tokens ajoutés.

Comme nous avons pu l'observer, les requêtes initiales présentes globalement un taux de recopie très élevé (en moyenne 0,9) et un taux de modification très faible (en moyenne 0,017). Ainsi, sur les 1592 tokens traités, 1475 étaient étiquetés comme recopies et 25 comme modifications. Logiquement, le reste des tokens correspondent à des opérations d'ajouts. On observe alors un total de 92 ajouts au sein des 300 requêtes observées. Le taux moyen d'ajouts observé est de 0,0857, avec un intervalle de confiance à 95 % compris entre 0,0633 et 0,1081. La répartition est relativement homogène selon la tâche avec 30 tokens ajoutés dans TWD, 27 dans TMR et 35 dans les requêtes initiales de la tâche TSF. De plus, on observe également une corrélation significative entre le fait de permuter des tokens de l'énoncé et d'en ajouter de nouveau (p-value de 0,0378 selon le coefficient de corrélation de Spearman). On peut encore émettre l'hypothèse que les participants vont mettre à profit ces éléments ajoutés afin de les combiner avec d'autres types d'opérations. Cependant, on observe une corrélation non-significative entre le fait de composer sa requête avec des ajouts et de formuler cette dernière selon un type spécifique de requêtes. On peut donc attendre une grande diversité au sein des productions comportant des ajouts. À l'analyse des résultats, on distingue 4 types d'ajouts qui répondent chacun à un besoin différent du participant.

- **Ajouts navigationnels**

On nomme ici ajouts navigationnels, les tokens ajoutés qui visent à accéder à un site internet spécifique. En lien donc avec les requêtes du même nom décrites par Broder (2002) et Rose & Levinson (2004). Abu Onq et al. (2025) mentionnent également ce type d'ajout sous le nom de "Ressource". Ces occurrences sont alors particulièrement rares car on en dénombre que trois sur l'ensemble des 300 requêtes initiales. Les trois requêtes en question visent toutes à accéder à une encyclopédie collaborative en ligne, comme on peut le voir dans les exemples **(72)** et **(73)**. À noter que les tokens ajoutés sont mis en relief en gras. On observe que chacune d'entre elles découlent de la tâche de recherche TWD. Ainsi, les trois participants qui ont opéré des ajouts de ce type ont par la suite totalement abandonné cette stratégie pour la suite de l'expérience. On remarque que le taux de suppression de ces requêtes est alors très élevé. Les participants choisissent d'isoler un segment très précis de l'énoncé afin d'accéder à une ressource spécifique **(72)**, ou alors indiquent directement le site visé **(73)**.

(72) *walking dead* ***wiki***
(73) ***wikipedia***

- **Ajouts structurels**

Les ajouts structurels font ici référence à l'ajout de divers mots-clés qui permettent aux participants de réorganiser la structure de l'énoncé afin de se le réapproprier. On dénombre 20 tokens qui répondent à cet objectif. Par exemple, dans la requête **(74)**, le participant ajoute deux mots-outils afin d'accompagner la modification du verbe "séduire". L'immense majorité de ces requêtes sont alors logiquement formulées en langage naturel. L'exemple (75) correspond à une requête déjà observée auparavant qui comporte l'ajout du pronom interrogatif "quel", la tâche de recherche TMR ne prenant pas la forme d'une interrogation directe. Les mots-outils formulés dans d'autres langues que le français ont également été adjoints à cette catégorie. On observe alors des ajouts tels que "in" ou "of" à des fréquences très faibles.

(74) *peinture héracles* ***qui*** *séduit* ***une*** *femme en animal*
(75) *en* ***quel*** *animal s'est transformé le père d'héraclès pour séduire les femmes mortelles*

- **Ajouts thématiques**

Ici, on fait référence à ce que Abu Onq et al. (2025) définissaient comme des ajouts "d'entités". C'est-à-dire des ajouts en lien avec la thématique de recherche, tokens qui prennent donc des formes assez variées. Concrètement, cette catégorie rejoint celle des ajouts structurels, à l'exception que celle-ci concerne l'ajout d'éléments ayant une valeur sémantique clairement déterminée et en lien avec la ou une des thématiques de recherche. Le participant s'affranchit ainsi de l'énoncé en insérant un élément qu'il juge pertinent quant à l'accès à l'information qu'il souhaite obtenir. Dans la requête **(76)**, le participant a supprimé "long métrage d'animation" afin d'ajouter "*dessin animé*". La requête **(77)** réalise une opération similaire en ajoutant le token "*logo*" en référence au groupe de mots "représenté par" présent dans la tâche TMR. 8 des 34 cas d'ajouts thématiques correspondent à l'incorporation du token "logo". Ainsi, les participants semblent systématiquement réorganiser l'énoncé en privilégiant certains termes thématiques qui diffèrent de ceux présents dans l'énoncé. L'exemple (78) représente une requête qui a ajouté le token "inspirés". Le participant arrive ainsi à passer outre la taille imposante de la tâche TMR en sélectionnant un certain nombre de tokens qu'il va ensuite recopier, avant d'ajouter un certain nombre de tokens qui vont ainsi pouvoir optimiser le processus de matching entre la requête et le moteur de recherche. On observe encore une volonté de représentativité par rapport à l'énoncé tout en diminuant le plus possible le nombre d'éléments présents dans la requête. En effet, les deux thématiques sont ici représentées. De plus, le token "inspirés" permet ici de relier ces deux dernières sans avoir à recopier la totalité du grand nombre de tokens de la tâche TMR.

(76) ***dessin animé*** *pete docter*
(77) *système d'exploitation* ***logo*** *spirale*
(78) *peintres* ***inspirés par*** *la mythologie grecque* ***dans*** *la renaissance italienne*

- **Inférences contextuelles**

Cas légèrement plus fréquents que les ajouts thématiques avec 35 occurrences, les inférences contextuelles sont les ajouts les plus couramment observés. Ces derniers correspondent à l'ajout d'éléments inférés par le participant grâce à une connaissance antérieure à la session de recherche. En effet, au fil de sa session, le moteur de recherche va fournir au participant un certain nombre d'informations en lien avec le ou les thématiques de recherche. Néanmoins, il arrive que, même avant d'interagir avec le système, le participant ait déjà en sa possession une expertise plus ou moins importante d'une thématique de recherche, ce qui va pouvoir lui permettre d'adjoindre des éléments en rapport avec celui-ci. Sur les 35 inférences contextuelles, 14 concernent l'ajout du token "The". En effet, la tâche de recherche TWD fait référence à "la série Walking Dead" du vrai nom "The Walking Dead". Ces participants visiblement familiers avec ce produit culturel ou du moins sa dénomination, ont alors adjoint le token "The" afin de possiblement optimiser le processus de matching. C'est notamment le cas dans l'exemple **(79)**. Dans l'exemple (80), on observe que la requête "*zeus europe taureau tableau*" possède un taux d'ajout de 0,75. En effet, le participant ajoute le token "*zeus*" pour faire référence au "père d'Héraclès". Il ajoute également les tokens "*europe*" et "*taureau*" pour faire référence à la peinture qui représente Zeus se transformant en taureau pour séduire la déesse Europe. Ainsi, le participant mobilise des connaissances qui lui sont antérieures à l'interaction avec le système, ce qui va lui permettre l'avantager lors de la recherche. La requête en lien avec l'exemple (81) montre une requête similaire où le participant a ici ajouté "*linux*" afin de faire référence au "système d'exploitation […] associé avec le pingouin". On peut donc imaginer que les participants qui mobilisent ce type d'ajouts vont effectuer des sessions plus courtes et possiblement effectuer plusieurs de ces ajouts au fil de leurs requêtes.

(79) *fille de hershel greene* ***the*** *walking dead*
(80) ***zeus europe taureau*** *tableau*
(81) ***linux***

Ainsi, les ajouts ne constituent qu'une part limitée des requêtes initiales, mais ils révèlent néanmoins plusieurs stratégies distinctes. Certains participants mobilisent des ajouts navigationnels afin d'accéder directement à une ressource identifiée, tandis que d'autres réorganisent l'énoncé à l'aide d'ajouts structurels ou de termes thématiques jugés plus représentatifs de leur besoin informationnel. Enfin, les inférences contextuelles témoignent de la capacité des participants à mobiliser des connaissances préalables afin d'enrichir leur requête. Globalement, ces stratégies d'ajouts permettent aux participants de se réapproprier l'énoncé en l'adaptant à leur propre représentation de la tâche de recherche.

4.3. De la requête initiale aux reformulations : Continuités et divergences

Maintenant que nous nous sommes penchés sur les caractéristiques formelles et lexicales des requêtes initiales, il nous est possible d'examiner la manière dont celles-ci s'inscrivent dans la dynamique globale de la session de recherche. Nous chercherons ainsi à déterminer dans quelle mesure les caractéristiques observées précédemment se maintiennent au fil des reformulations. L'objectif est alors d'observer et de mesurer à quel point la requête initiale se distingue du reste de la session. On peut supposer de fortes divergences entre la requête initiale et l'ensemble des reformulations. Comme l'affirme Marchionini (2006), la requête initiale correspond à une introduction exploratoire au sein d'une ou plusieurs thématiques de recherche. Ainsi, nous cherchons ici à identifier les aspects de la requête initiale qui sont conservés ou modifiés au cours des reformulations, cela d'un point de vue structurel, puis ensuite lexical. Nous pouvons alors mettre à profit le traitement exhaustif opéré sur l'ensemble des requêtes du corpus PRIVaThe. Cependant, nous ne prendrons pas en compte les 203 requêtes de la tâche de recherche TWD, ces dernières n'étant pas assez nombreuses afin d'opérer une comparaison entre la requête initiale et le reste des productions.

4.3.1. D'un point de vue structurel

Comme nous avons pu l'observer, les requêtes initiales présentent des caractéristiques structurelles relativement marquées. Leur taille varie selon la tâche de recherche et semble directement liée à celle de l'énoncé. Plus l'énoncé est court, plus les participants auront tendance à recopier des segments plus importants de celui-ci, impactant la taille des requêtes. Les participants privilégient également très largement les requêtes formulées sous forme de mots-clés, tandis que les permutations de tokens demeurent relativement rares. Les requêtes initiales apparaissent ainsi comme des productions fortement ancrées dans la structure de l'énoncé, tout en privilégiant des formulations concises et syntaxiquement peu élaborées.

En premier lieu, nous pouvons observer si la taille des requêtes initiales diffère de celle des reformulations. Les deux tâches étudiées ayant des caractéristiques formelles très distinctes, en particulier au niveau de la taille, nous avons choisi de réaliser une analyse différentielle. Les résultats nous indiquent ainsi une tendance totalement inversée pour les deux tâches. Tendances qui se rejoignent paradoxalement autour d'une logique commune. Pour rappel, les requêtes initiales découlant de la tâche TMR faisaient en moyenne 3,68 tokens. On observe alors que les reformulations pour cette même tâche ont tendance à être légèrement plus longues. Pour cette tâche, la moyenne des reformulations s'établit à 4,37 tokens par requête, avec un intervalle de confiance à 95 % compris entre 4,10 et 4,64 tokens. Au contraire, les requêtes de la tâche de recherche TSF avaient en moyenne une longueur de 6,31 tokens. Ici, on se rend compte que cette valeur est en nette baisse lors des reformulations. Les reformulations de la tâche TSF présentent ainsi une moyenne de 4,70 tokens par requête, avec un intervalle de confiance à 95 % compris entre 4,51 et 4,88 tokens. Là où pour la tâche TMR, les reformulations apparaissent comme plus longues que les requêtes initiales, le résultat est alors opposé pour la tâche TSF. La Figure 9 illustre ces deux tendances.

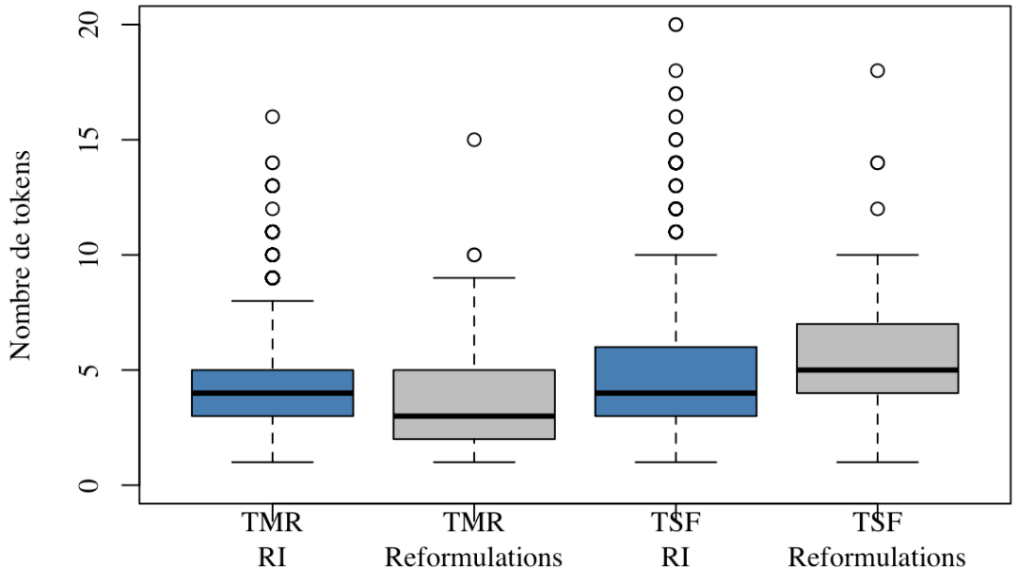

Figure 9 - Moyenne de la longueur des requêtes initiales et des reformulations selon la tâche

On observe alors que les valeurs moyennes de longueur des reformulations se stabilisent dès la deuxième requête, malgré une grande variabilité observée lors de la première requête. On remarque également que les valeurs changent immédiatement après la requête initiale pour se stabiliser autour d'une longueur moyenne 4,52 tokens pour l'ensemble des reformulations. Par exemple, les secondes requêtes découlant de la tâche TMR, font en moyenne 4,22 tokens, contre 4,62 tokens lors des troisièmes requêtes. À l'inverse, une régression progressive s'observe au sein de la tâche TSF. Pour conclure, les requêtes initiales présentent de manière attendue une variabilité bien plus importante que les reformulations, cela en lien avec la dissolution progressive du besoin d'information au fil de la session de recherche.

En second lieu, nous pouvons maintenant nous intéresser au type de requête adopté par les participants au cours de la session, en comparant la requête initiale aux requêtes qui suivent. On observe alors que la comparaison des types de requêtes entre les requêtes initiales et les reformulations ne révèle pas de différence significative. Cela concerne la tâche TMR (p-value = 0,13) ou pour la tâche TSF (p-value = 0,66). Si les proportions de requêtes formulées en langage naturel varient davantage entre les deux tâches au niveau des requêtes initiales (15 % pour TMR contre 27 % pour TSF), elles tendent à converger lors des reformulations (22 % et 25 % respectivement). Néanmoins, l'absence de significativité statistique ne permet pas de tirer de conclusions fermes à ce sujet. Ces résultats indiquent ainsi que la stratégie de formulation adoptée dès la requête initiale a globalement tendance à se maintenir de façon stable tout au long de la session.

Afin de mettre à l'épreuve cette observation, nous avons cherché à déterminer dans quelle mesure le type de requête adopté lors de la requête initiale influence la stratégie de formulation adoptée par la suite au niveau du participant. Le test du chi-deux révèle une relation significative (p-value = 3.001e-09) entre le type de requête adopté lors de la requête initiale et le type de requête majoritairement produit au cours des reformulations qui suivent. Ainsi, le choix de formulation opéré dès la première requête tend à influencer la stratégie adoptée tout au long de la session de recherche. On observe que 93 % des participants ayant formulé leur requête initiale en mots-clés produisent également une majorité de reformulations en mots-clés, cela à l'échelle de leur session. Également, 45 % des participants ayant formulé leur requête initiale en langage naturel maintiennent cette stratégie comme type majoritaire lors de leurs reformulations, malgré le faible taux d'apparition de ce type de formulation. Ce résultat vient renforcer l'idée que la requête initiale joue un rôle structurant dans la dynamique globale de la session. Le choix initial de formulation tendant à influencer le déroulé de la session, bien que la composition en mots-clés reste un choix privilégié qui se confirme lors des reformulations.

Enfin, on peut s'interroger sur l'évolution des opérations de permutation effectuées par les participants au fil de la session de recherche. Le Tableau 20 présente ainsi les valeurs moyennes du Tau de Kendall pour chaque tâche, en distinguant les requêtes initiales des reformulations. On observe des valeurs très proches entre les deux contextes, quelle que soit la tâche ou la variante observée. Les participants ne semblent donc pas modifier la manière dont ils réordonnent les termes de l'énoncé entre la requête initiale et les reformulations qui suivent. Le test de Wilcoxon confirme cette observation, ne révélant pas de différence significative entre les deux groupes, avec une p-value égale à 0,655. Ainsi, les faibles valeurs du Tau de Kendall présentées dans le Tableau 20 confirment que les opérations de permutation restent rares tout au long de la session, sans distinction significative entre la requête initiale et les reformulations. Cette stabilité peut s'expliquer par le fait que l'énoncé, constamment visible par le participant tout au long de la session, fixe durablement la représentation que ce dernier se fait du besoin d'information associé. Ainsi, la hiérarchisation des éléments de l'énoncé opérée lors de la requête initiale tend à se maintenir lors des reformulations, le participant continuant à juger les mêmes éléments comme prioritaires quant à la résolution de son besoin d'information.

Tâche	Requêtes initiales	Reformulations
TMR Variante 1	0,26	0,22
TMR Variante 2	0,04	0,05
TSF Variante 1	0,18	0,19
TSF Variante 2	0,05	0,06

Tableau 20 - Pourcentage du Tau de Kendall moyen entre requêtes initiales et reformulations selon la tâche

4.3.2. D'un point de vue lexical

Nous pouvons maintenant nous pencher sur l'évolution des caractéristiques lexicales des requêtes au fil de la session de recherche, en cherchant à déterminer dans quelle mesure les stratégies de recopie, de modification et d'ajout observées lors des requêtes initiales se maintiennent ou évoluent lors des reformulations. Pour rappel, nous avions observé que, lors de la formulation de la requête initiale, les participants se détachent rarement de l'énoncé en opérant presque uniquement des opérations de recopie. Les modifications demeurent particulièrement marginales, tandis que les ajouts, bien que plus fréquents, concernent un nombre limité de tokens. Les requêtes initiales apparaissent ainsi comme des productions fortement dépendantes du contenu de l'énoncé, dont elles reprennent l'essentiel des éléments lexicaux. Le traitement exhaustif des données nous permet ainsi d'évaluer si la requête initiale occupe une position particulière à cet égard, où si elle ne diffère pas de potentielles reformulations.

On observe immédiatement que le taux de recopie moyen observable au sein des reformulations est nettement plus faible que celui attribué aux requêtes initiales. En effet, là où ces dernières étaient composées à presque 90 % de recopies en moyenne, les reformulations obtiennent un taux de recopie moyen égal à 0,49 avec un intervalle de confiance situé entre 0,48 et 0,50. Le taux de modification moyen reste quant à lui totalement inchangé avec une valeur commune de 0,01 entre les requêtes initiales et le reste de la session. Au contraire, le taux d'ajout moyen augmente fortement passant de 0,09 à 0,49 avec un intervalle de confiance similaire à celui du taux de recopie. La composition lexicale des requêtes initiales semble donc fortement différer de celles des reformulations. Comme l'illustre la Figure 10, on observe une tendance progressive d'inversement des valeurs de recopie et d'ajout au fur et à mesure de la session de recherche. La requête initiale, correspondant au premier contact entre le participant et le besoin informationnel, incite le participant à interagir de manière plus directe avec l'énoncé en recopiant des éléments de celui-ci. Au contraire, au fur et à mesure des interactions avec le système, qui va progressivement fournir des éléments de résolution quant au besoin d'information, le participant va alors avoir la possibilité d'adjoindre des tokens qu'il aura rencontrés sur les pages de résultats. Le taux de modification restant encore extrêmement faible, on observe que quand un des deux autres taux baisse, l'autre augmente. Le coefficient de corrélation de Spearman confirme une corrélation importante entre le taux de recopie et le taux d'ajout à l'échelle de la session (rho = 0,983).

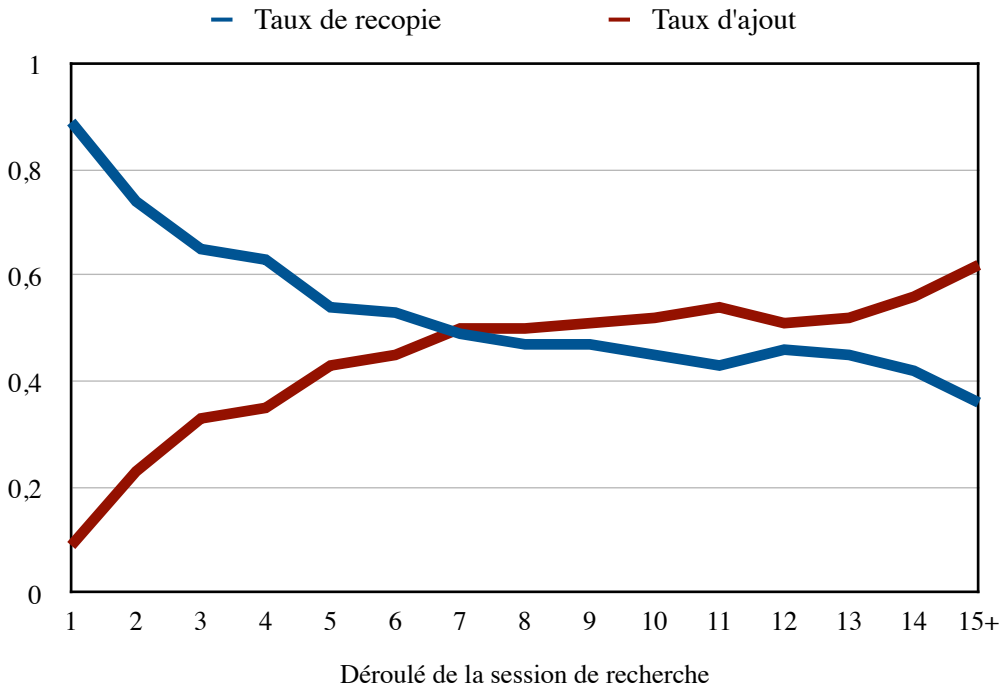

Figure 10 - Évolution du taux moyen de recopies et d'ajouts des tâches TMR et TSF au fil de la session de recherche

4.4. Vers un continuum entre représentativité de l'énoncé et économie de formulation

Un constat peut être établi sur la base de l'ensemble des observations effectuées jusqu'à maintenant. Les participants, à chacune de leur requête, semblent se positionner au sein d'un continuum représentant la manière dont chacun d'entre eux interagit avec l'énoncé de la tâche de recherche. Ainsi, chaque requête peut être située à une distance plus ou moins éloignée des deux pôles qui constituent ce continuum. Ces derniers représentent les deux options qui s'offrent à l'utilisateur. Le participant, voulant s'assurer que le système identifie avec précision son besoin d'information, peut alors décider de représenter avec exhaustivité ce dernier. On nommera ce pôle Représentativité. En faisant cela, il sacrifie une certaine économie de formulation (EDF). Plus il choisit de représenter exhaustivement le besoin d'information, plus il devra effectuer un nombre de gestes de saisie importants, ce qui augmentera le temps nécessaire à la formulation de la requête. Au contraire, il peut pencher vers une approche tournée vers l'économie, nom du pôle opposé à la représentativité. Cette approche le conduit à s'affranchir de l'organisation structurelle et lexicale qui existe au sein de l'énoncé de la tâche de recherche. Cela lui permet alors de gagner du temps et de réduire l'effort qu'il doit effectuer afin de transmettre son besoin au système.

Comme le soulignent Sanderson & Croft (2012), les systèmes de recherche ont constamment évolué avec le temps. Cette évolution a progressivement conduit au développement de systèmes davantage capables d'interpréter des requêtes concises et efficaces, réduisant ainsi la nécessité pour l'utilisateur de représenter exhaustivement son besoin d'information. Ainsi, les comportements des utilisateurs ont nécessairement muté en ce sens. Les moteurs de recherche sont aujourd'hui associés à une philosophie de réduction des gestes de saisies conduisant les utilisateurs à minimiser leurs efforts lors de la formulation de leurs requêtes. En théorie, les requêtes présentent alors globalement des caractéristiques les amenant plus proches du pôle EDF. Néanmoins, les observations que nous avons pus effectuer nous amènent, une nouvelle fois, à considérer la requête initiale comme une entité à part entière. Avant de rentrer en détail au sein de cette perspective analytique, il apparaît comme important de recontextualiser les différentes analyses ayant pu être effectuées du point de vue de ce continuum. La Figure 11 correspond ainsi à une représentation concrète de ce dernier, les numéros entre crochets correspondent aux exemples qui seront cités au sein de cette section. Ces derniers sont ainsi situés en relation avec leur positionnement au sein du continuum.

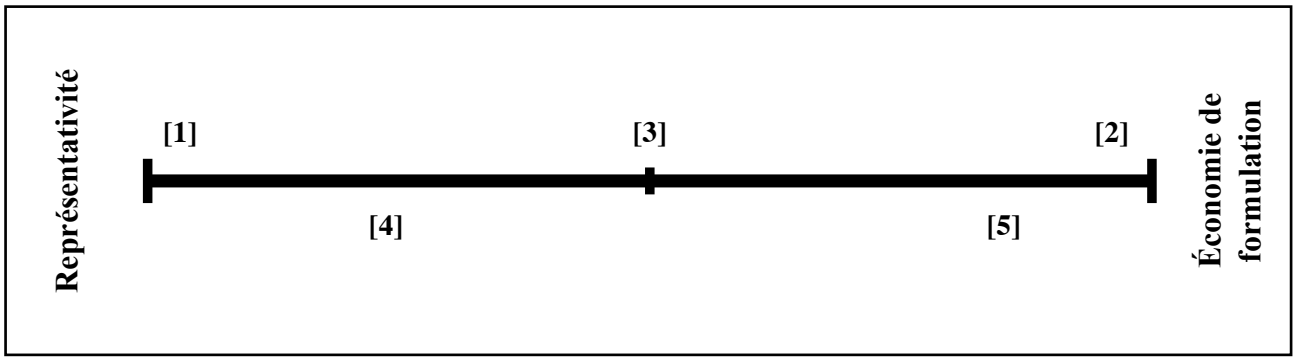

Figure 11 - Représentation du continuum entre représentativité et économie de formulation

Plusieurs caractéristiques formelles et lexicales peuvent être associées à une philosophie représentative. En premier lieu, une requête comportant un fort taux de recopie aura nécessairement tendance à s'orienter vers le pôle Représentativité. En effet, plus la proportion de la requête correspond à un ensemble d'opérations de recopies, moins cette dernière s'éloigne de l'énoncé de la tâche. De plus, une requête longue de taille comportant un grand nombre de ces recopies aura d'autant plus tendance à s'inscrire dans une démarche de représentativité. L'exemple **[1]** contient ainsi une requête s'inscrivant à l'extrémité du pôle Représentativité, comme cela est visible dans la Figure 11.

[1] *quel est le lien entre la version 1.2 du système d'exploitation représenté par une spirale associé à celui avec le pingouin et le long métrage d'animation réalisé par Pete Docter où Henri Guybet prête sa voix ?*

On observe que le participant a ici formulé en requête en langage naturel une longue de 39 tokens, chacun correspondant à des opérations de recopie. De plus, on remarque que cette production présente un pourcentage de suppression de seulement 2 %. On se trouve alors dans un cas extrême de représentativité. On remarque logiquement que le participant n'a effectué aucune action de permutation et que sa requête ne contient aucune opération de modification ou d'ajout. On considère ainsi qu'une requête contenant une des deux types de stratégies lexicales s'éloignera d'une démarche de représentativité de l'énoncé. Néanmoins, une opération d'ajout apparaît comme plus radicale qu'une stratégie de modification. L'exemple **[2]** correspond à une requête se trouvant à l'extrémité du pôle EDF. On aperçoit que cette dernière n'est composée que d'un unique token, à savoir un ajout navigationnel. Elle apparaît alors comme diamétralement opposée à la requête **[1]**. Aucune composante de la tâche TWD n'est ici représentée dans la requête. Le participant a donc choisi de s'éloigner totalement de l'énoncé en ne composant sa requête qu'à partir d'un ajout, référence au site "Wikipédia". Cette approche permet à l'utilisateur de s'affranchir entièrement de l'énoncé de la tâche et de construire une requête autonome, fondée uniquement sur une intention de recherche externe, optimisant ainsi le gain de temps et réduisant le nombre de mouvements de saisie au détriment de toute représentativité du besoin initial.

[2] *wikipedia*

Nous pouvons ainsi relier les opérations de recopies à une idée de représentativité. Au contraire, les stratégies d'ajouts peuvent être associées à une philosophie d'économie de gestes de saisies. On peut identifier les stratégies de modifications, en particulier les transformations morphologiques, comme un entre deux entre ces différentes approches. De plus, la présence d'opérations de permutations peut également indiquer une intention d'économie, tout comme la composition en mots-clés. Néanmoins, comme nous avons pu l'observer, une requête en langage naturel peut elle aussi indiquer une intention d'économie. C'est par exemple le cas dans la requête **[3]** qui a déjà pu être observée et qui contient plusieurs mots-outils ajoutés par le participant. Ces ajouts structurels permettent de réorganiser l'énoncé particulièrement dense de la tâche TMR. L'ajout du token "inspirés" joue également le même rôle et permet d'encapsuler la densité sémantiques propre au besoin d'information de cet énoncé. On observe également une permutation du token "peintre" qui est ici également modifié grâce à une transformation morphologique. Cette requête obtient néanmoins un taux de recopie de 0,60. On remarque en effet qu'il a adjoint les tokens correspondant aux deux thématiques de recherche propres à la tâche. On considère alors que cette requête se situe au centre du continuum, le participant faisant autant de compromis en faveur d'une représentativité de l'énoncé que d'une EDF dans la formulation de sa requête. L'exemple **[4]** correspond à une requête également formulée en langage naturel mais cette fois-ci avec un taux de recopie de 0,92 % à cause du seul ajout inférentiel du token "the". On peut alors affirmer que cette requête est moins représentative que l'exemple [1] mais également moins efficace que l'exemple **[3]**. L'exemple **[5]** est quant à lui assez similaire au cas de la requête **[2]**. Seulement, il est légèrement plus représentatif grâce à la recopie des deux groupes de tokens "Henri Guybet" et "Pete Docter", qui sont tous les deux intervertis. On peut donc affirmer que cette requête s'inscrit dans une démarche d'économie tout en comportant quelques caractéristiques propres à une production représentative, à savoir plusieurs recopies.

[3] *peintres inspirés par la mythologie grecque dans la renaissance italienne*
[4] *nom de la deuxième fille de Hershel Greene dans the Walking dead*
[5] *henri guybet et pete docter*

L'ensemble des exemples cités dans cette section provenant de requêtes initiales, nous pouvons encore une fois affirmer que ces dernières adoptent des formes particulièrement variées. Néanmoins, comme nous avons pu l'observer, la requête initiale aura le plus souvent tendance à être composée d'un ensemble de recopies, souvent organisées en mots-clés. Or, le nombre de recopies baisse drastiquement au fil de la session, le participant nourrissant ses requêtes d'ajouts permis par les interactions successives avec le système. De ce fait, il s'éloigne de l'énoncé sous sa forme initiale et optimise ses requêtes jusqu'à ce que le système réponde au besoin dans son intégralité. La requête initiale semble donc privilégier une démarche de représentativité du besoin d'information. Au fil des reformulations, à l'inverse, le nombre d'ajouts augmente et les requêtes tendent à être plus courtes. On peut donc supposer que les requêtes initiales relèvent davantage d'une démarche représentative, tandis que les reformulations s'inscrivent plutôt dans une logique d'économie, même si l'écart qui les sépare sur le continuum reste limité.

Conclusion

Ce mémoire nous a ainsi permis de concevoir une automatisation du traitement de requêtes formulées sur un moteur de recherche. Cette automatisation a alors permis une analyse structurelle et lexicale d'un ensemble de productions découlant d'un contexte expérimental. Ces actions ont finalement rendu possibles l'obtention de plusieurs informations concernant la requête initiale en RI traditionnelle. Nous avons donc pu répondre à nos deux problématiques de recherche initiales.

En premier lieu, nous nous interrogions sur les multiples caractéristiques observables au sein des requêtes initiales. Les observations effectuées suite au traitement nous démontrent que la forme que prend l'énoncé va immédiatement influencer le comportement de l'utilisateur. En effet, plus l'énoncé de la tâche de recherche est dense, plus ce dernier aura tendance à soumettre des requêtes courtes de taille. Un énoncé trop long aura tendance à dissuader le participant d'en recopier une taille importante, impactant ainsi la taille de sa première requête. Globalement, on observe également que les résultats observés divergent, sur cet aspect, de ceux d'Aula (2003). Cette dernière affirme que les requêtes initiales sont en moyenne longues d'environ 3 termes, un chiffre bien plus faible que le nôtre. On peut attribuer ceci au fait que les participants du corpus PRIVaThe possédaient déjà un besoin d'information en partie stabilisé grâce à l'énoncé en langage naturel fourni. On peut donc théoriser que ce facteur a tendance à augmenter la taille des requêtes initiales. Également, les SRI ont évolué depuis, en accord avec le comportement des utilisateurs.

Néanmoins, on observe que la grande majorité des productions initiales sont composées à l'aide d'une concaténation de mots-clés, en accord avec ce qu'avait pu observer Broder (2002) ou Aula (2003). De plus, les résultats nous indiquent que les participants tendent, la plupart du temps, à maintenir l'ordre canonique des éléments de l'énoncé au sein de leurs requêtes initiales.

D'un point de vue lexical, nous avons pus observer que les requêtes initiales ne sont constituées quasiment que de recopies. En effet, presque 90 % des tokens traités correspondent directement à un token de l'énoncé de la tâche. Cependant, en accord avec la forte proportion de requêtes en mots-clés, les participants tendent très rarement à recopier des segments longs de l'énoncé, en particulier lorsque la tâche est très longue. À l'inverse, les opérations de modifications sont rarissimes. Nous avons pu remarquer que les types de modifications distingués différaient largement l'un de l'autre. En ce sens, la majorité des modifications effectuées correspondent à des transformations morphologiques qui visent à modifier le sens d'un token de l'énoncé afin d'optimiser la potentielle performance de la requête initiale. Enfin, les stratégies d'ajouts apparaissent légèrement plus fréquentes et concernent majoritairement des cas d'inférences qui sous-entendent que le participant possède déjà une expertise plus ou moins importante concernant une ou plusieurs des thématiques de recherche contenues dans l'énoncé. On observe également l'apparition d'ajouts dits thématiques qui visent à optimiser certains éléments sémantiques de l'énoncé via des reformulions diverses.

La requête initiale apparaît alors comme porteuse d'un grand nombre de variations tout en présentant des tendances clairement identifiables. À savoir un taux de recopie très élevé, un faible nombre d'opérations de permutations mais également une tendance non négligeable à la formulation en mots-clés. Nous avons alors pu nous pencher sur notre seconde problématique. Celle-ci concernait alors les différentes continuités et ruptures entre les requêtes initiales et les reformulations du jeu de données PRIVaThe. L'application du traitement automatisé sur l'ensemble des données nous a ainsi permis de vérifier notre hypothèse initiale. Pour rappel, nous avions théorisé que les caractéristiques observables au sein des reformulations se distingueraient de celles caractérisant les requêtes initiales.

Cette hypothèse s'est rapidement vérifiée sur certains aspects. En effet, nos observations ont rapidement mis en avant la restructuration lexicale opérée au fur et à mesure de la session. Plus la session de recherche se poursuit, plus le taux de recopie diminue, faisant place à une proportion proportionnellement plus importante d'ajouts. De plus, d'un point de vue formel, nous avons pu remarquer une stabilisation progressive de la taille des requêtes dont la longueur tend à se stabiliser après quelques reformulations. Ainsi, l'impact de la taille de l'énoncé sur la taille des requêtes semble être rendu inexistant à partir d'un certain seuil. Cela démontre que la requête initiale tend à afficher plus de variations que les reformulations.

Nous avons également pu identifier plusieurs continuités entre ces deux entités. Le type de formulation des requêtes ne semble pas être impacté par le déroulé de la session de recherche. C'est également le cas pour les différentes opérations de permutations de tokens ou de groupes de tokens. Globalement, notre hypothèse apparaît donc comme vérifiée. La proximité entre le besoin d'information et la première requête formulée par l'utilisateur conduit celui-ci à adopter un comportement différent que celui qu'il affichera lors de ces reformulations. Cet effet semble se décupler au fil des reformulations.

Nous avons ensuite proposé une représentation sous la forme d'un continuum afin de situer les différents archétypes de requêtes ayant pu être observées. Ce continuum représente ainsi entre ces deux pôles, les stratégies variables mises en place par les participants. Là où certains adoptent une démarche de représentativité de l'énoncé, d'autres choisissent de sacrifier une part plus importante des informations présentes dans la tâche afin d'opter pour une stratégie tournée vers une certaine économie. On s'aperçoit alors que les requêtes initiales s'inscrivent de manière plus marquée au sein d'une démarche de représentativité. L'énoncé, comme nous l'avons mentionné, est plus souvent et largement représenté en première requête. Au contraire, les reformulations affichent plus souvent des caractéristiques propres à une philosophie d'économie des gestes de saisies en réalisant notamment un nombre plus conséquent d'ajouts.

Plusieurs continuités de recherche peuvent être envisagées. Premièrement, il serait intéressant d'optimiser le traitement en accord avec les limites identifiées en *Section 3.3.2*. Ce dernier pourrait alors être appliqué à un ensemble plus important de requêtes. Il serait pertinent de vérifier l'impact de la tâche en recueillant des données provenant de plus d'énoncés différents. Plus globalement, cette approche pourrait renforcer la véracité des résultats obtenus. Néanmoins, il aurait également été intéressant de mettre à profit les verbalisations effectuées par certains des participants du corpus PRIVaThe. Ces dernières auraient pu nous renseigner avec plus de certitude sur les intentions ayant conduit les participants à réaliser certaines actions. Deuxièmement, le continuum présenté en *Section 4.4* pourrait gagner en fiabilité. On pourrait par exemple trouver un moyen précis et moins arbitraire de mesurer clairement la position de chaque requête sur ce dernier. Cela nous renseignerait alors de manière plus précise sur les comportements globaux des utilisateurs.

Bibliographie

- Abu Onq, N. (2023). *How do human and contextual factors affect the way people formulate queries?* [Doctoral dissertation, RMIT University].

- Abu Onq, N., Sanderson, M., & Scholer, F. (2025). Classifying term variants in query formulation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)* (pp. 2395–2406). Association for Computing Machinery.

- Alaofi, M., Gallagher, L., Mckay, D., Saling, L. L., Sanderson, M., Scholer, F., Spina, D. et White, R. W. (2022). Where Do Queries Come From? *In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 2850–2862), New York, NY, USA. Association for Computing Machinery.

- Aula, A. (2003). Query formulation in web information search. In P. Isaías & N. Karmakar (Eds.), *Proceedings of the IADIS International Conference* (pp. 403–410).

- Broder, A. (2002). A Taxonomy of Web Search. *SIGIR Forum*, 36(2), 3-10.

- Dosso, C., Moreno, J. G., Chevalier, A., & Tamine, L. (2021). Cost: An annotated data collection for complex search. In *Proceedings of the 30th ACM international conference on information & knowledge management* (pp. 4455–4464). Association for Computing Machinery.

- Hearst, M. (2009) *Search User Interfaces*, Cambridge University Press.

- Hiemstra, D. (2009). Information Retrieval Models, 1–19. Wiley.

- Huang, J., & Efthimiadis, E. N. (2009). Analyzing and evaluating query reformulation strategies in Web search logs. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 77-86). ACM.

- Ibarboure, C. (2025). *Typologies des parcours de recherche d'information sur le Web : étude des variations thématiques dans les sessions complexes*. Doctorat en Sciences du langage, Université de Toulouse Jean Jaurès.

- Latour, M. (2014). *Du besoin d'informations à la formulation des requêtes : étude des usages de différents types d'utilisateurs visant l'amélioration d'un système de recherche d'informations*. Doctorat en Sciences de l'Information et de la Communication, Université de Grenoble.

- Marchionini, G. (2006). Exploratory search: from finding to understanding. *Communications of the ACM, 49*(4), 41–46.

- Mitra, B. (2015). Exploring Session Context Using Distributed Representations of Queries and Reformulations. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 3–12), New York, NY, USA. Association for Computing Machinery.

- Rose, D., & Levinson, D. (2004). Understanding user goals in web search. *Proceedings of the 13th International Conference on World Wide Web (WWW '04)* (pp. 13–19). ACM.

- Sanderson, M., & Croft, W. B. (2012). The history of information retrieval research. *Proceedings of the IEEE, 100*(Special Centennial Issue), 1444–1451.

- Sharit, J., Hernández, M. A., Czaja, S. J. et Pirolli, P. (2008). Investigating the roles of knowledge and cognitive abilities in older adult information seeking on the web. *ACM Trans. Comput.-Hum. Interact., 15*(1).

- Simonnot, B. (2006). *Le besoin d'information : principes et compétences* [Présentation PowerPoint]. Actes de la journée thémat'IC : Information, besoins et usages, Strasbourg-Illkirch.

- Strohmaier, M., Prettenhofer, P., & Lux, M. (2008). Different degrees of explicitness in intentional artifacts: Studying user goals in a large search query log. In *CSKGOI'08 International Workshop on Commonsense Knowledge and Goal Oriented Interfaces, in conjunction with IUI'08*.

- Tamine, L. & Calabretto, S. (2008). *Recherche d'information contextuelle et Web*. In Savoy, J. et Boughanem, M., éditeurs : Recherche d'information : état des lieux et perspectives, Recherche d'information et web, (pp. 201–224). Hermes Lavoisier.

- Wilson, T. (1981). On user studies and information needs. *Journal of Documentation*, 37(1), 3–15.

Annexes

Annexe 1 - Tables de fréquences des lemmes dans les requêtes initiales du corpus

Lemme	Fréquence absolue	Présence dans l'énoncé
pete	*57*	Présent
docter	*56*	Présent
exploitation	*36*	Présent
système	*36*	Présent
métrage	*35*	Présent
long	*35*	Présent
guybet	*31*	Présent
henri	*30*	Présent
spirale	*28*	Présent
par	*27*	Présent
animation	*23*	Présent
le	*17*	Présent
réaliser	*14*	Présent
représenter	*13*	Présent
un	*13*	Présent
version	*11*	Présent
de	*9*	Présent
pingouin	*8*	Présent
logo	*7*	Absent
linux	*7*	Absent
avec	*6*	Absent
voix	*6*	Présent
lien	*5*	Présent
entre	*5*	Présent
ou	*5*	Présent
sa	*5*	Présent
prêter	*5*	Présent
film	*4*	Absent
et	*4*	Présent
à	*4*	Présent
quel	*4*	Présent
être	*4*	Présent
associer	*4*	Présent
celui	*4*	Présent
movie	*2*	Absent
dessin	*1*	Absent
animer	*1*	Absent
différent	*1*	Absent
avoir	*1*	Absent
with	*1*	Absent
ordinateur	*1*	Absent
realisateur	*1*	Absent
os	*1*	Absent

TSF

Tableau 21 - Table de fréquence des lemmes dans les requêtes initiales de la tâche TSF

Lemme	Fréquence absolue	Présence dans l'énoncé
hershel	*86*	Présent
dead	*82*	Présent
walking	*82*	Présent
greene	*81*	Présent
fille	*57*	Présent
de	*45*	Présent
le	*29*	Présent
deuxième	*28*	Présent
nom	*26*	Présent
dans	*15*	Présent
the	*14*	Absent
daughter	*8*	Absent
est	*8*	Présent
série	*8*	Présent
quel	*7*	Présent
personnage	*2*	Absent
wiki	*2*	Absent
casting	*2*	Absent
name	*1*	Absent
family	*1*	Absent
prenom	*1*	Absent
quelle	*1*	Absent
généalogie	*1*	Absent
personne	*1*	Absent
wikipedia	*1*	Absent
?	*1*	Présent
ma	*1*	Absent

Tableau 22 - Table de fréquence des lemmes dans les requêtes initiales de la tâche TWD

Lemme	Fréquence absolue	Présence dans l'énoncé
héraclès	*79*	Présent
père	*55*	Présent
transformation	*25*	Présent
animal	*22*	Présent
le	*15*	Présent
peinture	*14*	Présent
renaissance	*12*	Présent
zeus	*12*	Absent
mythologie	*11*	Présent
italienne	*10*	Présent
grecque	*9*	Présent
seduction	*7*	Présent
femme	*6*	Présent
en	*6*	Présent
tableau	*5*	Présent
de	*5*	Présent
séduire	*5*	Présent
être	*4*	Présent
pour	*3*	Présent
qui	*3*	Absent
transformer	*3*	Absent
quel	*2*	Absent
taureau	*2*	Absent
painting	*2*	Absent
représenter	*1*	Absent
et	*1*	Présent
of	*1*	Absent
in	*1*	Absent
oeuvre	*1*	Présent
mortel	*1*	Présent
un	*1*	Présent
europe	*1*	Absent
peintre	*1*	Présent
inspirer	*1*	Absent
par	*1*	Présent
dans	*1*	Présent
3	*1*	Absent

Tableau 23 - Table de fréquence des lemmes dans les requêtes initiales de la tâche TMR

Annexe 2 - Description du format de sorties des données traitées

La Figure 12 correspond au traitement d'une requête initiale découlant de la tâche TSF. Comme mentionné en Section 2.3, le système produit une sortie au format JSON. Les résultats du traitement, contenus alors dans trois fichiers JSON différents, soit un par tâche, sont systématiquement agencés de la même manière.


```
"idParticipant": "NV9772",
"Contexte": "1",
"idTache": "TSF",
"variante": "1",
"requete": "logo systèmes exploitation",
"Dimensions": {
  "Dimension structurelle": {
    "Nombre de tokens": 3,
    "Type de requête": "mots-clés",
    "Kendall Tau": 0.0,
    "Pourcentage de suppressions": 95,
    "Damerau Levenshtein": 189
  },
  "Dimension lexicale": {
    "Nombre de recopies": 1,
    "Taux de recopies": 0.33,
    "Nombre de modifications": 1,
    "Taux de modifications": 0.33,
    "Nombre d'ajouts": 1,
    "Taux d'ajouts": 0.33,
    "detail": [
      {
        "token": "logo",
        "position": 1,
        "statut": "ajout",
        "correspondance": null
      },
      {
        "token": "systèmes",
        "position": 2,
        "statut": "modification : transformation morphologique",
        "correspondance": "système"
      },
      {
        "token": "exploitation",
        "position": 3,
        "statut": "recopie",
        "correspondance": "exploitation"
      }
```


Figure 12 - Exemple d'une sortie au format JSON pour la requête "logo système exploitation"

En premier lieu, le système indique plusieurs métadonnées déjà présentes dans le fichier texte d'entrée. "idParticipant" correspond à l'identifiant du participant. Cela permet de lier les productions d'un participant entre les tâches. "Contexte" indique le contexte de la requête au sein de la session, c'est-à-dire le moment où elle est soumise. Ici, le contexte est de 1, il s'agit donc d'une requête initiale. Lors de l'analyse des résultats, c'est cette métadonnée qui a permis de systématiquement différencier les requêtes initiales des autres productions. Est ensuite indiquée la "Variante" de la tâche, seulement pour les tâches de recherche TMR et TSF, comme cela est mentionné en Section 2.1. Enfin, la balise "requête" contient la production brute du participant, celle qui est concernée par le traitement.

En second lieu, nous retrouvons les différentes caractéristiques structurelles observées, toutes décrites dans la Section 3.2. En premier lieu, la balise "Nombre de tokens" correspond au compte des opérations lexicales, c'est-à-dire la taille de la requête en tokens. Ici, nous verrons que la requête contient une recopie, une modification et un ajout, donc 3 tokens. Est ensuite indiqué le "type de requête" qui correspond à la distinction mots-clés/langage naturel. La balise "Kendall Tau" contient ensuite le Tau de Kendall calculé pour la requête en question. Le calcul précis est explicité en Section 3.2.3. À noter que si le calcul ne peut pas etre effectué, le système renvoie alors la valeur "Null". Le "Pourcentage de suppressions" est ensuite indiqué, suivi par une distance d'édition qui mesure la distance entre la requête et l'énoncé. Cette dernière mesure n'a finalement pas été mise à profit au cours de l'analyse des résultats.

Enfin, nous retrouvons les éléments relatifs au traitement lexical des requêtes tels que décris au sein de la Section 3.1. Sont alors successivement indiqué le nombre et le taux de chacune des stratégies lexicales considérées. Ces balises suivent alors l'ordre : recopie ; modification ; ajout. La balise "détail" contient alors une analyse individuelle de chacun des tokens de la requête. Le token brut est d'abord indiqué. Ces derniers sont ordonnées selon leur ordre d'apparition au sein de la production du participant. Cela peut etre vérifié par le biais de la balise "position" qui indique le positionnement du token dans la requête. Par exemple, ici, le token "logo" correspond au premier token soumis par le participant, il obtient donc une position de 1. Nous retrouvons ensuite le "statut" de chaque token, c'est-à-dire l'étiquette lexicale qui lui a été attribué. La balise "correspondance" permet d'indiquer, dans les cas de recopie et de modification, quel token de l'énoncé est concerné par l'opération. Dans le cas d'un ajout, comme pour le token "logo", la correspondance renverra une valeur "Null" car le token n'a aucune correspondance dans l'énoncé.

Déclaration sur l'honneur de non-plagiat

(à joindre au mémoire à la fin du document)

Je soussigné.e,

Nom, Prénom : **Breton Hugo**
Régulièrement inscrit.e à l'Université de Toulouse II Jean Jaurès
N° étudiant : **22203413**

Année universitaire : **2025-2026**

certifie que le document joint à la présente déclaration est un travail original, que je n'ai ni recopié ni utilisé des idées ou des formulations tirées d'un ouvrage, article ou mémoire, en version imprimée ou électronique, sans mentionner précisément leur origine et que les citations intégrales sont signalées entre guillemets.

Fait à : **Castelnau-d'Estretefonds**

Le : **13/06/2026**

Signature :