UNIVERSITÉ DU QUÉBEC À MONTRÉAL

# IMPACT DES VARIATIONS MORPHOLOGIQUES SUR LA RECHERCHE D'INFORMATION SUR LE WEB

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN LINGUISTIQUE

PAR

SAID EDDAMOUN

MARS 2009

# REMERCIEMENTS

La réalisation d'un travail de recherche est une démarche qui requiert le concours de nombreuses personnes. À ce sujet, je tiens à remercier tous ceux et toutes celles qui ont contribué, de près ou de loin, à l'accomplissement de cette recherche. Mes remerciements s'adressent plus particulièrement à :

- Ma directrice de mémoire, Mme Louisette Emirkanian, professeure au département de linguistique et didactique des langues à l'UQÀM, pour la qualité de son encadrement, sa disponibilité, ses précieux commentaires et sa rigueur scientifique.
- Mme Claire Lefebvre et M. Emmanuel Chieze pour avoir accepté d'évaluer ce travail de recherche.
- M. Bertrand Fournier pour sa précieuse aide concernant l'analyse statistique des données.
- Ma famille, de l'autre côté de l'Atlantique, pour son soutien moral.
- Rachel Therrien pour sa présence et sa compréhension.

## TABLE DES MATIÈRES

## LISTE DES TABLEAUX

## LISTE DES FIGURES

## LISTE DES ABRÉVIATIONS ET ACRONYMES

| | |
|---|---|
| **BI** | **Besoin en Information** |
| **RI** | **Recherche d'Information / Repérage d'Information** |
| **SRI** | **Système de Recherche d'Information** |
| **SAS** | **Statistical Analysis System** |
| **SCAD** | **Service de Consultation en Analyse de Données** |
| **TAL** | **Traitement Automatique des Langues** |
| **TREC** | **Text REtrieval Conference** |
| **URL** | **Uniform Resource Locator** |

# RÉSUMÉ

Notre travail de recherche est de type exploratoire. Il traite de l'apport des connaissances linguistiques à la recherche d'information sur le Web. Plus spécifiquement, nous avons étudié l'impact des variations morphologiques, notamment les variantes dérivées, en termes de fréquence, sur la pertinence des documents rapportés. À ce sujet, nous avons vérifié s'il y a une corrélation entre la fréquence des termes et des variantes morphologiques extraits des documents rapportés et la pertinence de ces mêmes documents. Les résultats obtenus n'ont pas permis de confirmer, d'une façon évidente, cette corrélation. En d'autres termes, si les données brutes laissent croire que, globalement, il y a une corrélation entre la fréquence des variables et la pertinence des documents, ce n'est pas le cas après l'examen des requêtes d'une façon individuelle, et, aussi, après l'application du test statistique de Jonckheere-Terpstra. En somme, la présence ou non d'une telle corrélation dépend, en partie, de la requête, des mots de la requête, de la nature et de la qualité des variantes.

Mots-clés : recherche d'information, connaissances linguistiques, variations morphologiques, reformulation de requêtes, traitement automatique des langues, Web.

# INTRODUCTION

Internet est un outil qui contient une quantité gigantesque de données numériques : texte, audio, vidéo et image. C'est une source d'information en évolution constante. L'explosion des données sur le Web complique la tâche de l'utilisateur (internaute), car l'accès à l'information souhaitée, dans un environnement ouvert et non contrôlé, devient de plus en plus difficile. En ce sens, même si les informations recherchées par l'utilisateur sont présentes sur le Web, leur consultation est une tâche ardue, voire impossible. Cette situation est due notamment à la nature structurelle du Web où les informations disponibles sur le réseau croissent d'une façon exponentielle. Aussi, la variété des contenus des pages Web, leurs formats et la diversité des langues présentes sur la "toile" sont des éléments qui rendent l'accès à l'information plus ardu.

L'évolution du Web a été accompagnée par le développement de plusieurs systèmes de recherche d'information (SRI). Ceux-ci ont pour objectif d'aider l'utilisateur à accéder à l'information qu'il recherche ; la rapidité dans l'exécution et la pertinence de l'information sont deux paramètres importants dans ce processus. Si le besoin d'information est plus ou moins satisfaisant dans le contexte d'un SRI classique (domaine fermé), ce n'est pas le cas dans le cadre de la recherche d'information sur le Web (domaine ouvert). En effet, la tâche d'un SRI traditionnel consiste à représenter, stocker et organiser d'une façon automatique les informations pour en faciliter l'accès. Il est question d'un système dit dédié, c'est-à-dire qu'il est conçu pour traiter d'un domaine spécifique, voire d'une thématique particulière. De même, la base de données interrogée par ces systèmes est généralement fixe et contient des collections de petite taille. Dans le contexte de la RI sur le Web, les moteurs de recherche œuvrent dans une base de données ouverte et où les informations sont en évolution constante. La tâche principale de ces moteurs consiste notamment à pouvoir trier les documents, choisir les plus pertinents et les rendre accessibles et consultables en réponse au besoin d'information de l'utilisateur.

En réponse à une requête de l'utilisateur qui est exprimée généralement par des mots-clés, le moteur de recherche rapatrie une quantité considérable de documents. Ceux-ci sont considérés comme pertinents par le système dans la mesure où ils contiennent les

mots de la requête exécutée. Toutefois, la présence des termes de la requête dans les documents retournés n'est pas toujours synonyme de pertinence. L'ambiguïté caractérisant la langue (synonymie, polysémie, homonymie, etc.) rend la tâche du tri des documents pertinents souvent aléatoire ; l'une des raisons de la piètre qualité des résultats retournés est liée au fait que les termes de la requête formulée par l'utilisateur sont traités par le système de recherche comme une chaîne de caractères et non pas comme des unités linguistiques. En d'autres termes, la prise en compte des connaissances syntaxiques, morphologiques et sémantiques est insuffisante, voire rare dans le processus de recherche d'information.

Un autre aspect, qui est d'ordre ergonomique, pose un réel problème au niveau de l'interaction entre le système et l'utilisateur. Il est question ici de la lisibilité et consultabilité des résultats rapportés par un moteur de recherche. Ces résultats sont volumineux, à l'image du Web, ils sont présentés à l'utilisateur sous forme de pages Web et chaque page contient plusieurs documents ; comme il est communément admis, un internaute qui cherche un accès rapide à l'information ne va pas aller, dans sa consultation des résultats, au-delà de la troisième page ; alors que les pages non consultées pourraient bien contenir des documents pertinents qui répondraient à ses besoins informationnels. Cependant, la qualité ergonomique des moteurs de recherche ne fait pas partie de nos objectifs dans la présente étude

En effet, notre intérêt porte particulièrement sur les systèmes traitant des données textuelles. Ces systèmes sont confrontés à des difficultés notamment d'ordre linguistique (nature des langues et leur complexité). La piètre qualité de plusieurs systèmes de recherche d'information est liée, entre autres, au fait que ces systèmes omettent d'intégrer des connaissances linguistiques, mais cette affirmation ne fait pas l'unanimité parmi les chercheurs dans ce domaine. En effet, l'amélioration des performances des systèmes de recherche d'information et une interaction efficace avec l'utilisateur passent, en partie, par des traitements linguistiques (morphologie, syntaxe et sémantique) et cognitifs. Certes, l'intégration des outils du TAL (Traitement Automatique de la Langue) dans les SRI est souhaitée par une grande majorité de chercheurs (surtout des linguistes-informaticiens), mais que les résultats liés à l'efficacité de ces outils sont souvent contradictoires.

Notre recherche se situe dans le cadre de l'apport des connaissances linguistiques à la recherche d'information sur le Web. La question que nous nous posons est celle de savoir si la prise en compte des connaissances linguistiques peut permettre d'améliorer la performance des SRI sur le Web, en termes de pertinence. Parmi ces connaissances linguistiques, nous ne traitons que celles liées à la morphologie. Il s'agit des variations morphologiques (dérivation et flexion) caractérisant les termes de la requête initiale. La fréquence des variantes morphologiques et leur impact sur la qualité (pertinence) des documents rapportés seront évalués. Ces variations, qui sont extraites des documents rapportés, seront examinées dans une optique de la reformulation de la requête initiale. Cette démarche a pour objectif d'aider l'utilisateur à la reformulation des requêtes pour mieux exprimer ses besoins en informations. Toutefois, le présent travail ne traitera pas spécifiquement de la reformulation des requêtes.

Notre travail sera composé de quatre chapitres. Le premier chapitre traitera de la problématique et des objectifs de recherche. Nous passerons en revue des questions liées à la recherche d'information, et cela dans deux contextes, classique et sur le Web. Nous aborderons également des aspects linguistiques et cognitifs auxquels est confronté ce domaine. Nous évoquerons également l'importance que représente l'intégration des outils du TAL aux systèmes de recherche d'information. Les objectifs de recherche consistent à examiner, d'une part, l'impact des variantes morphologiques sur la pertinence des documents rapportés et, d'autre part, la possibilité d'utiliser ces variantes dans une optique de reformulation de requêtes.

Le deuxième chapitre présentera le cadre théorique. Il s'agit d'un bref état de l'art des fondements théoriques liés au domaine de RI. À ce sujet, nous parlerons des notions de modèle et de pertinence en RI. De même, nous aborderons la question des variations linguistiques (morphologie, syntaxe et sémantique). Un intérêt particulier portera sur les variations morphologiques, notamment les dérivés.

Le troisième chapitre exposera le cadre méthodologique adopté pour mener notre étude. À cet égard, nous présenterons les différents choix adoptés concernant le corpus, les outils techniques ainsi que la démarche expérimentale suivie.

Le quatrième chapitre présentera les résultats obtenus et leur analyse. Nous commencerons par présenter les résultats bruts, suivis de leur interprétation. Nous

proposerons, ensuite, les résultats issus du test statistique de Jonckheere-Terpstra ainsi que leur interprétation. Nous présenterons, enfin, une analyse linguistique des résultats obtenus.

Notre conclusion mettra en exergue les points les plus saillants relevés des résultats obtenus, les limites de l'étude ainsi que les perspectives de recherche future.

# CHAPITRE I

# PROBLÉMATIQUE ET OBJECTIFS

Rechercher de l'information, plus particulièrement sur le Web, qui est un domaine ouvert et dynamique, est un exercice délicat. C'est un processus qui fait appel à divers paramètres : linguistiques, cognitifs, etc. Dans le contexte de la RI sur le Web, à la différence d'un système de recherche d'information classique, le besoin en information de l'utilisateur reste insatisfait. Ainsi, le présent chapitre se propose d'étudier la problématique de la RI sur le Web en comparaison avec le SRI traditionnel. Nous aborderons également la question du besoin d'information (BI) de l'utilisateur et l'interaction entre ce dernier et le SRI. Une autre section de ce chapitre portera sur les objectifs et les questions de recherche.

## 1.1 Problématique

La recherche d'information (RI) est un domaine pluridisciplinaire. Son but est de pouvoir automatiser, par le biais d'outils informatiques, l'accès à l'information. Ce domaine, qui a une longue histoire notamment dans le champ de la recherche documentaire, communément appelé la bibliothéconomie, a connu un essor considérable avec l'arrivée du Web. Cet engouement a eu pour conséquence la conception de nombreux outils logiciels (moteurs de recherche, entre autres), et cela afin de gérer une base de données où l'afflux d'informations est considérable et de seconder l'utilisateur dans ses recherches.

L'expansion du Web a été accompagnée par le développement de divers systèmes de recherche d'information. Certes, les moteurs de recherche sur le Web ont montré leurs limites, mais l'intégration d'une panoplie d'outils de traitement automatique des langues rend ces systèmes de plus en plus performants. Néanmoins, si les techniques du TAL ont montré leur efficacité dans les SRI classiques, c'est pour des raisons liées notamment aux caractéristiques de ces systèmes ; il est question, entre autres, de la taille des collections, de la nature des documents (texte), du caractère fixe et fermé de la base de collections, etc. L'application de ces outils et leur efficacité dans le contexte

de la RI sur le Web passe par une adaptabilité à cet univers qui est ouvert, dynamique, hétérogène et incontrôlé. À vrai dire, la prise en compte des techniques de TAL en RI est relativement récente, mais le rôle de ces techniques a un impact réel et attendu ; ce rôle sera de plus en plus présent dans des tâches de la RI (Jacquemin et Zweigenbaum, 2000).

### 1.1.1 Recherche d'information : SRI classique vs RI sur Web

Un système de recherche d'information (SRI) est un outil logiciel qui permet, par le biais d'un ensemble d'opérations, de rechercher, de collecter et de rapporter des documents (informations) en réponse à la requête d'un utilisateur. En ce sens, parler de la recherche d'information c'est parler également du repérage d'information, c'est-à-dire que les deux processus sont inter-reliés. L'utilisateur, qui est un acteur essentiel dans le processus de RI, interagit avec le SRI via une interface ; celle-ci lui permet d'exprimer son besoin d'information par le biais de requêtes formulées en langage naturel (liste de mots, phrase, formule booléenne, etc.). La satisfaction de ce besoin dépend largement de la qualité de formulation des requêtes (choix des mots, précision thématique, etc.). Si ce besoin d'information est plus ou moins satisfait en présence d'un contexte de recherche d'information classique, où les différents éléments du système sont contrôlés (profil utilisateur, collection de petite taille, sources connues et structurées, représentation des documents normalisée, etc.), ce n'est pas le cas dans le contexte de la recherche d'information sur le Web. Celui-ci contient et héberge une collection gigantesque et dynamique de documents (en différents formats et en diverses langues) qu'il est difficile de couvrir complètement. Dans ce contexte (Web), le défi d'un SRI consiste à pouvoir repérer les documents recherchés, c'est-à-dire les documents pertinents répondant au besoin de l'utilisateur. Ce qui s'avère difficile eu égard à l'hétérogénéité caractérisant le Web. En effet, la notion de pertinence est un critère essentiel dans le processus de RI, car la qualité du système en dépend.

Les SRI traditionnels ont été conçus pour traiter des documents de type textuel, c'est le cas, par exemple, d'un système de recherche documentaire qui permet de rechercher et sélectionner l'information dans un fonds documentaire conçu et structuré en tenant compte des paramètres propres à l'utilisateur (modèle utilisateur), mais, avec l'avènement du Web, le concept de document a pris d'autres dimensions en intégrant

d'autres médias comme les images, les sons et les vidéos. La diversification des médias a rendu les processus d'indexation (représentation du document), d'interrogation (formulation d'une requête) et de correspondance (appariement entre requête et document), plus ardus et complexes. Cela a un impact réel sur la performance des SRI traitant de ces informations.

Pour illustrer notre propos concernant un SRI classique, nous présentons à la Figure 1.1, ci-dessous, le fonctionnement d'un tel système ; il s'agit d'un schéma adapté de celui de Chbeir (2001) :

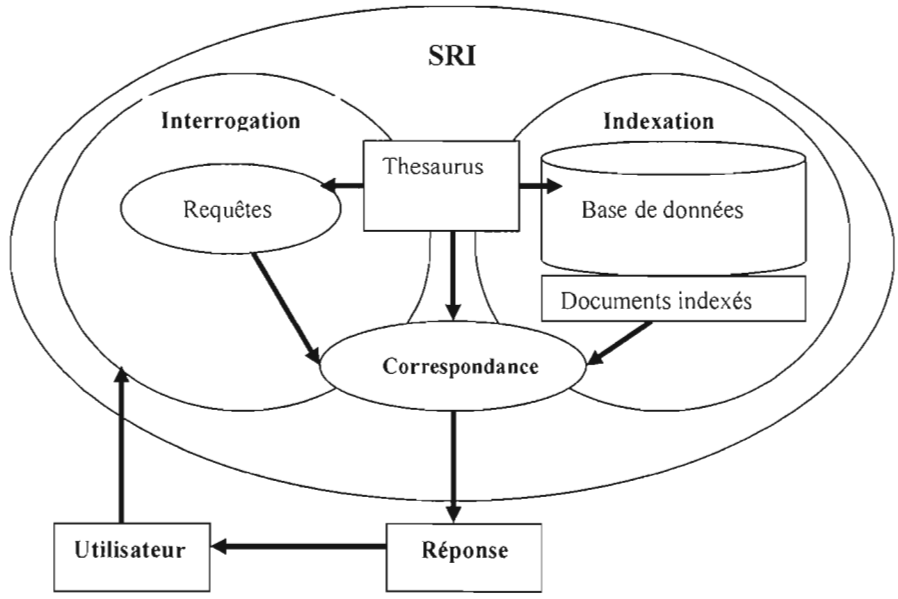



**Figure 1.1 : Schéma d'un système de recherche d'information classique**

Si la recherche d'information d'ordre textuel pose moins de problèmes dans un contexte de RI classique, qui se démarque par son homogénéité (sources connues, informations structurées et normalisées, etc.), la recherche de données textuelles sur le Web est confronté à plusieurs problèmes. Ces derniers sont le résultat de l'environnement dans lequel évolue le processus de RI. Par exemple, l'ajout de documents et le retrait d'autres se font d'une façon continuelle. De même, un même document peut être présent sur la toile en plusieurs versions (langues, formats, etc.). Dans ce contexte incontrôlable, les diversités linguistiques et thématiques caractérisant la collection des documents sur le Web engendrent de l'ambiguïté et rendent le

processus de RI plus délicat. La question qui se pose est de pouvoir repérer des documents pertinents qui sont noyés dans une collection en constante évolution. À cet égard, la tâche la plus délicate pour un SRI est d'établir une correspondance ou un appariement entre les termes de la requête formulée par l'utilisateur (l'information recherchée) et l'ensemble des documents pertinents disponibles. La mise en correspondance d'une façon efficace des termes de la requête avec les termes représentant les documents est un processus complexe auquel un SRI se trouve confronté. L'ambiguïté, caractéristique des langues naturelles, est l'un des problèmes d'ordre linguistique auquel ce processus est confronté. L'homographie constitue un des phénomènes contribuant à l'ambiguïté linguistique ; il est question de deux formes ou deux mots qui s'écrivent de la même façon (même graphie) sans avoir forcément le même sens. C'est l'exemple en français de : ***sens*** (Nom) et ***sens*** (Verbe), ***couvent*** (Nom) et ***couvent*** (Verbe), ***tour*** (Nom féminin) et ***tour*** (Nom masculin), etc. Ce phénomène (l'homographie) peut caractériser deux langues différentes, c'est-à-dire que des mots appartenant à deux langues différentes peuvent avoir une même représentation graphique, mais des significations distinctes ; par exemple, le mot ***store***, qui existe en français et en anglais, signifie ***magasin*** en anglais et ***rideau*** en français ; un autre exemple est celui du mot ***accommodation*** qui signifie ***logement*** en anglais et ***adaptation*** en français.

Ce type de phénomènes (homographie) a un impact sur la langue des documents rapportés notamment dans le cas de requêtes courtes, c'est-à-dire que l'utilisateur va avoir, parmi les résultats rapportés, des documents répondant à sa requête dans sa langue et, aussi, des documents dans d'autres langues. Pour limiter le risque lié à cette ambiguïté, le choix des termes de la requête et leur nombre s'avèrent pertinents. Certes, des moteurs de recherche offrent la possibilité de choisir la langue de recherche, c'est-à-dire de paramétrer l'outil de recherche pour n'avoir, par exemple, que des documents en français. Cette option existe dans les moteurs de recherche comme Google ; c'est vrai, cela réduit les risques, mais ne règle pas le problème d'ambiguïté pour autant. Une autre possibilité permettant de limiter l'ambiguïté sémantique associée à l'homographie est la formulation de la requête sous forme d'expression. Toutefois, le recours à cette dernière possibilité, c'est-à-dire la formulation de la requête sous forme d'une expression, aura comme effet de limiter le

nombre de documents pertinents rapportés, ce qui se traduit par l'augmentation de la précision.

Un autre phénomène d'ambiguïté est celui associé à la polysémie. Celle-ci est définie comme la propriété d'un mot qui peut référer à des concepts différents ; en d'autres termes, un mot qui peut avoir deux ou plusieurs sens différents. C'est le cas, par exemple, de ***vol*** (*délit ; déplacement* (*espace*)...), de ***verre*** (*matière ; récipient*...), de ***bureau*** (*table de travail ; pièce où est installée la table de travail*...). De même, la synonymie est un autre phénomène linguistique auquel la RI sur le Web fait face. Ce phénomène désigne la propriété d'un mot (ou concept) qui peut être reformulé différemment ; autrement dit, les liens sémantiques de proximité qu'un mot entretient avec les autres mots de la même langue ; c'est l'exemple de ***crise*** (*malaise, conflit, tension*...), de ***combat*** (*lutte, bataille, assaut*...). Ces phénomènes d'ambiguïté sémantique entraînent inévitablement la récupération de documents ne répondant pas au besoin de l'utilisateur, c'est-à-dire des documents non pertinents. Un autre phénomène linguistique très présent dans le cadre de la RI concerne les variations morphologiques (ou graphiques) des mots. Cela implique la possibilité de formuler différemment le même terme. Il s'agit de termes qui sont proches d'un point de vue graphique et sémantique. Le problème de la variation morphologique concerne aussi bien la dérivation que la flexion. Par exemple, en français, le mot ***exporter,*** qui est un verbe, est lié aux variantes *exporte, exportes, export, exportation, exportateur, exportatrice, etc.* Ainsi, en RI le problème qui se pose c'est la prise en compte de la variation morphologique des mots de la requête. Cette variation peut empêcher, par exemple, le rapatriement de documents pertinents juste parce que l'appariement des termes ne prend pas en compte les variantes. C'est le cas d'une requête contenant le terme ***crime*** mais qui ne sera pas appariée avec des documents contenant, par exemple, les variantes ***criminalité, criminel, crimes,*** *etc.* Cette variation morphologique peut être aussi exploitée pour enrichir les requêtes de l'utilisateur dans le contexte de la reformulation de requêtes. Pour cela, les techniques du TAL ont apporté des solutions qui ont montré leur efficacité, mais la question de la reformulation des requêtes par le biais des variantes morphologiques pose un réel problème de glissement sémantique. Nous soulevons, dans la section suivante, la

question liée à l'intégration en RI de paramètres d'ordre cognitif[1]. Il s'agit de la prise en compte des caractéristiques de l'utilisateur pour mieux optimiser son BI, et cela en l'aidant, par exemple, à la tâche de reformulation des requêtes ainsi que celle de la visualisation des résultats rapportés.

### 1.1.2 Utilisateur et besoin d'information

La prise en compte de l'utilisateur par les systèmes de recherche d'information est un paramètre important. L'intégration de ce paramètre peut avoir un impact réel sur la qualité des interactions entre le système et l'utilisateur ; il est question ici de l'utilisabilité et de l'adaptabilité des systèmes à l'utilisateur (problème d'exploitation et visualisation des résultats de la requête). Le problème de l'accessibilité des documents retournés est présent dans le contexte de la RI sur le Web. À ce sujet, l'ordre de présentation des documents diffère d'un moteur de recherche à l'autre et les documents recherchés pourraient très bien ne pas figurer dans les premières pages retournées ; cette situation priverait l'utilisateur d'accéder à des informations pertinentes surtout que ce dernier ne visualise que les premiers résultats rapportés (Chieze, 2006), généralement les trois premières pages (Bonnel et Moreau, 2005). Le comportement de l'utilisateur et l'accessibilité des résultats retournés sont des problématiques associées au SRI ; ces deux aspects ne feront pas partie de notre objectif de recherche dans la présente étude.

La reformulation des requêtes et leur extension en fonction des besoins en information de l'utilisateur sont des éléments à prendre en considération dans ce processus. À cet égard, des techniques de reformulation de la requête de l'utilisateur sont adoptées afin de maximiser la qualité des réponses rapportées. L'une de ces techniques (Moreau et Claveau, 2006) consiste à enrichir les termes de la requête en utilisant leurs variantes morphologiques ; il s'agit d'une méthode d'extension de requêtes par le biais des variations morphologiques. Une autre technique (Chieze, 2006) consiste à réécrire les requêtes de base en utilisant les expressions saillantes extraites des premiers documents de la requête initiale. Encore une fois, les techniques de reformulation ont

[1] Pour une analyse plus approfondie des aspects cognitifs en RI, nous vous référons à la thèse de Chieze (2006).

montré leurs limites dans le cadre de la RI sur Web. Contrairement à un SRI traditionnel, établir un modèle utilisateur dans le contexte du Web pour mieux répondre à ses attentes reste un défi majeur pour ne pas dire utopique.

D'un point de vue cognitivo-linguistique, quand un utilisateur a un BI donné, il le traduit par des mots en interrogeant un moteur de recherche. En d'autres termes, avant d'interroger le SRI, l'utilisateur construit une représentation mentale de son BI et il l'exprime, après, par le biais de requêtes en langue naturelle. De même, le BI de l'utilisateur est évolutif, c'est-à-dire qu'il se réajuste et se précise avec la visualisation et l'exploration des documents rapportés (Chieze, 2006). La question qui se pose est la suivante : est-ce que les mots choisis par l'utilisateur lors de la formulation de sa requête reflètent vraiment son BI ? C'est à ce niveau qu'intervient l'intérêt de la reformulation de requêtes ; celle-ci consiste à réécrire ou étendre les termes de la requête initiale. La qualité de cette démarche dépend notamment du niveau de l'utilisateur et sa capacité à choisir les termes pertinents pour pouvoir optimiser les résultats rapportés. En ce sens, pouvoir déterminer la pertinence des termes est une tâche difficile pour un utilisateur qui a une connaissance limitée de la collection de documents présente sur le Web. En effet, le système peut satisfaire la requête telle qu'exprimée par l'utilisateur, mais il ne peut pas garantir efficacement le BI désiré. Ce dernier reste approximatif, voire aléatoire. Cette situation ne relève pas uniquement des problèmes d'ordre technique ou informatique, mais elle est le corollaire de la complexité du langage humain et les propriétés des langues.

## 1.2 Objectifs de recherche

Divers travaux ont porté sur la recherche de données textuelles sur le Web. Ces travaux ont traité d'aspects aussi bien informatiques que linguistiques. La présente recherche n'a pas pour objectif d'évaluer la performance d'un système de recherche d'information sur le Web. Nous nous intéressons, plutôt, à l'apport des connaissances linguistiques au domaine de la RI. Il est question, dans notre étude, des variations morphologiques. Notre objectif consiste à examiner et à évaluer l'impact éventuel des variantes morphologiques, notamment dérivationnelles, sur la pertinence des documents rapportés. Cette étude exploratoire va nous permettre de vérifier l'intérêt de la prise en compte des variantes morphologiques dans la RI en français sur le Web.

Notre étude prend appui, en partie, sur les travaux traitant des variations linguistiques, notamment morphologiques, dans la recherche d'information (Jacquemin & Tzoukermann, 1999 ; Daile et Morin, 2000 ; Namer, 2000 ; Chieze, 2000, 2006 ; Emirkanian & Chieze, 2003 ; Moreau et Sébillot, 2005 ; Moreau et Claveau, 2006).

Les questions que nous nous posons sont les suivantes :

- la présence et fréquence de variantes morphologiques (dérivées et fléchies) des termes de la requête de base dans les documents rapportés a-t-elle un impact sur la qualité de ces documents en termes de pertinence ?

  Nous postulons que la fréquence des variantes morphologiques, notamment dérivées, dans les documents rapportés est en corrélation avec leur degré de pertinence.

- dans une optique de reformulation des requêtes basée sur les variantes extraites des documents rapportés, quel serait l'impact de la prise en compte des variations morphologiques sur la pertinence des résultats ?

  Nous postulons que la pertinence dépend, en partie, de la nature et de la qualité des variantes.

  Il est à noter que notre travail a pour objectif premier de répondre à la 1ère question. La 2ème question, qui porte sur l'impact de la prise en compte des variations morphologiques dans une optique de reformulation des requêtes, sera seulement abordée sans pouvoir ainsi être implémentée.

En termes de variables, les questions de recherche posées ci-dessus nous permettent de dégager deux variables, l'une indépendante et l'autre dépendante. La *fréquence des variantes morphologiques* constitue la variable indépendante. La *pertinence des documents* est la variable appelée dépendante ; cette variable représente un comportement issu de l'action de la variable indépendante. Celle-ci est indépendante parce qu'elle est définie et mise en place par le chercheur. L'expérimentation envisagée dans notre recherche nous montrera le degré de corrélation entre les deux variables en question.

## 1.3 Conclusion

Nous avons passé en revue, dans ce chapitre, la question de la RI sur le Web par rapport à la RI classique. À ce sujet, nous avons mis l'accent sur les difficultés auxquelles la RI est confrontée. Ces difficultés touchent plus particulièrement la RI sur le Web. Ce dernier est considéré comme un espace complexe et hétérogène. Ce caractère hétérogène est le corollaire de la nature structurelle du Web ; il s'agit, entre autres, de la variété des contenus des pages, de leurs formats ainsi que de la diversité des langues. Ces paramètres, et bien d'autres, rendent l'accès à l'information recherchée par l'utilisateur plus difficile.

Le but d'un SRI est de rapporter des informations en réponse à une requête. Celle-ci est exprimée par l'utilisateur en langage naturel, et cela sous diverses formes : liste de mots-clés, phrases, formules booléennes, expressions, etc. Toutefois, la complexité des langues et leur ambiguïté (homographie, polysémie, synonymie, etc.) ne permettent pas un repérage d'information efficace et pertinent. Cette complexité est associée également aux capacités cognitivo-linguistiques de l'utilisateur. Le comportement de ce dernier lors du processus de la formulation et reformulation des requêtes, les connaissances encyclopédiques sur les domaines et thèmes de recherche et celles liées au fonctionnement des systèmes de recherche d'information s'avèrent importants dans une démarche de recherche d'information plus efficace. En ce sens, un utilisateur avisé serait capable d'exprimer ses besoins informationnels en choisissant et formulant les termes de la requête les plus adéquats. Nous entendons par adéquation, dans le contexte de la RI sur le Web, la correspondance entre les termes de la requête et les documents rapportés. Vu la complexité des langues, cette correspondance n'est pas forcément un gage de pertinence.

Le chapitre suivant sera consacré au cadre théorique de notre recherche. Nous présenterons un état de l'art du domaine de la RI. Nous aborderons également l'intégration des connaissances linguistiques, plus particulièrement les variations morphologiques, et leur apport à la RI.

# CHAPITRE II

# CADRE THÉORIQUE

Le présent chapitre traitera du cadre théorique de notre étude. Nous présenterons, d'abord, un bref état de l'art du domaine de la recherche d'information. Ce survol théorique abordera, ensuite, la notion du modèle en RI ainsi que celle de la pertinence. La dernière section (2.4) se proposera d'étudier une question fondamentale dans notre recherche, il s'agit des connaissances linguistiques et leur impact sur la recherche d'information ; à ce sujet, un intérêt particulier sera porté sur la question des variations morphologiques, objet de notre travail de recherche. Le but escompté est de pouvoir optimiser la recherche d'information par le biais de la reformulation des termes de la requête initiale en nous basant, pour réécrire les requêtes, sur la variation morphologique.

## 2.1 Recherche d'information : survol du domaine

L'expansion du nombre de documents dont le contenu est accessible par le biais d'outils informatiques est en perpétuel développement. Cet engouement est associé notamment au développement du Web. Pour faciliter l'accès à l'information et pouvoir retrouver des documents dans cet ensemble immense et dynamique qu'est le Web, divers systèmes de recherche d'information (SRI) ont été conçus. Un SRI permet à un utilisateur d'exprimer son besoin d'information au moyen d'une requête. Le SRI établit ensuite la liste des documents pertinents en fonction du besoin exprimé par l'utilisateur. La qualité des résultats retournés dépend notamment du thème de recherche et de la formulation de la requête. Celle-ci peut ne pas être satisfaisante, c'est-à-dire exprimant au mieux le besoin de l'utilisateur ; satisfaire ce besoin passe surtout par l'application de diverses reformulations des termes de la requête. Le but escompté est de pouvoir maximiser le nombre de documents pertinents rapportés et par le fait même minimiser le nombre de documents non pertinents.

Généralement, les moteurs de recherche sur le Web ne dévoilent pas leur mécanisme de fonctionnement (algorithmes). Ces moteurs sont considérés comme des boîtes

noires dont l'accès est limité aux seuls concepteurs. Le secret entourant les moteurs de recherche n'empêche pas les utilisateurs avisés et les spécialistes du domaine de deviner approximativement leur fonctionnement. Un moteur de recherche[2] «est un outil logiciel» qui comprend des robots ou agents. Ces derniers parcourent le Web (sites, forums, etc.) d'une façon automatique et à intervalles réguliers. En parcourant le Web, ces agents archivent les pages (textes, images, fichiers audio, etc.) trouvées dans les serveurs. L'indexation des pages s'effectue automatiquement en associant à chaque page les mots clés inscrits dans la page, la position des mots dans la page et leur répétition, l'indice de popularité de la page, etc. Lorsqu'un utilisateur interroge le système, sa requête est envoyée au moteur de recherche. La tâche de ce dernier consiste à consulter ses bases de données pour chacun des mots de la requête. Il récupère les réponses sous forme de liens vers des pages. Celles-ci sont présentées à l'utilisateur selon un ordre de pertinence décroissant, cela en fonction du degré de similarité des documents avec la requête, c'est-à-dire que les documents retournés à l'utilisateur sont définis sur la base du degré de correspondance entre le contenu des documents et celui des termes de la requête (Bonnel et Moreau, 2005).

La plupart de ces systèmes laissent à l'utilisateur le soin d'exprimer son besoin d'information au moyen d'une requête composée d'un ensemble de mots, d'une phrase en langage naturel, etc. À cet égard, ces systèmes représentent les documents et les requêtes par les mots les constituant (communément appelés ''sacs de mots''). L'efficacité du processus d'appariement entre requête et index est liée notamment aux mécanismes de simplification de la requête et des documents. Cette simplification se traduit, à titre d'exemple, par la suppression des mots grammaticaux (mots outils), mots fréquents, mots sans pouvoir discriminatoire, etc. Le but consiste à assurer une meilleure représentation du contenu des documents ; autrement dit, la pertinence des mots choisis pour l'indexation des documents a un impact direct sur la pertinence des résultats de la recherche. La ra cinisation et l'amalgame des mots sont parmi les méthodes de simplification adoptées par ces systèmes ; la racinisation est «une procédure qui vise à regrouper les mots sémantiquement proches à partir de ressemblances "graphiques"» (Jacquemin et Zweigenbaum, 2000). Cette procédure permet, par exemple, d'obtenir une forme (racine) commune aux différentes variantes

[2] http://www.ebsi.umontreal.ca/jetrouve/internet/moteur.htm

morphologiques ; c'est le cas des variantes morphologiques *cheval, chevaux, chevalier, chevalerie, chevaucher, etc.* qui sont ramenées à la racine ou forme tronquée "*cheva*". Cette façon de faire a l'avantage de regrouper diverses variantes dérivationnelles en les ramenant à la même racine à moindre frais, c'est-à-dire que cette opération se fait, dans la plupart des cas, au détriment du sens des variantes morphologiques et des lemmes initiaux. Nous ne pensons pas que les moteurs de recherche sur le Web intègrent cette procédure (racinisation), mais des outils d'aide à l'utilisateur comme la "recherche apparentée" sont présents dans quelques moteurs de recherche, c'est le cas de Google. Cette procédure consiste à proposer des suggestions de requêtes apparentées à la recherche initiale. Le problème c'est qu'on ne sait pas sur quoi se base Google pour afficher ce genre de suggestions.

D'autres paramètres entrent également en jeu, c'est le cas, par exemple, de la proximité des termes dans le document, leur position et leur ordre d'apparition. Le recours à une syntaxe spécifique permet d'appliquer des contraintes lors de la formulation des termes de la requête. À cet égard, les moteurs de recherche sur le Web proposent diverses options de recherche dans leur mode avancé (choix de la langue, possibilité d'exprimer la requête en mode booléen, etc.), mais ces options restent inexplorées par la majorité des utilisateurs.

Certes, les précurseurs de ce domaine avaient pour objectif de pouvoir automatiser la RI, et cela pour pouvoir faciliter l'accès à l'information. Mais le développement du Web a été accompagné d'une masse considérable de documents numériques, ce qui a constitué un défi pour les SRI. Autrement dit, le problème ne consiste pas seulement à accéder à l'information, mais il est plutôt question du filtrage des informations qui répondraient mieux au besoin de l'utilisateur, c'est-à-dire la possibilité de rapporter seulement les ''informations pertinentes'' présentes dans l'ensemble de la collection.

La section suivante abordera la notion de modèle en RI. Nous nous limitons à présenter les deux modèles les plus utilisés ; il s'agit du modèle vectoriel et du modèle booléen.

## 2.2 Modèles de RI

Le choix du modèle est central dans la recherche d'information textuelle. C'est le modèle qui permet de donner une interprétation aux termes des documents et des requêtes pondérés par l'indexation. Ainsi, la tâche d'un modèle consiste, d'une part, à créer une représentation interne des documents et de la requête (par le biais de termes) et, d'autre part, à établir une comparaison entre les deux représentations, et cela afin de déterminer le degré de correspondance entre les termes de la requête et ceux des documents. Dans le contexte de la RI plein texte, divers modèles sont utilisés par les SRI. Parmi les modèles les plus connus et les plus utilisés, citons les approches vectorielle et booléenne (Salton et McGill, 1983). La rapidité d'exécution constitue la qualité principale de ces approches. C'est le cas des moteurs de recherche sur le Web adoptant ces modèles (AltaVista, Google, etc.) ; ces moteurs sont capables de retrouver, en quelques secondes, des documents noyés dans des collections contenant des centaines de millions de documents.

- **Le modèle vectoriel** : ce modèle est apparu avec le SRI SMART (Salton, 1983). Dans ce modèle, les documents et la requête sont représentés par des vecteurs de termes d'indexation dans un espace vectoriel à $n$ dimensions. Cet espace représente l'ensemble de termes indexés par le système. Les termes d'indexations sont des mots-clés $[t_1, t_2, t_3 \ldots, t_n]$, où $t_n$ est le nombre de termes indexés. La correspondance entre les termes est basée sur une mesure de similarité entre les vecteurs représentant les documents et la requête. Dans cet espace multidimensionnel, chaque terme a un poids qui reflète son importance dans les documents et la requête ainsi que son caractère discriminant ou non. Le poids des termes peut être exprimé dans ce modèle par le biais des valeurs de *tf*- (fréquence de terme), et *idf*- (fréquence documentaire inverse). Ce modèle est basé sur une logique binaire, c'est-à-dire un document correspond ou ne correspond pas. De même, ce modèle permet d'atteindre de bons résultats en termes de pondération. Cela est dû notamment à la structure du modèle qui peut être qualifiée de compositionnelle ; en ce sens, les documents pertinents rapportés sont le corollaire de connaissances micro-structurelles ou atomiques associées à ces documents, il est question ici de mots-clés. En d'autres termes, la pertinence d'un document est déterminée par le nombre et la fréquence des termes partagés entre la requête et le document. Une autre caractéristique de ce modèle concerne le tri des documents

pertinents rapportés ; ces derniers sont triés par ordre décroissant de similarité (Martinet et al. 2002).

- **Le modèle booléen** : dans le modèle booléen, un document est représenté par un ensemble de termes. À la différence de l'approche vectorielle, la notion de fréquence n'intervient pas. Toutefois, ce modèle fait intervenir notamment le principe de proximité des termes dans les documents, et cela par le biais de l'opérateur NEAR ; c'est-à-dire que le sens d'un document (et sa pertinence) ne dépend pas seulement de la fréquence des termes (cas du modèle vectoriel) mais aussi de la proximité spatiale de ces termes. Ce principe requiert, par exemple, que deux termes soient proches (soit à une distance de *n* mots, soit dans la même phrase soit dans le même paragraphe). Le modèle booléen suppose que les termes des documents sont liés. De même, la fonction de correspondance, dans ce modèle, est fondée sur la notion d'implication, c'est-à-dire que la correspondance entre le document et la requête suppose une implication logique des propositions ; autrement dit, il s'agit d'un processus compositionnel où le sens global du texte (document) et sa pertinence sont le résultat d'une construction micro-structurelle impliquant les liens entre les termes et les phrases constituant le document. L'inconvénient de cette implication logique est lié au fait que les documents ne contenant pas tous les termes de la requête sont généralement considérés comme non pertinents. Par ailleurs, une autre limite du modèle booléen concerne les résultats rapportés. Ceux-ci ne peuvent être triés, ils sont retournés à l'utilisateur en vrac, c'est-à-dire que les documents retournés ne sont pas ordonnancés en fonction de leur degré de correspondance avec les termes de la requête.

Parler de modèles en RI suppose forcément la présence de la notion de pertinence. Celle-ci fera l'objet de la section suivante.

## 2.3 Notion de pertinence

L'objectif principal d'un SRI consiste à retrouver les documents pertinents qui répondent aux besoins formulés par l'utilisateur. La notion de pertinence est un élément clé et complexe dans le processus de recherche d'information. Avoir une définition claire et non équivoque de la pertinence n'est pas une tâche facile. De même, la notion de pertinence est conditionnée par divers paramètres et contextes relevant à la fois du système et de l'utilisateur. La pertinence liée au système peut être

qualifiée d'algorithmique ; il s'agit de trouver des documents dans une base en réponse à la requête formulée en langue naturelle par l'utilisateur. Les documents retournés sont souvent classés selon leur degré de pertinence. Par ailleurs, la pertinence relevant de l'utilisateur est plutôt d'ordre cognitivo-linguistique ; il est question de la représentation que l'utilisateur pourrait faire des documents retournés par le système ainsi que le degré de satisfaction en fonction de son besoin d'information ; autrement dit, la relation entre la requête et le document (Saracevic, 1970).

En effet, la pertinence en recherche d'information reste une notion vague et variable. C'est une notion qui varie en fonction de différents facteurs notamment ceux liés au système et à l'utilisateur. Dans cette optique, Saracevic (1970) a dressé une liste de quelques définitions de la pertinence en vigueur à cette époque. Pour Saracevic (1970), la pertinence est définie comme :

- la correspondance ou relation entre un document et une requête ;
- le degré de chevauchement entre un document et une requête ;
- le degré de surprise d'un document par rapport aux besoins de l'utilisateur, etc.

Saracevic (1970), en s'inspirant des définitions proposées par ses prédécesseurs, a émis une définition unificatrice. Celle-ci stipule que la pertinence est **A** de **B** existant entre **C** et **D** jugé par **E** (**A** = un intervalle de mesure du degré de pertinence (binaire ou multivaluée) ; **B** = la pertinence absolue ; **C** = un document ; **D** = le contexte de mesure de la pertinence ; **E** = le juge). Certes, cette définition permet d'identifier les acteurs principaux faisant partie de cette notion, mais elle ne fournit pas d'explications précises sur la nature de la pertinence, la relation logique entre un document et une requête ainsi que la nature des facteurs contextuels.

Les variations caractérisant l'évaluation de la pertinence sont liées notamment à l'évolution du domaine de la recherche d'information. En effet, le souci de vouloir tester et évaluer expérimentalement les méthodes adoptées en RI a toujours été présent, et cela depuis les premiers travaux en RI. En ce sens, l'évaluation de la pertinence a évolué en passant d'une évaluation dite classique, basée sur une approche binaire de la pertinence (pertinent ou non pertinent), à une évaluation récente basée sur

une approche multivaluée (divers degrés de pertinence). L' approche classique est considérée comme une évaluation simple qui se fait dans un environnement réduit et contrôlé (collection des documents de petite taille) ; c'est le cas des expériences de Cranfield, premières expériences portant sur l'évaluation dans les années 60 (Cleverdon, 1967). Le projet Cranfield avait pour but de tester l'efficacité de nouvelles méthodes d'indexation et de recherche de documents. La collection de tests comprenait des articles (18.000) et des requêtes (1200) ; les requêtes ont été évaluées par des experts, et cela afin d'établir des correspondances, en termes de pertinence, entre les articles de la collection et les requêtes. Les principes d'évaluation adoptés dans le projet Cranfield ont marqué le domaine de RI à tel point que ces mêmes principes sont encore adoptés de nos jours dans les systèmes de RI.

Notre approche est celle d'une évaluation multivaluée. Cette approche inspirée de Chignell et al. (1999) et Sormunen (2002) consiste à évaluer la pertinence des documents en nous basant sur une échelle comprenant 4 paliers : 0, 1, 2 et 3. La valeur 0 correspond à un document non-pertinent, la valeur 1 correspond à un document peu pertinent (seulement le BI est mentionné), la valeur 2 correspond à un document partiellement pertinent et, finalement, la valeur 3 correspond à un document pertinent.

Par ailleurs, d'autres notions sont utilisées dans le processus d'évaluation des documents en RI, il est question des principes de précision et de rappel (c'est une approche d'évaluation classique). La précision est une mesure correspondant à la proportion de documents pertinents parmi tous les documents retrouvés. Le rappel est une mesure désignant la proportion de documents pertinents retrouvés parmi tous les documents pertinents de la collection. L'idéal serait d'avoir un système de recherche d'information dont la précision et le rappel soient égaux (ce qui est possible en présence d'une base de données de petite taille), c'est-à-dire que le système rapporte uniquement les documents pertinents d'une collection (Loupy, 2001). Ces deux mesures complémentaires sont plus adaptées à des expériences où la taille de la collection testée est généralement petite et statique. Ce qui n'est pas le cas actuellement avec la présence de collections de plus en plus grandes, comme c'est le cas du Web. Le recours à des approches d'évaluation plus récentes et plus adaptées à la taille des collections s'est avéré nécessaire. Ces nouvelles approches ont été

développées et adoptées grâce aux expériences comme TREC, AMARYLLIS, etc. (Lespinasse et al., 1999).

La section suivante abordera le thème lié aux connaissances linguistiques et leur apport à la RI. Un intérêt particulier portera sur les variations morphologiques.

## 2.4 Prise en compte des connaissances linguistiques

L'apport des connaissances linguistiques à la recherche d'information est un sujet de débat qui a accompagné l'évolution de ce domaine. À ce sujet, divers travaux ont traité, entre autres, des variations linguistiques (morphologiques, syntaxiques et sémantiques) et de leur impact sur la recherche d'information. Parmi ces travaux nous citons, à titre d'exemple, ceux effectués par : Jacquemin et Tzoukermann, 1999 ; Jacquemin, 1997 ; Assadi et Bourigault, 1996 ; Dias et al., 2000 ; Namer, 2000 ; Gaussier et al., 2000 ; Chieze, 2000 ; Zweigenbaum et al., 2001 ; Emirkanian et Chieze, 2003, etc. Ainsi, Jacquemin et Tzoukermann (1999) ont développé un système permettant une indexation automatique des données dédié au français ; ce système a pour caractéristique de prendre en compte les variations morphologiques ainsi que le contexte syntaxique des termes (combinaison des termes). Ainsi, les auteurs distinguent les variantes morphologiques (par exemple, "*gene is located...*" est une variante de "*gene location*"), les variantes syntaxiques (par exemple, "*diseases of the lower urinary tract*" est une variante syntaxique de "*urinary tract disease*") ainsi que les variantes morpho-syntaxiques (par exemple, "*translational inhibition*" est une variante morpho-syntaxique de "*translation inhibitor*"). Les résultats obtenus par ce système ont montré l'apport réel, en termes de précision, de la combinaison entre la morphologie et la syntaxe.

Namer (2000) a développé un analyseur morphologique dédié au français ; il s'agit plus exactement d'un lemmatiseur du français appelé FLEMM et qui est basé sur des règles. C'est un programme qui opère sur un texte étiqueté. De même, des connaissances linguistiques sont prises en compte dans le processus de lemmatisation, c'est-à-dire qu'il ne s'agit pas d'une démarche opérant par troncation. Une autre caractéristique de FLEMM est sa robustesse ; celle-ci se traduit par sa capacité à analyser les mots inconnus. Ainsi, les résultats des tests de l'analyseur ont montré la pertinence et l'utilité d'un analyseur morphologique en RI ; en ce sens, cet analyseur

est un outil du TAL, parmi d'autres, qui peut être bénéfique aux SRI ; cet outil permettra, par exemple, d'analyser les unités lexicales absentes des dictionnaires, d'élargir la famille lexicale des mots analysés (variantes morphologiques, utiles pour la reformulation des requêtes), etc. Chieze (2000) a traité des variations morphologiques flexionnelles et de leur apport au repérage d'information sur le Web ; il s'agit d'un processus de reformulation de requêtes appliquant une démarche automatisée d'intégration de la morphologie flexionnelle du français. Cette démarche consiste à «transformer une requête vectorielle [sans enrichissement morphologique] en un ensemble de sous-requêtes booléennes [avec enrichissement morphologique]» (Chieze, 2000, p.52). L'une des conclusions de cette étude montre que la reformulation des requêtes de base par le biais d'un enrichissement morphologique flexionnel a un impact minime sur le RI sur le Web. Dans la même optique, c'est-à-dire la reformulation des requêtes, Emirkanian et Chieze (2003) ont mené une étude traitant des variations morphologiques, syntaxiques, sémantiques et de leur impact sur le RI sur le Web. Les résultats des tests liés à cette étude ont donné lieu à des conclusions selon lesquelles les variations morphologiques et sémantiques (synonymiques) permettent dans certains cas d'améliorer le rappel, mais cela au détriment de la précision (cela dépend de la requête). De même, au niveau syntaxique, cette étude a montré que la prise en compte du contexte d'utilisation des termes de la requête et de leur structure syntaxique a un impact réel sur la précision. Une autre conclusion intéressante de cette étude a montré que la précision augmente nettement en présence d'un contexte verbal, c'est-à-dire le contexte «où le MOT1 (1$^{er}$ mot de la requête) est un verbe». Toutefois, ce contexte (V-SP et V-SN) est rare, puisque qu'il ne représente que 15%, et dépend notamment du corpus étudié (Emirkanian et Chieze, 2003, p.149).

Les auteurs des travaux cités ci-dessus s'entendent sur le fond pour dire que l'apport des connaissances linguistiques à la recherche d'information est à prendre en considération, et cela malgré les contradictions caractérisant souvent les résultats obtenus. Ces connaissances contribuent à améliorer, entre autres, l'indexation des documents en créant une représentation plus riche de leur contenu. En effet, ces paramètres linguistiques constituent un avantage important dans la mesure où les termes des documents et de la requête seront considérés comme des unités linguistiques et non pas seulement comme de simples chaînes de caractères. Cela

permet d'améliorer l'appariement entre l'information recherchée (formulée par le biais d'une requête) et les documents de la collection. De même, ce processus offre à l'utilisateur la possibilité d'enrichir ses requêtes (reformulation des requêtes de base en recourant à des informations morphologiques, syntaxiques et sémantiques), et pouvoir obtenir, ainsi, davantage de documents répondant à son besoin d'information (Moreau et Sébillot, 2005).

Au niveau morphologique, le français fa it partie des langues réputées pour leurs irrégularités caractérisant leurs unités lexicales. En effet, la graphie des mots subit plusieurs variations d'ordre flexionnel (nombre, genre, conjugaison, etc.) et dérivationnel (formation de nouveaux mots à partir d'un radical : préfixation, suffixation, etc.). Certes, des irrégularités peuvent avec des unités lexicales surtout celles construites par dérivation. Mais ces irrégularités ne sont pas systématiques, c'est-à-dire qu'elles constituent des exceptions ; par exemple, une même unité lexicale peut avoir des significations grammaticales et sémantiques différentes, comme c'est le cas de *antidémocratique, antimondialisation*, etc. "*anti-*" a le sens de *opposé à,* alors que *antimoine, antilope,* etc. "*anti-*" n'a pas la même signification (*opposé à*). De même, cette irrégularité fait en sorte que les informations dérivationnelles sont partiellement productives. C'est-à-dire que des mots appartenant à la même classe peuvent ne pas subir les mêmes dérivations. Ainsi, pourquoi *mangeable* et non **comportable*[3], etc. (Bouillon et al., 1998). Ce phénomène reste marginal en français qui est une langue où les irrégularités sont considérées comme des exceptions.

La contre performance associée aux variations d'ordre dérivationnel est liée au fait que le processus de dérivation entraîne de nombreux glissements sémantiques, c'est le cas par exemple des variations suivantes : ***populaire*** signifie *propre au peuple* ou *avoir du succès, être célèbre*, alors que la variante ***impopulaire*** signifie *quelque chose mal vu, qui n'a pas de succès*. De même pour l'exemple de ***planète*** (*globe, terre...*) et ***planétaire*** (*international, mondial...*), etc. Les mots reliés morphologiquement ont généralement une racine commune, et leur différence se situe au niveau de leurs affixes (préfixes et suffixes) ; cette variation dérivationnelle, en plus des modifications sémantiques, entraîne des changements dans les catégories grammaticales des mots

[3] C'est un sujet de débat. Des critères syntaxiques ou sémantiques peuvent justifier cette irrégularité.

(*stabiliser* [V] ↔ *stabilisation* [N] ; *différent* [ADJ] ↔ *différemment* [ADV] ; *espace* [N] ↔ *spatial* [ADJ], etc.). On assiste au même phénomène du glissement sémantique avec un degré moindre au niveau des variations morphologiques flexionnelles. C'est le cas par exemple des variantes comme *fond* (au sens de partie la plus basse d'un objet creux) et *fonds* (au sens de capital, somme d'argent, etc.) ; *affaire* (au sens politique) et *affaires* (au sens économique), *stupéfiant* (au sens d'étonnant et aussi de drogue) et *stupéfiants* (au sens de drogue), etc. Nous constatons un changement de catégorie grammaticale du dernier exemple où *stupéfiant* au sens d'étonnant est un adjectif, tandis que *stupéfiant ou stupéfiants* au sens de drogue(s) sont des substantifs (Chieze, 2000). L'ambiguïté sémantique engendrée par ces variations peut avoir un impact sur la pertinence des résultats rapportés. A propos de ce glissement sémantique qui survient lors de la formation des mots (morphologie dérivationnelle), Corbin (1991) soutient l'idée d'un processus associatif et stratifié qui fait en sorte que le sens d'un mot est construit en même temps que sa structure morphologique ; il est question d'un processus compositionnel qui reflète une construction simultanée de la structure et du sens.

Devant les irrégularités morphologiques caractérisant le français, le choix de la méthode d'analyse morphologique et son intégration au SRI comme outil d'aide à l'utilisateur (en termes de reformulation des requêtes), se posent. À ce sujet, serait-il préférable d'opter pour une morphologie à base de règles et contrôlée au risque de ne pas avoir une large couverture des mots de la langue ? Un autre choix possible consiste à adopter une morphologie aléatoire ; certes, le recours à une morphologie aléatoire permettra d'avoir une large couverture de la langue, mais, avec ce dernier choix, les risques liés au phénomène de surgénération (déviations) sont bien réels.

Pour pouvoir extraire les variations morphologiques des termes de requêtes de base des documents rapportés, nous avons opté pour la deuxième méthode, c'est-à-dire l'application d'une morphologie aléatoire ou plutôt une démarche par troncation ; c'est une démarche qui consiste à ramener ''l'ensemble'' des variantes morphologiques à une seule forme. Autrement dit, la possibilité d'atteindre toutes les variantes à partir d'une seule racine. C'est la méthode que nous avons appliquée pour pouvoir extraire les variantes morphologiques des documents rapportés. C'est le cas, par exemple, de la racine ***port*** pour atteindre des variantes *export, exporter, exporte,*

*exportation, import, importation, etc.* Notons que, dans notre cas, cette démarche est appliquée d'une façon manuelle. De même, nous convenons qu'il s'agit d'une démarche risquée, c'est-à-dire que le rapatriement de variantes qui n'ont aucun lien sémantique avec les termes de base est fort probable.

La prise en compte des informations morphologiques dans un SRI a pour but de reconnaître les différentes formes d'un mot présentes dans les documents et les requêtes et de pouvoir les apparier. Ces variations morphologiques ont pour conséquence l'augmentation du taux de rappel. En effet, les variations linguistiques (morphologiques, syntaxiques et sémantiques) n'étaient pas prises en compte dans les SRI classiques. Dans ce contexte, les termes de la requête et des documents sont considérés comme de simples graphies, c'est-à-dire que chaque variation morphologique (d'un même mot) représente un mot distinct (une nouvelle forme équivaut à un nouveau mot). Une des solutions linguistiques utilisées pour pallier relativement cette variation, c'est le recours à des procédures comme la lemmatisation ; c'est une procédure qui consiste à ramener les mots à leurs formes canoniques (lemmes). Par exemple, ramener un verbe conjugué à sa forme infinitive, etc., et la racinisation ou *stemming*, une procédure qui consiste à ramener les différentes variantes morphologiques à une seule forme tronquée et commune à toutes les variantes (suppression des flexions et des suffixes) ; c'est le cas, par exemple, de la forme '*déménag*' (ou '*déménage*') qui regroupe, entre autres, les formes suivantes : *déménageur, déménageurs, déménagement, déménagements, déménage,* etc. (Loupy, 2001 ; Savoy, 1993 ; Paice, 1996 ; Jacquemin et Tzoukermann, 1999 ; Namer, 2000).

Namer (2000) traite amplement de cette procédure (racinisation) appliquée notamment à la morphologie du français. L'auteure montre l'importance et la pertinence d'intégrer aux tâches de recherche d'information des outils d'analyse morphologique comme celui qu'elle a développé 'FLEMM' (Namer, 2000). C'est un système fonctionnant à base de règles. L'auteure dresse des listes d'exceptions regroupant des mots irréguliers ou non assujettis aux règles morphologiques établies. De même, les cas de néologismes et fautes de frappe sont traités par FLEMM comme des mots réguliers, ce qui pose le problème de surgénération, c'est-à-dire qu'un mot bien formé est analysé correctement par le système même si le mot construit n'existe pas dans la langue ; c'est le cas par exemple des mots comme *recevation* et *éduquation*, ces deux

mots, n'existant pas dans le dictionnaire, peuvent être analysés correctement par le système sur des bases flexionnelles *recev* et *éduqu*. D'autres chercheurs sont partisans de cette voie dans l'analyse morphologique, c'est-à-dire la possibilité d'implanter des règles morphologiques de telle sorte que le système analyse correctement les mots bien formés malgré le risque de surgénération. Dans cette optique, Jacquemin et Tzoukermann (1997), ont implanté un système d'analyse automatique de la morphologie dérivationnelle du français. Ce système fonctionne à l'aide de transducteurs à états finis. Dans ce modèle, les formes lexicales sont construites au moyen de règles morpho-phonologiques. Ce modèle d'analyse permet de passer d'une représentation lexicale à celle de surface par le biais des opérations de concaténation (transduction d'un état à un autre). La formation des bases dérivationnelles par le système passe par trois étapes :

1. le calcul des bases flexionnelles à l'aide des règles morpho-phonologiques. Par exemple, les expressions /{reg_v91-1} : reç{verb}/ ; /{reg_v91-2} : recev{verb}/ représentent les bases du verbe *recevoir*) ;

2. la formation des variantes dérivationnelles à partir des bases flexionnelles. Par exemple, les bases du verbe *recevoir* (*reç- et recev-*) vont produire aussi bien les flexions verbales que les variantes dérivées (*recev-able, recev-eur, recev-abilité)* ;

3. l'écriture des règles allomorphiques destinées aux bases dérivationnelles manquantes. Par exemple, pour former des dérivés tels que *réception, récepteur, réceptacle*, etc. une autre base s'avère nécessaire ; la base adoptée ici est *récept-*, cette base est le résultat de la règle allomorphique selon laquelle le "*v*" se réécrit en "*pt*" dans le contexte où les deux caractères précédents sont "*ce*".

Par ailleurs, pour limiter les anomalies liées au phénomène de la surgénération, Jacquemin et Tzoukermann (1997) intègrent dans leur système une panoplie de filtres permettant de sélectionner les mots "possibles". Trois types de filtres sont adoptés par le système : un filtrage lexical (emploi du dictionnaire), un filtrage d'attestation sur corpus et un filtrage sémantique collocatif (contexte d'apparition des mots dérivés d'une même base).

Divers analyseurs morphologiques, en fonction de la langue traitée, intègrent des algorithmes d'analyse morphologique. L'un des algorithmes le plus connu est celui de Porter (1980). Il s'agit d'un algorithme de désuffixation mis au point pour l'anglais. Il est composé d'une cinquantaine de règles de désuffixation classées en plusieurs phases. Le mot analysé passe par toutes les étapes. La désuffixation intègrent des règles de recodage (transformation) ainsi que des règles de contexte. Ces dernières indiquent les conditions de suppression d'un suffixe (par exemple, "*troubling*" devient "*troubl*", alors que "*sing*" ne change pas, c'est-à-dire que "*ing*" n'est pas supprimé). La morphologie de cette langue (anglais) est relativement simple et l'application de ce genre d'algorithmes se prête bien à l'analyse morphologique. Ce qui n'est pas le cas pour le français, une langue à morphologie plus complexe. Ainsi, des algorithmes comme celui de Porter appliqué à l'anglais, sont rares, voire inexistants pour le français. À cet égard, les systèmes traitant de la morphologie du français, faute de ce type d'algorithmes, recourent à des outils qui s'avèrent souvent lourds et coûteux, il s'agit, par exemple, de l'utilisation des dictionnaires, de l'étiquetage morphosyntaxique des mots présents dans les documents, etc.

Pour revenir à la procédure de racinisation (*stemming*), divers travaux indiquent que l'intégration de cette procédure aux systèmes de recherche d'information a un impact sensible sur les performances de ces derniers, et cela aussi bien en indexation des documents qu'en extension des requêtes. Ces conclusions peuvent varier en fonction de la langue, des techniques et approches linguistiques adoptées. À cet égard, la racinisation peut être considérée comme un processus ad hoc, c'est-à-dire dédié à une langue et à des tâches déterminées ; c'est une procédure qui recourt, aussi, à des algorithmes approximatifs qui ne sont pas toujours basés sur une théorie linguistique ou morphologique viable et reconnue. De même, les heuristiques mises en œuvre n'ont pas une couverture large de la langue ; ce qui peut avoir comme conséquence la surgénération, la sous-génération, la non-reconnaissance de mots nouveaux (néologismes), association erronée des formes, ambiguïté liée aux glissements sémantiques des mots, etc. Généralement, ces phénomènes sont plus fréquents en présence des variations dérivationnelles que flexionnelles. Ainsi, notre intérêt, dans cette étude, portera plus particulièrement sur les variations dérivationnelles (leur fréquence, leur qualité, etc.). Par ailleurs, les variations d'ordre flexionnel seront prises en compte dans le processus d'analyse. Au sujet des variations flexionnelles,

Chieze (2000) avait traité, dans son mémoire de maîtrise, de l'apport des variations morphologiques flexionnelles au RI sur le Web où il était question de transformer, par le biais d'un processus automatique, les requêtes de base en requêtes booléennes suivi d'un enrichissement morphologique flexionnel.

Notre objectif consiste à vérifier la corrélation entre la fréquence des variations morphologiques et la pertinence des documents retournés. En d'autres termes, une grande partie de notre travail portera sur l'analyse des documents rapportés ; cette analyse va nous permettre, d'abord, d'établir le degré de pertinence des documents et, ensuite, d'extraire les variantes morphologiques et de calculer leur fréquence, etc. En fonction de leur type et de leur qualité, ces variantes seront considérées dans une perspective de reformulation de requêtes, une démarche qui ne sera pas implémentée dans le présent travail.

## 2.5 Conclusion

Nous avons présenté, dans le présent chapitre, le cadre théorique de notre travail de recherche. En ce sens, nous avons établi un bref historique qui met en exergue le domaine de la recherche d'information notamment dans le contexte du Web. En effet, le développement du Web qui s'est traduit par la présence d'une quantité considérable, voire illimitée de données de tout genre (texte, audio, vidéo et image), a rendu l'accès à l'information de plus en plus difficile ; cette expansion a été accompagnée par la conception de nouveaux outils de RI sur le Web, comme c'est le cas des moteurs de recherche d'information. Pour pouvoir aider l'utilisateur dans le processus de recherche d'information, les outils de recherche ont intégré dans leur système un ensemble de procédures aussi bien informatiques que linguistiques. Il est question, par exemple, de la conception de nouveaux algorithmes de plus en plus performants, de l'intégration de divers modèles de RI, de la prise en compte, de plus en plus fréquente, des connaissances linguistiques dans les SRI, etc.

Nous avons évoqué dans ce chapitre deux aspects importants associés à la RI sur le Web ; il s'agit de la notion de pertinence et l'apport des connaissances linguistiques, notamment morphologiques à la RI. À propos de la notion de pertinence, nous avons remarqué que c'est une notion qui reste vague et imprécise ; c'est une notion dont la définition diffère en fonction des critères liés à la fois au système et à l'utilisateur. Au

niveau du système, l'évaluation de la pertinence est plutôt algorithmique et au niveau de l'utilisateur, cette évaluation relève d'un processus cognitivo-linguistique (degré de satisfaction à un besoin d'information. De même, l'évaluation de la pertinence a accompagné l'évolution du domaine de la RI, et cela en passant, par exemple, d'une évaluation classique où la pertinence est d'ordre binaire (pertinent ou non pertinent), à une évaluation récente où la pertinence est exprimée en divers degrés (multivaluée).

Concernant les connaissances linguistiques, nous avons proposé un survol théorique lié à l'apport des connaissances linguistiques (syntaxe, morphologie et sémantique) à la RI. Un intérêt particulier a porté sur l'apport des connaissances morphologiques (dérivationnelles et flexionnelles), sujet de notre travail de recherche. En effet, l'apport du linguistique à la RI a fait l'objet de plusieurs études. Celles-ci ont accompagné l'évolution de ce domaine et ont permis la conception des procédures et des outils automatiques capables de traiter des connaissances linguistiques. L'intégration de ces connaissances aux SRI reste une nécessité pour pouvoir améliorer la qualité des résultats rapportés en termes de satisfaction des besoins de l'utilisateur. En effet, la plupart des travaux abordant ce sujet confirment l'idée selon laquelle l'impact des connaissances linguistiques sur la RI est bien réel, c'est-à-dire que l'intégration des connaissances linguistiques aux SRI permet d'améliorer la qualité des résultats retournés.

# CHAPITRE III

# MÉTHODOLOGIE

Pour vérifier la validité des questions de recherche émises précédemment, un travail expérimental s'avère nécessaire. Nous exposerons, dans le présent chapitre, la méthodologie et les outils adoptés pour mettre en œuvre notre expérimentation. Nous aborderons, d'abord, le choix du corpus retenu pour effectuer les tests ; ensuite, nous évoquerons les outils utilisés pour mener à bien notre expérimentation ; il est question, en particulier, du choix du corpus, du choix du moteur de recherche sur le Web et du choix de l'outil statistique. Nous présenterons, par la suite, les différentes étapes concernant la démarche expérimentale adoptée.

## 3.1 Choix du corpus

Pour mener notre recherche nous avons opté pour les requêtes TREC[4] (Text REtrieval Conference) conçues pour le français. Le corpus choisi comprend une liste de cinquante requêtes en français faisant partie de TREC 6, TREC 7 et TREC 8 (TREC, 2000). TREC représente une série de conférences qui consiste à tester des méthodes et des systèmes de recherche d'information en utilisant des collections de grandes tailles. Ces conférences organisées annuellement ont contribué au développement du domaine de la recherche d'information. Autrement dit, elles ont stimulé ce domaine (RI), cela en nous basant sur des collections de tests réalistes et en concevant une nouvelle méthodologie d'évaluation.

TREC fait partie de la lignée des projets promus par l'agence fédérale de recherche technologique de la défense américaine (DARPA). L'enjeu de ces conférences organisées annuellement, et cela depuis 1992, consiste à établir des protocoles et des méthodes d'évaluation communs et portables au sein de la communauté scientifique. Autrement dit, ce cycle de conférences permet de construire des ressources (corpus, protocoles expérimentaux, méthodes d'évaluation, etc.), qui sont utilisées et partagées

[4] http://trec.nist.gov/data/topics_noneng/index.html

par divers systèmes. L'intérêt principal de ce partage est de pouvoir capitaliser les connaissances et le savoir-faire dans le domaine de la recherche d'information (Lespinasse et al., 1999).

Deux raisons ont motivé le choix du corpus (TREC) ; la première est liée au fait que les requêtes sont en français ; elles ont été conçues dans le cadre des expériences multilingues. Le simple fait d'avoir des requêtes en français permet d'éviter des problèmes associés à des requêtes issues d'une traduction. La deuxième raison est liée à la valeur de ce corpus (TREC). Ce dernier est considéré comme un corpus de référence dans le domaine de la recherche d'information, c'est-à-dire que de nombreux chercheurs et diverses études en RI ont eu recours au corpus TREC pour tester et évaluer les méthodes adoptées ainsi que l'efficacité des systèmes en recherche d'information.

Pour les besoins de notre étude, nous nous sommes limité à tester une liste de cinquante requêtes (Tableau 3.1). Le nombre total de requêtes TREC (TREC 6, 7 et 8) en français atteint 108. Certes, Le nombre de requêtes choisi est sujet à discussion car il soulève la question de la représentativité du corpus par rapport à une collection aussi volumineuse que celle du Web, mais, s'agissant d'une étude exploratoire, nous considérons que ce nombre fixé à cinquante reste raisonnable pour mener la présente étude. De même, un autre critère lié à la longueur des requêtes a été pris en compte, c'est-à-dire que nous avons opté pour des requêtes qui contiennent deux mots pleins au minimum ; le corpus TREC comprend, en effet, des requêtes à un seul mot, c'est le cas, par exemple, de la requête CL5 (*acupuncture*). Ce choix est conditionné par l'environnement hétérogène du Web où plusieurs langues se côtoient ; cette diversité linguistique peut engendrer des interférences d'ordre linguistique notamment en présence de requêtes courtes. C'est le cas, du phénomène de l'homonymie où un mot qui a la même graphie dans plusieurs langues, mais pas forcément le même sens (cf. chapitre I). Ainsi, le choix des requêtes contenant plus d'un mot plein pourrait limiter le risque d'ambiguïté linguistique.

Comme les requêtes (TREC) ont été conçues pour une base de données spécifique, nous avons retenu notamment des requêtes qui restent, à notre avis, d'actualité dans le contexte d'une base évolutive et dynamique comme celle du Web. Les requêtes TREC sélectionnées n'ont pas été exécutées comme telles, c'est-à-dire sous forme

d'expressions où l'ensemble des éléments de la requête sont pris en compte (mots pleins et mots vides), mais plutôt sous forme de mots-clés ; en ce sens, nous avons supprimé les articles et les prépositions contenus dans les requêtes originales. Par exemple, la requête de base *législation sur la protection de la nature* sera transformée en mots-clés *législation protection nature* et exécutée comme telle, sans articles ni prépositions ; les coordonnants sont également exclus au moment de l'exécution de la requête. Rappelons que les prépositions jouent un rôle sémantique important en marquant les liens entre les mots, notamment les prépositions dites fortes ; leur suppression reste, donc, un pari risqué. Le fait d'opter pour les requêtes par mots-clés en supprimant les 'mots vides' (articles, prépositions et coordonnants) peut comporter des risques notamment au niveau du sens ; autrement dit, l'absence des déterminants des termes des requêtes peut engendrer des déviations sémantiques. Par exemple, la transformation de la requête de base (*exportation d'armes à la Turquie*) en mots-clés génère une requête sans déterminants, ni préposition (*exportation armes Turquie)*. La suppression des mots dits vides peut être l'origine d'ambiguïté sémantique ; ainsi, la requête en question peut avoir deux sens : la Turquie exporte des armes ou la Turquie importe des armes. Dans un autre exemple, l'ambiguïté sémantique est présente dans les requêtes sans suppression des déterminants ; c'est l'exemple de la requête 15 ([*déportation* [*des étrangers*] *en Autriche*]). Cette requête peut avoir le sens de : déportation des étrangers vers l'Autriche ou déportation des étrangers de l'Autriche vers une autre destination.

Par ailleurs, il est admis que le système de recherche d'information utilisé pour notre recherche, en l'occurrence Google, ne prend pas en compte les mots dits vides (déterminants, prépositions et coordonnants), probablement pas tous, dans le filtrage des documents. Ainsi, notre choix (mots-clés) résulte de deux facteurs, d'abord, le recours aux mots-clés est considéré comme la méthode la plus courante adoptée par l'utilisateur pour interroger un moteur de recherche ; ensuite, nos connaissances concernant le fonctionnement de Google sont limitées, ce qui ne nous permet pas de savoir comment sont traités réellement les mots vides.

## 3.2 Choix des outils

Pour mener notre étude, nous avons eu recours à divers outils notamment d'ordre logiciel. Le moteur de recherche d'information Google est l'outil adopté pour tester nos requêtes. Pour pouvoir établir la fréquence des mots présents dans les documents et la longueur des documents nous avons utilisé le Fréquencier[5] (version 8), outil logiciel d'ordre statistique, il permet de dresser une liste comportant le nombre et la fréquence des mots d'un texte. C'est un outil important dans la mesure où il nous permet d'épargner beaucoup de temps, en termes d'analyse de données. Les logiciels Word et Excel sont d'autres outils utilisés pour stocker et organiser les données recueillies. Pour effectuer les tests statistiques requis, nous avons eu recours au logiciel de statistiques SAS (Statistical Analysis System). Nous détaillons ces différents outils utilisés dans les sections ci-dessous.

### 3.2.1 Moteur de recherche

Les moteurs de recherche sont des outils incontournables pour accéder à l'information sur le Web. En d'autres mots, un moteur de recherche est un logiciel qui a pour fonction de rechercher des informations sur Web. Nombreux sont les moteurs de recherche oeuvrant sur le Web, c'est le cas, à titre d'exemple, de Yahoo, Altavista, Hotbot, Google, etc. La qualité de certains moteurs de recherche a eu un impact considérable sur le développement d'Internet. Ces moteurs ont accompagné cette évolution en permettant un accès illimité à un volume gigantesque de documents. Notre étude n'a pas pour objectif d'établir une étude comparative sur la performance des moteurs de recherche sur Internet. Mais il s'agit plutôt d'étudier la pertinence des documents rapportés en réponse à des requêtes. Celles-ci sont exécutées par le biais d'un seul moteur de recherche qui est l'un des moteurs les plus performants et les plus populaires : le moteur de recherche Google.

Google est le système de recherche d'information choisi pour effectuer les tests. Le choix de ce système tient compte de divers paramètres aussi bien techniques que fonctionnels. C'est un moteur de recherche[6] qui :

---

[5] http://www.lextutor.ca/freq/fr/

[6] http://www.ebsi.umontreal.ca/jetrouve/internet/mot_goo.htm, page consultée le 05-02-07)

- est généraliste et populaire ;
- a une large couverture du Web ;
- fonctionne selon le modèle booléen (par défaut) permettant le filtrage et le tri des documents ; l'opérateur AND est inséré implicitement entre les mots-clés de la requête ;
- inclut dans ses recherches des documents d'autres formats que les pages Web ; il s'agit de fichiers aux formats PDF, PS, PPT, etc. ;
- possède une interface simple et facile d'utilisation (page d'accueil) ; l'interface de Google permet une visualisation complète offrant à l'utilisateur la possibilité d'effectuer sa recherche avec plus de rapidité;
- présente dans les résultats rapportés des segments de textes contenant les termes de la requête de l'utilisateur ;
- se caractérise par sa rapidité d'exécution ; la moyenne de ce moteur de recherche est de 0,29 secondes par recherche ;
- privilégie les pages dans lesquelles les termes de la requête apparaissent proches les uns des autres ;
- calcule et affiche les pages en fonction de leur popularité ;
- ne tient pas compte des accents et de la casse ;
- offre la possibilité de choisir la langue de recherche et le nombre de résultats à afficher (de 10 à 100 par page) ;
- comprend un dictionnaire qui propose une autre orthographe si un mot est mal orthographié ;
- suggère 10 expressions populaires (Google Suggest) au moment de taper un terme de la requête ; une autre nouveauté de Google (2007), la recherche apparentée ; il consiste à proposer des suggestions de requêtes apparentées à la requête initiale, un autre moyen pour aider l'utilisateur à trouver "rapidement" l'information recherchée.
- etc.

Malgré ces caractéristiques, le moteur de recherche Google ne livre pas tous ses secrets. En effet, les mécanismes de son fonctionnement, notamment ceux liés au tri des documents et l'algorithme utilisé, ne relèvent pas du domaine public. L'environnement évolutif et non contrôlé du moteur de recherche adopté permet à notre étude d'être plus réaliste et proche de la situation à laquelle est confrontée un utilisateur dans sa recherche d'information sur le Web. Ainsi, malgré les quelques zones d'ombre caractérisant Google, nous pensons que c'est l'outil de recherche le plus adéquat parmi ceux disponibles pour mener notre étude. Cela ne veut pas dire que les autres moteurs de recherche (Yahoo, Altavista, Hotbot, etc.) ne réunissent pas les caractéristiques adéquates pour notre étude ; ils ont aussi des qualités et des zones d'ombre comme Google.

Nous présentons, ci-dessous (Figure 3.1), un schéma décrivant, d'une façon globale, le fonctionnement d'un moteur de recherche sur le Web, comme c'est le cas de Google. Ce schéma est une adaptation de celui de Bonnel et Moreau (2005).

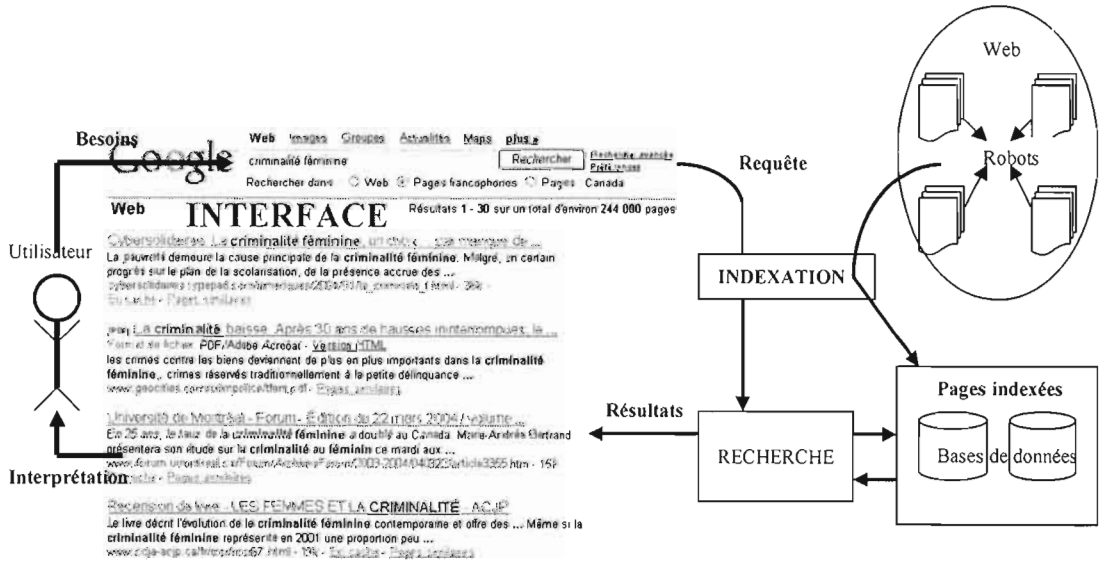



**Figure 3.1 : Schéma du fonctionnement d'un moteur de recherche**

Deux niveaux interagissent lors de la recherche d'information, l'utilisateur et le SRI. L'utilisateur exprime son besoin d'informations par le biais d'une interface du moteur de recherche Google, phase d'interrogation. Au niveau du SRI, nous distinguons trois modules : collecte des données, indexation et recherche. La collecte des données se fait au moyen de robots logiciels qui parcourent le Web d'une façon permanente et recensent les pages (adresses) visitées. L'indexation est un module qui consiste à stocker le contenu (mots-clés, adresses, etc.) des pages visitées par les robots dans un index, c'est-à-dire une base de données des pages indexées. Le dernier module est celui de recherche ; il consiste, en réponse à la requête formulée par l'utilisateur, à interroger la base de données (index) et à présenter les résultats à l'utilisateur. Ce dernier va interpréter et évaluer les résultats rapportés par le système.

### 3.2.2 Autres outils

Nous nous servons également des logiciels Excel et Word pour stocker et organiser les données recueillies après l'exécution des requêtes et l'évaluation des résultats. Le stockage des données, notamment les pages Web, est nécessaire pour notre étude ; cela est dû au contexte évolutif et non contrôlé du Web. Par ailleurs, pour calculer la longueur des documents rapatriés nous avons eu recours à un outil statistique appelé Fréquencier (version 8) disponible sur le Web[7] et à accès libre. C'est un outil qui permet d'établir une liste de fréquence des mots et le nombre de mots constituant un texte. Nous présentons, Figure 3.2 ci-dessous, une copie d'écran représentant l'interface de ce logiciel. Le rapatriement des documents associés à chaque requête se fait individuellement en enregistrant chaque document rapporté dans un dossier (base de données). L'accès ultérieur à ces documents se fait hors connexion.

[7] http://www.lextutor.ca/freq/fr/

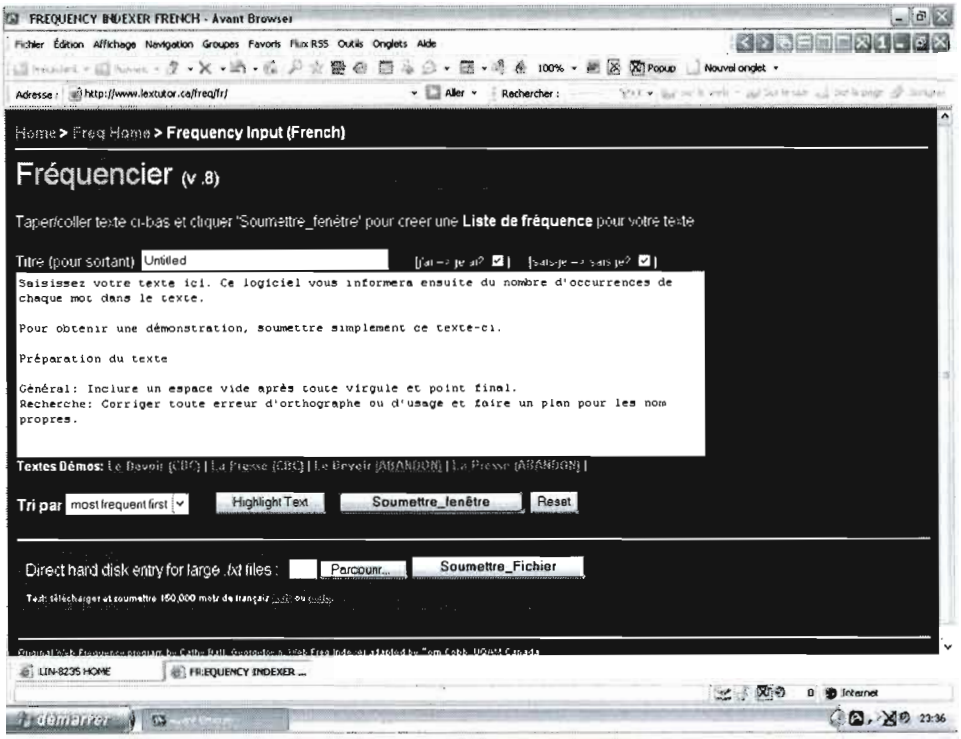



**Figure 3.2 : Page écran de l'interface du Fréquencier**

Les tests statistiques sont faits par le biais du logiciel de statistiques SAS[8] (Statistical Analysis System). Le SAS est un outil logiciel qui permet de traiter, analyser et générer des données informatiques indépendamment de leur format ou de leur plate-forme (support). Ce logiciel comprend différents modules associés aux différentes opérations (gestion de données statistiques, analyse descriptive, modélisation, etc.). Le choix des tests statistiques et leur réalisation ont été effectués par M. Bertrand Fournier de l'équipe du SCAD (Service de Consultation en Analyse de Données de l'UQAM).

## 3.3 Test statistique

Mener une recherche scientifique passe inévitablement par l'émission d'hypothèses issues de nos connaissances théoriques. Une fois les hypothèses établies, vient le processus de vérification. Celui-ci consiste à récolter des données empiriques en vue

[8] http://www.google.ca/search?hl=fr&q=sas+introduction&meta=lr%3Dlang_fr

d'apporter des informations sur l'acceptabilité de telle ou telle hypothèse. L'acceptation d'une hypothèse dépend notamment de l'objectivité avec laquelle le chercheur fait ses choix. Atteindre cette objectivité passe par l'application des procédures objectives. Le choix du test statistique est une étape importante dans le processus de signification des données et la vérification des hypothèses.

Le test statistique appliqué aux échantillons de notre recherche est celui de Jonckheere-Terpstra, il s'agit d'un test non paramétrique (Hollander et al., 1973). Ce dernier est facile à appliquer ; cette simplicité découle notamment du remplacement des valeurs obtenues par des variables alternatives ou par des rangs ; en ce sens, les valeurs observées sont rangées selon un ordre croissant. Le choix de ce test est lié à la nature des données recueillies. Il s'agit d'échantillons aléatoires, variables et indépendants les uns des autres. Ce test, qui est disponible sous sa version exacte en SAS® (Statistical Analysis System, version 9.01), consiste à rejeter l'hypothèse nulle d'absence d'association par rapport à d'autres possibilités ordonnées.

Avec le test Jonckheere-Terpstra appliqué aux échantillons aléatoires simples, nous postulons comme hypothèse nulle que les distributions des populations d'origine sont semblables ; autrement dit, les échantillons sont uniformes et proviennent tous de la même population :

$$\mathbf{H_0} : \mu_1 = \mu_2 = \ldots = \mu_K$$

Par ailleurs, nous posons l'hypothèse alternative selon laquelle il existe un ordre a priori croissant. Dans cet ordre, au moins une des inégalités est stricte :

$$\mathbf{H_1} : \mu_1 \leq \mu_2 \leq \ldots \leq \mu_K$$

Dans le contexte de notre étude, le niveau de pertinence (0, 1, 2, 3) est considéré comme quatre mesures ordonnées (nous y reviendrons à la section suivante). Cette démarche permet de vérifier s'il existe un ordre a priori croissant des différentes mesures de variation linguistique (*termes de la requête, variantes dérivées, variantes fléchies et longueur du texte*). Le test de Jonckheere-Terpstra est appliqué avec le principe selon lequel une déviation de l'hypothèse nulle $\mathbf{H_0}$ (où les distributions d'origine sont semblables), serait dans la direction alternative ; celle-ci suppose une

forte présence de variation linguistique qui serait associée à un accroissement du niveau de pertinence des documents.

## 3.4 Démarche expérimentale

Une fois la liste des requêtes établie et le moteur de recherche choisi, nous entamons le processus expérimental de notre étude. Notre démarche expérimentale comprend les étapes suivantes :

### Étape 1

**Paramétrage du moteur de recherche** : nous avons évoqué précédemment plusieurs caractéristiques du moteur de recherche Google. Celui-ci propose à l'utilisateur la possibilité de paramétrer l'outil de recherche. Parmi les options proposées par Google, citons notamment le choix de la langue de recherche. Compte tenu que notre étude porte sur la RI en français, nous utilisons le mode avancé du moteur de recherche Google pour paramétrer notre recherche en conséquence, c'est-à-dire la récupération de documents en français. Une autre option nous permet de choisir le nombre de documents à afficher par page. Pour les besoins de notre étude nous avons opté pour l'affichage de 30 documents par page. Parmi les 30 documents affichés, nous n'en gardons que 20. Autrement dit, nous examinons uniquement les 20 premiers documents, cela en excluant les doublons, les documents inexistants et tout autre document ne comportant pas de texte (vidéos, images, etc.).

### Étape 2

**Exécution des requêtes :** à cette étape, nous procédons à l'exécution des 50 requêtes TREC contenues dans notre corpus (Tableau 3.1 ci-dessous). Précisons que les mots des requêtes sont exécutés avec les accents. Toutefois, nous n'avons pas pris en compte les articles, les prépositions et les coordonnants (''mots vides'') présents dans les requêtes de base. Cela malgré l'intérêt que peuvent avoir ces unités linguistiques, et surtout les prépositions dites fortes, sur le plan sémantique. Comme il est admis, le Web est un environnement dynamique et non contrôlé ; ainsi, pour minimiser les risques liés à ce contexte (Web), nous avons pris soin de pouvoir effectuer nos tests

dans un laps de temps raisonnable. C'est pour cette raison aussi que nous rapatrions les documents retenus.

**Tableau 3.1 : La liste des 50 requêtes TREC**

| N° | Requête | N° | Requête |
|---|---|---|---|
| 1 | La culture écologique | 26 | Terrorisme international |
| 2 | anti-sémitisme en Allemagne après 1945 | 27 | Maltraitance des enfants |
| 3 | Exportation de médicaments dangereux | 28 | Exploitation économique du fond marin |
| 4 | Criminalité féminine | 29 | Exportation d'armes à la Turquie |
| 5 | Politique extérieure de l'Autriche | 30 | Les débris spatiaux |
| 6 | Conditions de vie des immigrants | 31 | Séparatistes catalans et galiciens |
| 7 | Production minimale au Japon | 32 | Coopération internationale des entreprises |
| 8 | Extrémisme de droite et racisme | 33 | Réfugiés de Bosnie |
| 9 | Jeunesse et politique | 34 | Accidents dans l'industrie minière |
| 10 | Criminalité des adolescents | 35 | Transport public local |
| 11 | Charte sociale européenne | 36 | La pollution causée par l'automobile |
| 12 | Violence des adolescents | 37 | Vitesse sur les autoroutes en Suisse |
| 13 | Législation sur la protection de la nature | 38 | L'homosexualité et la loi |
| 14 | La politique économique de la Slovénie | 39 | Processus de paix au Moyen-Orient |
| 15 | Déportation des étrangers en Autriche | 40 | Effets du chocolat sur la santé |
| 16 | Les communistes au Parlement Européen | 41 | Ouvriers étrangers en Europe |
| 17 | Destruction de la forêt tropicale en Amérique du Sud | 42 | Conséquences de la réunification allemande |
| 18 | Famine au Soudan | 43 | Conversion de la dette pour la Pologne |
| 19 | Protection des animaux | 44 | Statut militaire de l'Allemagne unifiée |
| 20 | Limitations des importations de l'UE | 45 | Lutte contre la corruption |
| 21 | Unité franco-allemande | 46 | Protection de l'environnement au sein des entreprises |
| 22 | Immigration et racisme | 47 | Maintien de la paix par l'OUA |
| 23 | les accidents de la route | 48 | L'industrie européenne du film |
| 24 | L'éducation sexuelle | 49 | Normes de protection professionnelle |
| 25 | Les effets de la déforestation | 50 | Traitement des déchets nucléaires |

**Étape 3**

**Rapatriement des documents :** à ce stade, nous commençons, d'abord, par copier coller dans Word la première page de Google contenant les résultats liés à chaque requête, c'est-à-dire les 30 documents (réponses) rapportés ; nous rapatrions ensuite chaque document retenu, parmi ceux rapportés, et cela jusqu'à atteindre 20 documents requis par requête. Le rapatriement des documents s'est fait individuellement ; autrement dit, chaque document retenu est enregistré dans un dossier en mode hors connexion pour une analyse future. Les documents récupérés peuvent être de formats divers : Word, PDF, PPT, HTML, etc. Rappelons que les doublons et les documents inexistants sont exclus, et par le fait même ils ne comptent pas dans l'analyse. Est exclu également tout document ne contenant pas du texte (vidéos, images, etc.). Malgré les exclusions nous avons toujours eu les 20 documents par requêtes. De même, nous enregistrons l'adresse URL de chaque document rapatrié. La requête 4 (*criminalité féminine*) nous servira d'exemple pour illustrer notre démarche. Nous présentons ci-dessous (Figures 3.3, 3.4 et 3.5) des copies d'écrans représentant une partie de cette démarche :

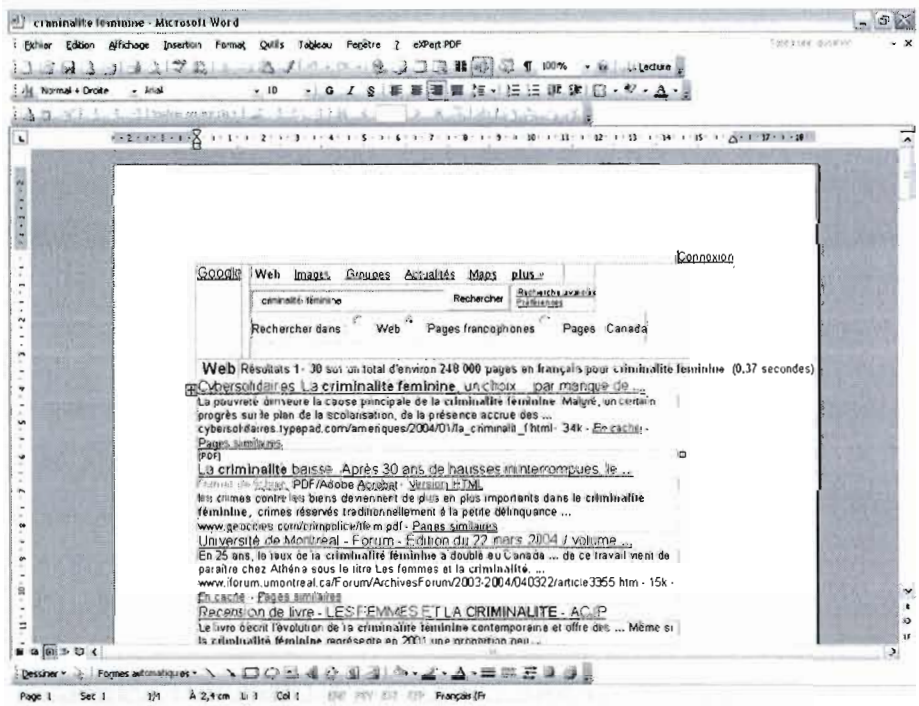



**Figure 3.3 : Résultats rapportés après l'exécution d'une requête**

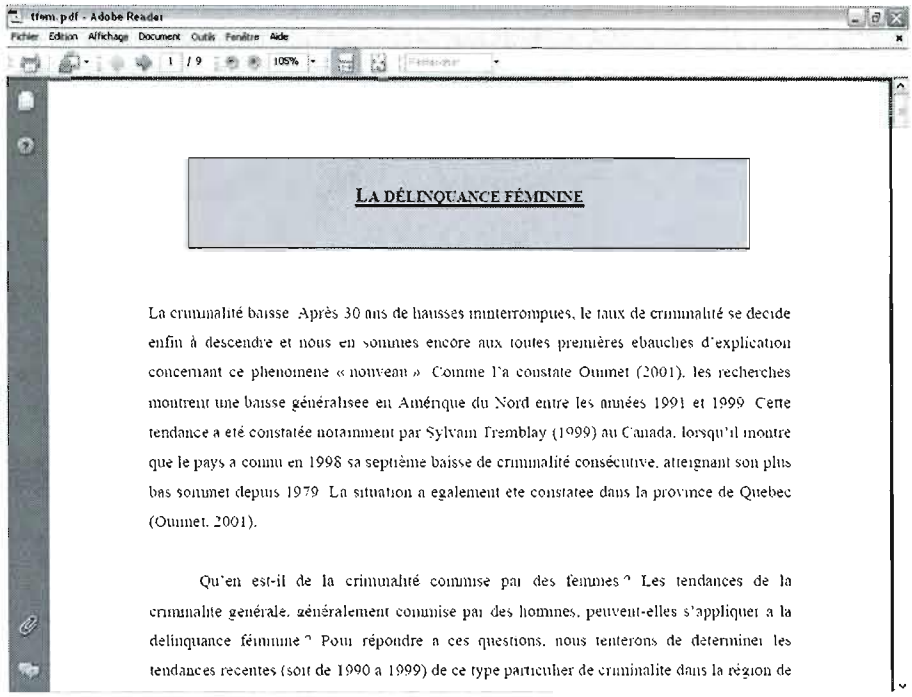



**Figure 3.4 : Document récupéré après l'exécution d'une requête**

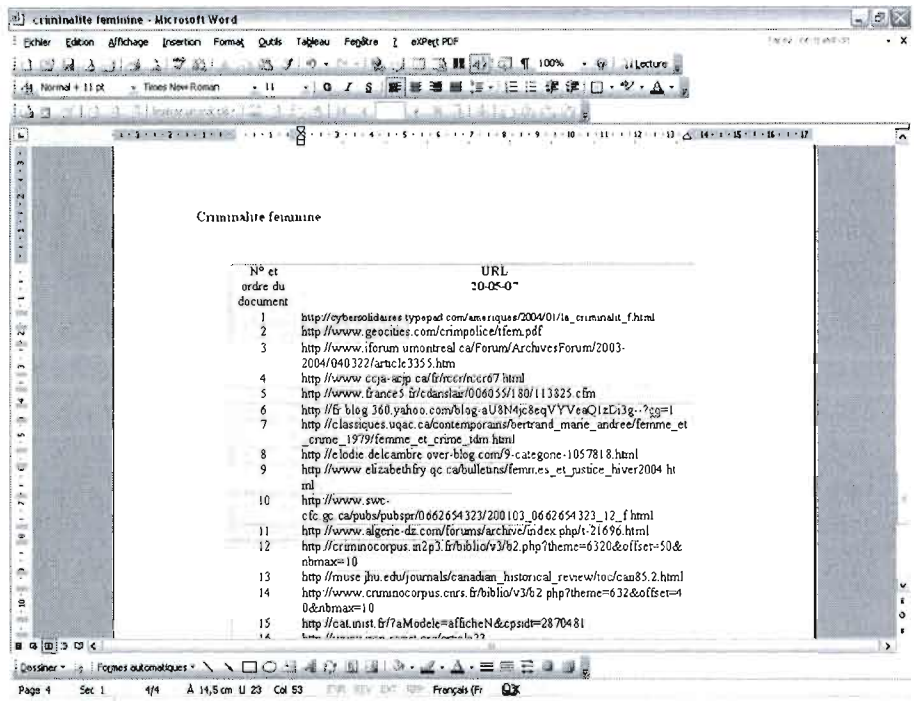



**Figure 3.5 : Adresses URL des documents récupérés**

**Étape 4**

**Processus d'évaluation des documents :** à cette étape de l'expérimentation, nous procédons à l'évaluation de chaque document rapatrié. Il s'agit ici d'établir, outre le numéro de la requête et le numéro assigné à chacun des documents, une évaluation de la pertinence. La méthode d'évaluation de la pertinence des documents est basée sur une échelle de pertinence à 4 paliers, de 0 à 3 (0 = non-pertinent ; 1 = seul le BI est mentionné ; 2 = partiellement pertinent ; 3 = pertinent). Cette méthode multivaluée est inspirée de Chignell et al. (1999) et Sormunen (2002). Ainsi, la valeur 3 représente un document pertinent et implique que ce dernier évoque d'une façon approfondie le BI ou couvre plusieurs facettes du BI. La valeur 2 correspond aux documents partiellement pertinents ; il est question de documents qui évoquent seulement quelques facettes du BI (un document contenant un lien vers un autre document pertinent sera considéré comme partiellement pertinent). La valeur 1 correspond à un document où le BI est seulement mentionné, et cela sans avoir d'information traitant de ce BI. La valeur 0 représente un document non pertinent où le BI n'est pas mentionné. Pour évaluer les documents rapatriés, nous nous sommes appuyé sur la partie narrative des requêtes TREC[9] (les requêtes TREC avec les sections descriptives et narratives sont présentées à l'Annexe A). Pour illustrer notre propos, nous présentons, à l'Annexe B, l'évaluation des extraits de documents rapportés en fonction de leur degré de pertinence ; nous prenons comme exemple la requête 4 *criminalité féminine*. Nous concédons qu'une part de subjectivité est inévitable au moment d'établir le degré de pertinence des documents rapatriés ; l'idéal dans ce contexte aurait été de déléguer la tâche d'évaluation de la pertinence à des utilisateurs externes, mais, pour des raisons liées au temps et au coût, c'est nous-même qui avons évalué la pertinence des documents.

Nous présentons ci-dessous un exemple de requête TREC (*criminalité féminine*) avec les parties descriptive et narrative.

[9] http://trec.nist.gov/data/topics_noneng/index.html

**Tableau 3.2 : Parties *Narrative* et *Descriptive* d'une requête TREC**

```
<top>
<num> Number: 1

<F-title> Criminalité féminine

<F-desc> Description :
Quels sont les rapports, affaires, recherches
empiriques et études disponibles au sujet de la
criminalité et de la délinquence des femmes ?

<F-narr> Narrative :
Les documents concernés abordent les problèmes
particuliers de la criminalité des femmes, y compris
les problèmes de la réinsertion sociale et de
l'emprisonnement des femmes. Les études historiques
(avant 1945), les statistiques générales, les
réflexions sur la philosophie du droit et le
terrorisme ne sont pas pris en compte.
</top>
```

L'exemple cité ci-dessus (Tableau 3.2) représente une requête TREC. Celle-ci est composée d'un titre, d'une section descriptive et d'une section narrative. Le titre désigne les termes de la requête originale, ici *criminalité féminine*. La section descriptive établit la thématique sur laquelle porte la requête de base. La section narrative est d'ordre informatif, elle donne des informations et des précisions sur la thématique de la requête ; c'est précisément en nous basant sur la section narrative associée à chaque requête que nous établissons la pertinence des documents rapportés. Nous constatons également que, dans quelques cas, la section narrative peut être une source d'ambiguïté au niveau thématique ; autrement dit, il y a un décalage entre le titre de la requête originale et la partie narrative qui lui est associée ; ce décalage thématique fait en sorte que les résultats rapportés ne correspondent pas aux besoins exprimés par les termes de la requête. C'est le cas, par exemple, de la requête de base ***production minimale (au) Japon*** (testée le 06 juin 2007). La section narrative liée à cette requête est la suivante :

> « *Les documents pertinents traitent des questions particulières de la production minimale au Japon. Ces documents peuvent inclure la production* ***minimale*** **[mot corrigé]** *présentée comme concept de gestion et offrant des*

*exemples concrets d'applications et d'études au Japon. Les considérations générales sur la production minimale et les études spécifiques à d'autres pays ne sont pas prises en compte.* »

En se référant à cet énoncé narratif, nous nous apercevons que la chance d'avoir des documents répondant aux exigences thématiques exprimées dans la section narrative est minime. Cela apparaît d'une façon évidente dans le score de pertinence obtenu par ce genre de requêtes, c'est-à-dire que cette requête enregistre un score de pertinence des documents très faible (cf. chapitre IV, Tableau 4.2). L'inadéquation entre les termes de la requête et la section narrative pourrait être liée, entre autres, à la spécificité des requêtes et à la nature du Web. En effet, les requêtes en question ont été conçues pour un cadre de recherche documentaire bien précis (corpus d'articles de journaux de la Suisse romande) ; autrement dit, ces requêtes, à l'origine, n'étaient pas destinées à l'univers du Web. Une autre raison de cette inadéquation serait liée à la nature du Web ; celui-ci est considéré comme un espace dynamique et évolutif ; alors, vu qu'une partie de ces requêtes n'est plus d'actualité, la possibilité d'avoir des résultats satisfaisants reste infime.

**Étape 5**

**Analyse des documents :** après avoir évalué la pertinence des 20 premiers documents liés à chaque requête, nous procédons à l'analyse des autres variables. Cette analyse consiste à extraire les termes de la requête, les variantes dérivées, les variantes fléchies, les expressions, et leur fréquence. Nous établissons également la somme des termes et des variantes ainsi que la longueur des documents. Toutes les données associées à ces variables sont consignées dans un tableau[10]. Pour extraire les termes, les variantes et leur fréquence, et pour établir la longueur des documents, nous nous sommes basé sur le logiciel Fréquencier (version 8). Le processus d'analyse des documents s'effectue de la façon suivante : nous soumettons le document rapatrié au logiciel Fréquencier (Figure 3.6). Ce dernier analyse le document et donne comme réponse une liste de fréquence des mots contenant le document ainsi que d'autres éléments statistiques comme le nombre total de mots du document, l'ordre des mots, fréquence des mots en termes de pourcentage, etc. Pour aller chercher les termes et les

[10] Vu la grande quantité de données consignées dans le document qui est volumineux, nous ne le mettons pas en annexe.

variantes dans la liste des mots générée par le Fréquencier, nous utilisons l'option recherche (Ctrl + V), c'est une opération qui peut être considérée comme manuelle (Figure 3.7). Les données brutes recueillies sont enregistrées dans un tableau (Figure 3.8 ci-dessous).

Pour pouvoir atteindre (extraire) les termes et les variantes présents dans les documents rapportés, nous appliquons une démarche morphologique qui consiste à recourir à une racine ou une forme tronquée, idéalement, commune à toutes les variantes morphologiques (notamment dérivées). L'intérêt de cette technique consiste à offrir la possibilité de trouver des formes différentes au niveau morphologique et proches au niveau sémantique. Ainsi, le choix de la troncation est une façon non contrainte permettant d'avoir une large couverture des termes de la langue (élargissement de la famille lexicale liée aux mots). Le recours à plus d'une forme tronquée est possible si besoin est. Par exemple, pour extraire les termes et les variantes associés à la requête 4 (*criminalité féminine*), nous utilisons les formes tronquées *crim*, *fem* et *fém*. Cette démarche morphologique, qui peut être qualifiée d'aléatoire, peut avoir des inconvénients liés notamment aux glissements sémantiques. Autrement dit, cette façon de faire peut entraîner la prise en compte de variantes, notamment dérivées, qui diffèrent complètement, d'un point de vue sémantique, des termes de base. C'est le cas, par exemple, de la racine (ou forme tronquée) ***port*** utilisée pour atteindre les variantes du mot de base ***importations*** de la requête 20 (*limitations (des) importations (de) (l') UE*). Les déviations enregistrées concernant cette racine sont nombreuses, nous citons, entre autres, *portée, ports, comporter, supporter, apport, portantes, etc.* Nous relèverons d'autres cas lors de l'analyse des résultats obtenus (chapitre IV).

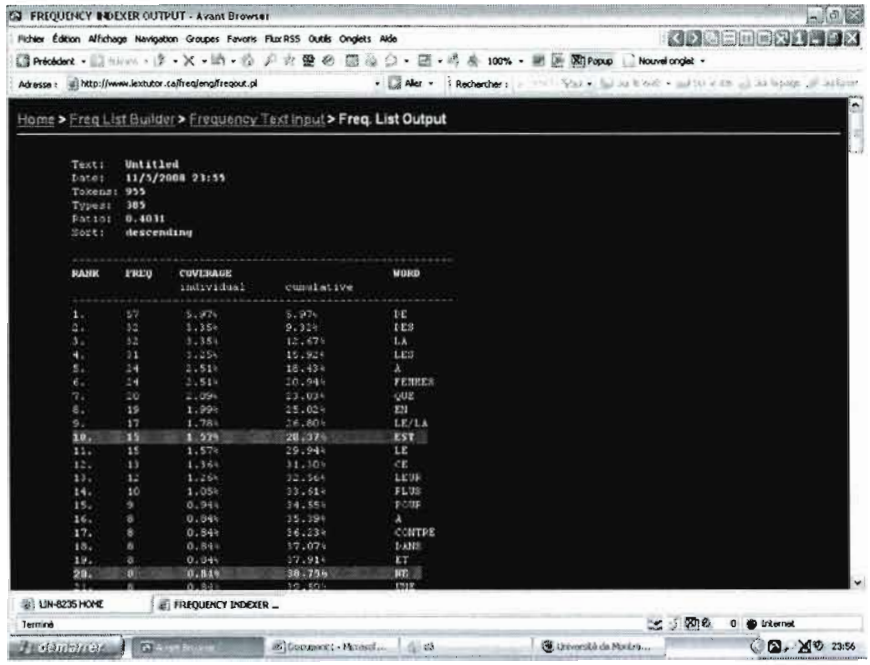



**Figure 3.6 : Résultat d'analyse d'un document par le Fréquencier**

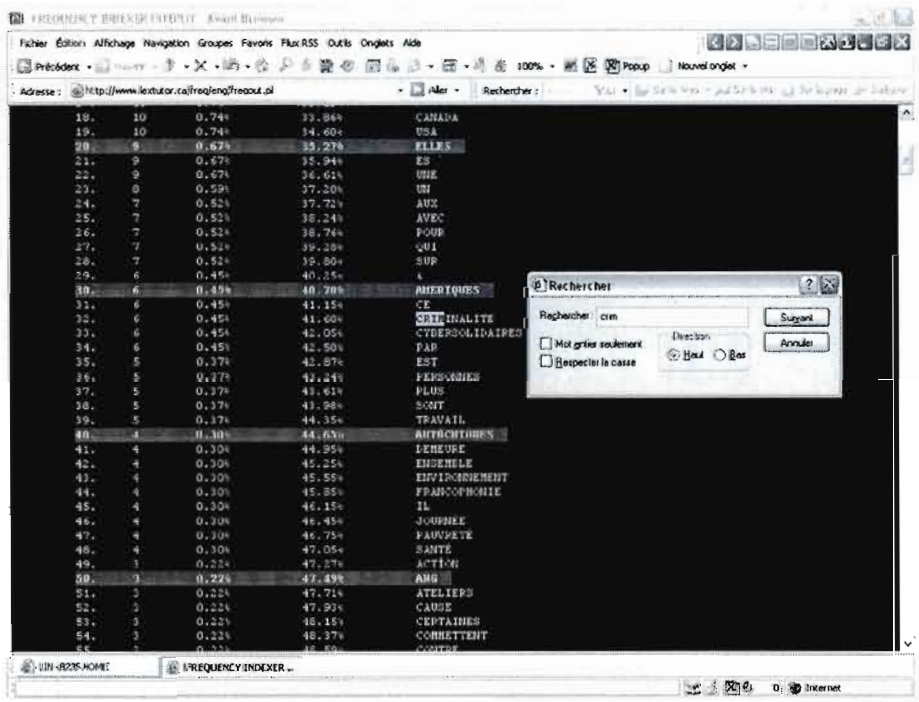
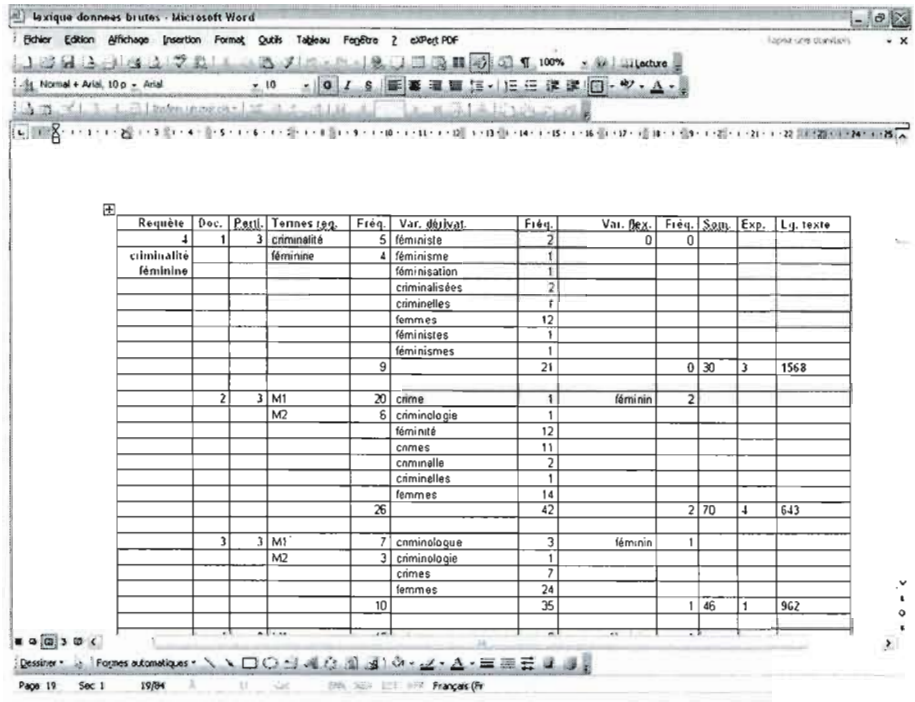



**Figure 3.7 : Recherche des termes et des variantes, *criminalité féminine***



| Requête | Doc. | Pertl. | Termes req. | Fréq. | Var. dérivat. | Fréq. | Var. flex. | Fréq. | Somm. | Exp. | Lg. texte |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 1 | 3 | criminalité | 5 | féministe | 2 | 0 | 0 | | | |
| criminalité | | | féminine | 4 | féminisme | 1 | | | | | |
| féminine | | | | | féminisation | 1 | | | | | |
| | | | | | criminalisées | 2 | | | | | |
| | | | | | criminelles | 1 | | | | | |
| | | | | | femmes | 12 | | | | | |
| | | | | | féministes | 1 | | | | | |
| | | | | | féminismes | 1 | | | | | |
| | | | | 9 | | 21 | | 0 | 30 | 3 | 1568 |
| | 2 | 3 | M1 | 20 | crime | 1 | féminin | 2 | | | |
| | | | M2 | 6 | criminologie | 1 | | | | | |
| | | | | | féminité | 12 | | | | | |
| | | | | | crimes | 11 | | | | | |
| | | | | | criminelle | 2 | | | | | |
| | | | | | criminelles | 1 | | | | | |
| | | | | | femmes | 14 | | | | | |
| | | | | 26 | | 42 | | 2 | 70 | 4 | 643 |
| | 3 | 3 | M1 | 7 | criminologue | 3 | féminin | 1 | | | |
| | | | M2 | 3 | criminologie | 1 | | | | | |
| | | | | | crimes | 7 | | | | | |
| | | | | | femmes | 24 | | | | | |
| | | | | 10 | | 35 | | 1 | 46 | 1 | 962 |

**Figure 3.8 : Tableau des données brutes recueillies (extrait)**

La Figure 3.8 ci-dessus présente les données recueillies après l'analyse des documents rapportés. Ces données représentent la requête, le numéro du document, le degré de pertinence, les termes de la requête et leur fréquence, les variantes dérivées et leur fréquence, les variantes fléchies et leur fréquence, la somme des termes et des variantes, les expressions et leur fréquence, et la longueur du texte.

## 3.6 Conclusion

Le présent chapitre passe en revue les différents choix et outils adoptés pour mener à bien le processus expérimental. Le choix du corpus de requêtes TREC (50 requêtes) est conditionné par le fait qu'il s'agit d'un corpus en français, ce qui nous épargne toute anomalie résultant d'un corpus traduit. Il s'agit d'un corpus de référence qui fait autorité dans le domaine de RI. Un autre outil important utilisé pour mener notre étude est celui du moteur de recherche sur le Web, Google. Le choix porté sur Google n'est pas aléatoire, il est plutôt lié aux qualités techniques et fonctionnelles qui font que ce moteur de recherche est l'un des plus généralistes, performants et populaires.

D'autres outils logiciels ont été adoptés, il s'agit du logiciel le Fréquencier (version 8) ; ce dernier permet une analyse statistique rapide des documents (fréquence des mots, nombre de mots dans le document, etc.), ce logiciel nous facilite également l'extraction des termes de la requête et des variantes morphologiques, et leur fréquence. Pour stocker et organiser les données recueillies, nous avons utilisé les logiciels Word et Excel. Finalement, nous avons fait appel au logiciel d'analyse statistique SAS (Statistical Analysis System) pour réaliser les tests statistiques requis.

L'intérêt principal de notre démarche méthodologique consiste à essayer de trouver des réponses aux questions de recherche que nous nous sommes posées. Rappelons que l'objectif de notre démarche expérimentale consiste à faire ressortir les variations morphologiques (des termes de base) des documents rapportés. En fonction de la fréquence (et de la qualité) de ces variantes, notamment dérivées, nous allons vérifier s'il y a une corrélation entre cette fréquence et la pertinence des documents rapportés. Sur la base de ces variantes et leur qualité, nous aurons une réponse sur la possibilité de considérer ces variantes dans une perspective de reformulation de requêtes. Le chapitre suivant va nous permettre d'examiner ces hypothèses et de vérifier leur validité.

# CHAPITRE IV

# RÉSULTATS DE L'ANALYSE

Le présent chapitre présentera les résultats obtenus. Nous commencerons par présenter les résultats bruts obtenus lors de l'analyse des données ; ces résultats englobent l'évaluation de la pertinence des documents associés à chacune des 50 requêtes ainsi que la fréquence des variables enregistrée à chaque requête. Un tableau récapitulatif établira une lecture globale, toutes requêtes confondues ; c'est une démarche qui consiste à calculer les pourcentages liés à chacune des variables en fonction du niveau de pertinence et en prenant en compte la longueur du document. Ensuite, nous introduirons les résultats de l'analyse statistique ; celle-ci dressera le bilan des corrélations entre la pertinence et les variables (*termes de la requête, variantes dérivées, variantes fléchies et longueur des documents*). Pour donner un sens à ces corrélations, nous présenterons les résultats du test statistique de Jonckheere-Terpstra. Enfin, nous établirons un bilan des résultats obtenus et celui de l'analyse linguistique.

## 4.1 Présentation des résultats

La présente section présente les résultats obtenus. Il s'agit des résultats bruts ainsi que ceux obtenus par le biais du logiciel de statistiques SAS (Statistical Analysis System). Rappelons que le test statistique de Jonckheere-Terpstra nous a été proposé par M. Bertrand Fournier du SCAD (Service de Consultation en Analyse de Données de l'UQAM). Il s'agit, dans cette section, de relever les points les plus saillants des résultats bruts et ceux issus du test statistique. Notre objectif consiste à étudier le score de pertinence en corrélation avec les autres variables (*termes de la requête, variantes dérivées, variantes fléchies et longueur du texte*). Nous reviendrons sur le score de pertinence lors de la présentation des résultats liés aux corrélations entre le degré de pertinence et les différentes variables citées auparavant.

### 4.1.1 Évaluation des documents rapportés

Le tableau ci-dessous (Tableau 4.1), présente les résultats liés au score de pertinence obtenus pour nos 50 requêtes où chaque requête comprend 20 documents. Autrement

dit, le Tableau 4.1 établit, pour chaque requête, le nombre de documents appartenant à chacun des niveaux de pertinence ainsi que la valeur correspondante en pourcentage. Rappelons que nous avons adopté une méthode d'évaluation multivaluée où quatre degrés de pertinence, allant de 0 à 3, sont pris en compte. La valeur 0 représente un document non pertinent où le BI n'est pas mentionné. La valeur 1 correspond à un document peu pertinent où le BI est seulement mentionné, et cela sans avoir d'informations traitant de ce BI. La valeur 2 correspond aux documents partiellement pertinents, il est question de documents qui évoquent seulement quelques facettes du BI (un document contenant un lien vers un autre document pertinent sera considéré comme partiellement pertinent). La valeur 3 représente un document pertinent et implique que ce dernier évoque d'une façon approfondie le BI ou couvre plusieurs facettes d'un BI.

Le calcul de la somme de toutes les valeurs obtenues (Tableau 4.1), à chaque niveau de pertinence, nous révèle que le score de pertinence est quasi-similaire entre les trois premiers degrés de pertinence (0, 1 et 2) ; en termes de pourcentage, ce score atteint 26 % à chacun des trois niveaux de pertinence en question (264 documents au niveau 0, 265 documents au niveau 1, 260 documents au niveau 2). Toutefois, ce score est légèrement moins élevé (-5 %) au niveau de la pertinence 3, car il atteint 21 % (211 documents sur 1000).

Par ailleurs, les résultats obtenus nous révèlent la présence de plusieurs écarts, c'est-à-dire que le score de pertinence associé à chaque requête indique des variations diverses entre les différents degrés de pertinence. Par exemple, pour le degré 0 de pertinence, le pourcentage varie entre **5** et **80 %** ; le nombre de documents ayant le degré de pertinence 1 varie entre **10** et **60 %** ; le nombre de documents ayant le degré de pertinence 2 varie entre **5** et **60 %** et le nombre de documents ayant obtenu le degré de pertinence 3 varie entre **5** et **45 %**.

Nous constatons que les pourcentages les plus élevés sont associés au niveau 0 de la pertinence, c'est-à-dire qu'un nombre important des documents rapportés ne sont pas pertinents. Par exemple, pour la requête 7 (*(la) production minimale (au) Japon)*, le pourcentage de documents considérés comme non pertinents atteint 80 %, c'est-à-dire que 16 documents sur 20 ne sont pas pertinents ; ces derniers obtiennent le degré 0 de pertinence. Au même niveau de pertinence (0), la requête 15 (*déportation (des)*

*étrangers (en) Autriche*), obtient 70 %, c'est-à-dire que 14 documents sur 20 ont le degré de pertinence 0. Les requêtes 16 et 41 ((*les) communistes (au) parlement européen ; ouvriers étrangers (en) Europe*), suivent avec 65 %, ce qui représente 13 documents sur 20 ayant obtenu le degré 0 de pertinence.

Au niveau de la pertinence 1, la valeur la plus élevée est 60 %, elle représente 12 documents sur 20. Ce score est associé à deux requêtes, 1 et 42 (*culture écologique* ; *conséquences (de) (la) réunification allemande*). Au niveau de la pertinence 2, une seule requête atteint le score de 60 % (12 documents sur 20), il s'agit de la requête numéro 17 (*destruction (de) (la) forêt tropicale (en) Amérique (du) sud*). Finalement, au niveau de la pertinence 3, la valeur maximale obtenue est 45 % ; celle-ci représente 9 documents sur 20. Les requêtes ayant obtenu ce score sont au nombre de 5, il s'agit des requêtes 8 (*extrémisme (de) droite (et) racisme*), 20 (*limitations (des) importations (de) (l') UE*), 24 ((*l') éducation sexuelle*), 26 (*terrorisme international*) et 38 ((*l') homosexualité (et) (la) loi*). Ainsi, le niveau de pertinence 3 enregistre le score le plus bas, en termes de pourcentage, par rapport aux autres niveaux (0, 1 et 2), ce qui reflète le fait qu'une grande partie des documents retournés n'atteint pas le degré de pertinence maximal.

Une autre lecture des résultats associés au score de pertinence des documents (Tableau 4.1), consiste à regrouper les valeurs obtenues par niveaux de pertinence ; c'est-à-dire regrouper les valeurs, en pourcentage, des documents ayant les niveaux de pertinence 0 et 1, de même pour les documents ayant les niveaux de pertinence 2 et 3. La tendance globale liée au regroupement des valeurs, montre que le nombre total de documents ayant obtenu les niveaux de pertinence bas (0 et 1) atteint 529 documents, ce qui représente 52% ; ce qui signifie que plus de la moitié des documents rapportés sont non pertinents ou ont un niveau de pertinence bas. Par ailleurs, le nombre de documents ayant obtenu les niveaux élevés de pertinence (2 et 3), atteint 471 documents, ce qui représente 47 % en termes de pourcentage. Vus individuellement, les pourcentages obtenus se caractérisent par diverses variations ; par exemple, les valeurs associées aux niveaux bas de la pertinence (0 et 1) varient entre 5 et 95 %, c'est la même tendance pour les valeurs dont les niveaux de pertinence sont élevées (2 et 3), c'est-à-dire qu'elles varient aussi entre 5 et 95 %. Nous citons l'exemple de la requête 40 (*effets (du) chocolat (sur) (la) santé*) ; celle-ci enregistre seulement 5 %, ce

qui représente un seul document non pertinent (0 et 1) sur un total de 20. Cela veut dire également que, pour cette requête, 95 % de documents font partie des niveaux de pertinence les plus élevés, en l'occurrence 2 et 3. Une autre requête ayant des valeurs extrêmes, il s'agit de la requête 15 (*déportation (des) étrangers (en) Autriche*) ; celle-ci comprend 19 documents ayant des niveaux de pertinence bas (0 et 1), ce qui équivaut à 95 %, et seulement 1 document ayant un niveau de pertinence 3, ce qui représente 5% en termes de pourcentage. Appartiennent à la même catégorie les requêtes 37 (*vitesse (sur) (les) autoroutes (en) Suisse*), 41 (*ouvriers étrangers (en) Europe*) et 49 (*normes (de) protection professionnelle*) En plus de ces cas extrêmes, d'autres requêtes obtiennent des valeurs similaires, comme c'est le cas de la requête 33 (*réfugiés (de) Bosnie*) où 50% de documents appartiennent aux niveaux bas de pertinence (0 et 1) et les autres 50% aux niveaux de pertinence élevés (2 et 3), sinon des valeurs proches où les pourcentages enregistrés varient entre 45 et 55 %, c'est l'exemple des requêtes 2 (*antisémitisme (en) Allemagne après 1945*), 10 (*criminalité (des) adolescents*), 18 (*famine (au) Soudan*), 21 (*unité franco-allemande*), etc.

## Tableau 4.1 : Score de pertinence des documents

Nb = nombre de documents ; PCT = pourcentage

| | | PERTINENCE | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | | 1 | | 0+1 | 2 | | 3 | | 2+3 |
| | Requête | Nb | PCT % | Nb | PCT % | % | Nb | PCT % | Nb | PCT % | % |
| 1 | culture écologique | 3 | 15% | 12 | 60% | 75% | 4 | 20% | 1 | 5% | 25% |
| 2 | antisémitisme Allemagne 1945 | 6 | 30% | 3 | 15% | 45% | 8 | 40% | 3 | 15% | 55% |
| 3 | exportation médicaments dangereux | 6 | 30% | 6 | 30% | 60% | 4 | 20% | 4 | 20% | 40% |
| 4 | criminalité féminine | 4 | 20% | 9 | 45% | 65% | 2 | 10% | 5 | 25% | 35% |
| 5 | politique extérieure Autriche | 11 | 55% | 4 | 20% | 75% | 1 | 5% | 4 | 20% | 25% |
| 6 | conditions vie immigrants | 2 | 10% | 3 | 15% | 25% | 8 | 40% | 7 | 35% | 75% |
| 7 | production minimale Japon | 16 | 80% | 2 | 10% | 90% | 1 | 5% | 1 | 5% | 10% |
| 8 | extrémisme droite racisme | 0 | 0% | 3 | 15% | 15% | 8 | 40% | 9 | 45% | 85% |
| 9 | jeunesse politique | 3 | 15% | 6 | 30% | 45% | 5 | 25% | 6 | 30% | 55% |
| 10 | criminalité adolescents | 9 | 45% | 2 | 10% | 55% | 4 | 20% | 5 | 25% | 45% |
| 11 | charte sociale européenne | 2 | 10% | 6 | 30% | 40% | 5 | 25% | 7 | 35% | 60% |
| 12 | violence adolescents | 7 | 35% | 5 | 25% | 60% | 4 | 20% | 4 | 20% | 40% |
| 13 | législation protection nature | 8 | 40% | 5 | 25% | 65% | 4 | 20% | 3 | 15% | 35% |
| 14 | politique économique Slovénie | 8 | 40% | 5 | 25% | 65% | 3 | 15% | 4 | 20% | 35% |
| 15 | déportation étrangers Autriche | 14 | 70% | 5 | 25% | 95% | 0 | 0% | 1 | 5% | 5% |
| 16 | communistes parlement européen | 13 | 65% | 3 | 15% | 80% | 4 | 20% | 0 | 0% | 20% |
| 17 | destruction forêt tropicale Amérique sud | 3 | 15% | 4 | 20% | 35% | 12 | 60% | 1 | 5% | 65% |

| | | PERTINENCE | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | | 1 | | 0+1 | 2 | | 3 | | 2+3 |
| | Requête | Nb | PCT % | Nb | PCT % | % | Nb | PCT % | Nb | PCT % | % |
| 18 | famine soudan | 3 | 15% | 8 | 40% | 55% | 6 | 30% | 3 | 15% | 45% |
| 19 | protection animaux | 3 | 15% | 10 | 50% | 65% | 4 | 20% | 3 | 15% | 35% |
| 20 | limitations importations UE | 3 | 15% | 2 | 10% | 25% | 6 | 30% | 9 | 45% | 75% |
| 21 | unité franco-allemande | 7 | 35% | 4 | 20% | 55% | 7 | 35% | 2 | 10% | 45% |
| 22 | immigration racisme | 2 | 10% | 9 | 45% | 55% | 6 | 30% | 3 | 15% | 45% |
| 23 | accidents route | 3 | 15% | 9 | 45% | 60% | 5 | 25% | 3 | 15% | 40% |
| 24 | éducation sexuelle | 1 | 5% | 3 | 15% | 20% | 7 | 35% | 9 | 45% | 80% |
| 25 | effets déforestation | 2 | 10% | 4 | 20% | 30% | 7 | 35% | 7 | 35% | 70% |
| 26 | terrorisme international | 2 | 10% | 4 | 20% | 30% | 5 | 25% | 9 | 45% | 70% |
| 27 | maltraitance enfants | 1 | 5% | 7 | 35% | 40% | 4 | 20% | 8 | 40% | 60% |
| 28 | exploitation économique fond marin | 2 | 10% | 4 | 20% | 30% | 8 | 40% | 6 | 30% | 70% |
| 29 | exportation armes Turquie | 7 | 35% | 6 | 30% | 65% | 7 | 35% | 0 | 0% | 35% |
| 30 | débris spatiaux | 4 | 20% | 3 | 15% | 35% | 7 | 35% | 6 | 30% | 65% |
| 31 | séparatistes catalans galiciens | 11 | 55% | 5 | 25% | 80% | 4 | 20% | 0 | 0% | 20% |
| 32 | coopération internationale entreprises | 6 | 30% | 5 | 25% | 55% | 6 | 30% | 3 | 15% | 45% |
| 33 | réfugiés Bosnie | 4 | 20% | 6 | 30% | 50% | 5 | 25% | 5 | 25% | 50% |
| 34 | accidents industrie minière | 7 | 35% | 2 | 10% | 45% | 6 | 30% | 5 | 25% | 55% |
| 35 | transport public local | 7 | 35% | 2 | 10% | 45% | 7 | 35% | 4 | 20% | 55% |
| 36 | pollution causée automobile | 3 | 15% | 8 | 40% | 55% | 4 | 20% | 5 | 25% | 45% |
| 37 | vitesse autoroutes Suisse | 11 | 55% | 8 | 40% | 95% | 0 | 0% | 1 | 5% | 5% |
| 38 | homosexualité loi | 2 | 10% | 5 | 25% | 35% | 4 | 20% | 9 | 45% | 65% |
| 39 | processus paix Moyen-Orient | 2 | 10% | 10 | 50% | 60% | 6 | 30% | 2 | 10% | 40% |
| 40 | effets chocolat santé | 1 | 5% | 0 | 0% | 5% | 10 | 50% | 9 | 45% | 95% |

| | | PERTINENCE | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | | 1 | | 0+1 | 2 | | 3 | | 2+3 |
| | Requête | Nb | PCT % | Nb | PCT % | % | Nb | PCT % | Nb | PCT % | % |
| 41 | ouvriers étrangers Europe | 13 | 65% | 6 | 30% | 95% | 1 | 5% | 0 | 0% | 5% |
| 42 | conséquences réunification | 1 | 5% | 12 | 60% | 65% | 5 | 25% | 2 | 10% | 35% |
| 43 | conversion dette Pologne | 10 | 50% | 6 | 30% | 80% | 4 | 20% | 0 | 0% | 20% |
| 44 | statut militaire Allemagne unifiée | 8 | 40% | 6 | 30% | 70% | 3 | 15% | 3 | 15% | 30% |
| 45 | lutte corruption | 1 | 5% | 6 | 30% | 35% | 7 | 35% | 6 | 30% | 65% |
| 46 | protection environnement sein entreprises | 2 | 10% | 3 | 15% | 25% | 8 | 40% | 7 | 35% | 75% |
| 47 | maintien paix OUA | 1 | 5% | 3 | 15% | 20% | 9 | 45% | 7 | 35% | 80% |
| 48 | industrie européenne film | 5 | 25% | 2 | 10% | 35% | 7 | 35% | 6 | 30% | 65% |
| 49 | normes protection professionnelle | 11 | 55% | 8 | 40% | 95% | 1 | 5% | 0 | 0% | 5% |
| 50 | traitement déchets nucléaires | 3 | 15% | 6 | 30% | 45% | 7 | 35% | 4 | 20% | 55% |
| **Toutes les requêtes** | | **264** | **26%** | **265** | **26%** | **52%** | **260** | **26%** | **211** | **21%** | **47%** |

Le Tableau 4.2, ci-dessous, présente les résultats du score de pertinence obtenus pour les 50 requêtes. Ce score est trié d'une façon ascendante : le minimum théorique = 0 ; le maximum théorique = 60 ; le score = somme des 20 cotes de pertinence (0, 1, 2, 3). Par exemple, pour la requête 7 (*(la) production minimale (au) Japon)*, le score obtenu est 7, il est le plus bas ; ce score est calculé de la façon suivante : 0 (pertinence) * 16 (nombre de documents) = 0 ; 1 * 2 = 2 ; 2 * 1 = 2 ; 3 * 1 = 3 ; la somme = 7. C'est ainsi pour le reste des requêtes.

**Tableau 4.2 : Score de pertinence des requêtes**

| N° | requête | Score pertinence |
|---|---|---|
| 7 | production minimale (au) Japon | 7 |
| 15 | déportation (des) étrangers (en) Autriche | 8 |
| 41 | ouvriers étrangers (en) Europe | 8 |
| 49 | normes (de) protection professionnelle | 10 |
| 16 | (les) communistes (au) parlement européen | 11 |
| 37 | Vitesse (sur) (les) autoroutes (en) Suisse | 11 |
| 31 | séparatistes catalans (et) galiciens | 13 |
| 43 | conversion (de) (la) dette (pour) (la) Pologne | 14 |
| 5 | politique extérieure (de) (l') Autriche | 18 |
| 29 | exportation (d') armes (à) (la) Turquie | 20 |
| 44 | statut militaire (de) (l') Allemagne unifiée | 21 |
| 13 | législation (sur) (la) protection (de) (la) nature | 22 |
| 1 | (la) culture écologique | 23 |
| 14 | (la) politique économique (de) (la) Slovénie | 23 |
| 21 | unité franco-allemande | 24 |
| 10 | criminalité (des) adolescents | 25 |
| 12 | violence (des) adolescents | 25 |
| 3 | exportation (de) médicaments dangereux | 26 |
| 32 | coopération internationale (des) entreprises | 26 |
| 19 | protection (des) animaux | 27 |
| 2 | antisémitisme (en) Allemagne (après) 1945 | 28 |
| 4 | criminalité féminine | 28 |
| 23 | (les) accidents (de) (la) route | 28 |
| 35 | transport public local | 28 |
| 39 | processus (de) paix (au) Moyen-Orient | 28 |
| 42 | conséquences (de) (la) réunification allemande | 28 |
| 18 | famine (au) Soudan | 29 |
| 34 | accidents (dans) (l') industrie minière | 29 |
| 22 | immigration (et) racisme | 30 |
| 17 | destruction (de) (la) forêt tropicale (en) Amérique (du) sud | 31 |
| 33 | réfugiés (de) Bosnie | 31 |
| 36 | (la) pollution causée (par) (l') automobile | 31 |
| 50 | traitement (des) déchets nucléaires | 32 |
| 9 | jeunesse (et) politique | 34 |
| 48 | (l') industrie européenne (du) film | 34 |

| | | |
|---|---|---|
| **30** | (les) débris spatiaux | **35** |
| **11** | charte sociale européenne | **37** |
| **28** | exploitation économique (du) fond marin | **38** |
| **45** | lutte (contre) (la) corruption | **38** |
| **25** | (les) effets (de) (la) déforestation | **39** |
| **27** | maltraitance (des) enfants | **39** |
| **6** | conditions (de) vie (des) immigrants | **40** |
| **38** | (l') homosexualité (et) (la) loi | **40** |
| **46** | protection (de) (l') environnement (au) sein (des) entreprises | **40** |
| **20** | limitations (des) importations (de) (l') UE | **41** |
| **26** | terrorisme international | **41** |
| **47** | maintien (de) (la) paix (par) OUA | **42** |
| **24** | (l') éducation sexuelle | **44** |
| **8** | extrémisme (de) droite (et) racisme | **46** |
| **40** | effets (du) chocolat (sur) (la) santé | **47** |

Le constat global concernant le Tableau 4.2 ci-dessus, c'est que les scores de pertinence associés aux requêtes se caractérisent par une importante variation. À ce sujet, les scores obtenus varient de 7, comme score le plus faible obtenu pour la requête 7 (***production minimale (au) Japon***), à 47, comme score le plus élevé obtenu pour la requête 40 (***effets (du) chocolat (sur) (la) santé***). À part quelques requêtes qui ont obtenu un score similaire, les autres requêtes enregistrent des scores variés. Le tri des scores de pertinence nous permet de faire ressortir, d'une façon évidente, les requêtes où le score de pertinence enregistre des valeurs statistiquement positives. Ainsi, en nous basant sur les résultats du Tableau 4.2, nous observons que les scores obtenus sont positifs, et cela à partir du score de pertinence 30 (sur 60, maximum théorique), en partant du principe que le score 30 constitue la valeur médiane. À cet égard, nous considérons que le score de pertinence enregistré par les 22 dernières requêtes (Tableau 4.2), qui se situe entre 30 et 47, est un score significatif. Nous remarquons également qu'un groupe de 8 requêtes obtiennent des valeurs intéressantes, c'est-à-dire que le score associé à ces requêtes est proche de la moyenne (30) ; il s'agit des valeurs 28 et 29 respectivement.

À ce stade d'analyse il reste à vérifier s'il y a une corrélation entre les scores de pertinence obtenus et les différentes variables prises en compte dans la présente étude. La section suivante présentera les données brutes obtenues pour les 50 requêtes. Ces données représentent la fréquence des différentes variables étudiées (*termes de la*

*requête, variantes dérivées et variantes fléchies*) ; le pourcentage de la fréquence des variables étudiées prend en compte la variable *longueur du texte.*

### 4.1.2 Données brutes associées aux requêtes

Nous présentons au Tableau 4.3, ci-dessous, les données brutes associées aux 50 requêtes. Il s'agit de données obtenues après l'analyse des 20 premiers documents rapportés. L'analyse prend en compte 4 variables : *les termes de la requête, les variantes dérivées, les variantes fléchies et la longueur du texte.* Pour pouvoir comparer la fréquence des variables liées aux requêtes avec les scores de pertinence obtenus, nous reprenons les scores présentés au Tableau 4.2. Les données sont présentées au Tableau 4.3 comme suit : à chaque requête est associé le score de pertinence, la somme des termes de la requête (chaque terme au sens de mot est calculé individuellement), la somme des variantes dérivées, la somme des variantes fléchies et la longueur du texte ; les sommes en question représentent la fréquence de chacune des variables dans les 20 documents analysés. Nous avons traduit également les valeurs obtenues en termes de pourcentage, et cela en nous basant sur la variable *longueur du document.* Celle-ci représente le nombre de mots composant le document, cette valeur est calculée au moyen du logiciel le Fréquencier. Cette dernière variable (*longueur du document*) est retenue pour calculer les pourcentages des résultats bruts associés à chaque variable (*les termes de la requête, les variantes dérivées et les variantes fléchies*).

**Tableau 4.3 : Données brutes**

| N° | Requête | Score P. | Termes | % | Dér. | % | Flex. | % | L. texte |
|---|---|---|---|---|---|---|---|---|---|
| 1 | (La) culture écologique | 23 | 262 | **0,73%** | 205 | **0,57%** | 43 | **0,12%** | 35979 |
| 2 | anti-sémitisme (en) Allemagne (après) 1945 | 28 | 297 | **0,47%** | 470 | **0,75%** | 1 | **0,00%** | 62533 |
| 3 | Exportation (de) médicaments dangereux | 26 | 651 | **0,88%** | 689 | **0,93%** | 177 | **0,24%** | 74307 |
| 4 | Criminalité féminine | 28 | 249 | **0,50%** | 775 | **1,55%** | 29 | **0,06%** | 49919 |
| 5 | Politique extérieure (de) (l') Autriche | 18 | 667 | **1,63%** | 126 | **0,31%** | 47 | **0,11%** | 40937 |
| 6 | Conditions (de) vie (des) immigrants | 40 | 332 | **0,64%** | 684 | **1,32%** | 35 | **0,07%** | 51781 |
| 7 | Production minimale (au) Japon | 7 | 273 | **0,93%** | 172 | **0,59%** | 22 | **0,08%** | 29281 |
| 8 | Extrémisme (de) droite (et) racisme | 46 | 620 | **1,80%** | 237 | **0,69%** | 68 | **0,20%** | 34416 |
| 9 | Jeunesse (et) politique | 34 | 917 | **2,68%** | 440 | **1,29%** | 69 | **0,20%** | 34234 |
| 10 | Criminalité (des) adolescents | 25 | 315 | **0,91%** | 246 | **0,71%** | 59 | **0,17%** | 34447 |
| 11 | Charte sociale européenne | 37 | 817 | **2,40%** | 154 | **0,45%** | 140 | **0,41%** | 33973 |
| 12 | Violence (des) adolescents | 25 | 1541 | **1,61%** | 324 | **0,34%** | 161 | **0,17%** | 95577 |
| 13 | Législation (sur) (la) protection (de) (la) nature | 22 | 430 | **1,07%** | 205 | **0,51%** | 2 | **0,00%** | 40305 |
| 14 | (La) politique économique (de) (la) Slovénie | 23 | 623 | **1,67%** | 249 | **0,67%** | 153 | **0,41%** | 37307 |
| 15 | Déportation (des) étrangers (en) Autriche | 8 | 252 | **0,48%** | 181 | **0,34%** | 64 | **0,12%** | 52979 |
| 16 | (Les) communistes (au) parlement européen | 11 | 293 | **1,28%** | 108 | **0,47%** | 167 | **0,73%** | 22965 |
| 17 | Destruction (de) (la) forêt tropicale (en) Amérique (du) Sud | 31 | 830 | **1,50%** | 420 | **0,76%** | 952 | **1,73%** | 55177 |
| 18 | Famine (au) Soudan | 29 | 231 | **1,89%** | 41 | **0,34%** | 0 | **0,00%** | 12194 |
| 19 | Protection (des) animaux | 27 | 290 | **1,90%** | 54 | **0,35%** | 72 | **0,47%** | 15230 |
| 20 | Limitations (des) importations (de) (l') UE | 41 | 243 | **1,07%** | 242 | **1,07%** | 50 | **0,22%** | 22683 |
| 21 | Unité franco-allemande | 24 | 185 | **0,70%** | 75 | **0,28%** | 77 | **0,29%** | 26511 |
| 22 | Immigration (et) racisme | 30 | 321 | **1,20%** | 178 | **0,67%** | 3 | **0,01%** | 26705 |
| 23 | (les) accidents (de) (la) route | 28 | 283 | **1,41%** | 80 | **0,40%** | 176 | **0,88%** | 20036 |
| 24 | (L') éducation sexuelle | 44 | 578 | **1,66%** | 334 | **0,96%** | 94 | **0,27%** | 34895 |
| 25 | (Les) effets (de) (la) déforestation | 39 | 333 | **0,84%** | 397 | **1,00%** | 83 | **0,21%** | 39635 |
| 26 | Terrorisme international | 41 | 489 | **1,33%** | 364 | **0,99%** | 86 | **0,23%** | 36776 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 27 | Maltraitance (des) enfants | 39 | 903 | **1,69%** | 482 | **0,90%** | 387 | **0,72%** | 53442 |
| 28 | Exploitation économique (du) fond marin | 38 | 283 | **0,87%** | 177 | **0,55%** | 268 | **0,83%** | 32377 |
| 29 | Exportation (d') armes (à) (la) Turquie | 20 | 500 | **1,91%** | 282 | **1,08%** | 87 | **0,33%** | 26177 |
| 30 | (Les) débris spatiaux | 35 | 415 | **0,87%** | 85 | **0,18%** | 105 | **0,22%** | 47768 |
| 31 | Séparatistes catalans (et) galiciens | 13 | 113 | **0,10%** | 308 | **0,26%** | 124 | **0,11%** | 117723 |
| 32 | Coopération internationale (des) entreprises | 26 | 461 | **1,42%** | 114 | **0,35%** | 160 | **0,49%** | 32567 |
| 33 | Réfugiés (de) Bosnie | 31 | 325 | **1,78%** | 14 | **0,08%** | 7 | **0,04%** | 18242 |
| 34 | Accidents (dans) (l') industrie minière | 29 | 650 | **1,04%** | 682 | **1,09%** | 280 | **0,45%** | 62631 |
| 35 | Transport public local | 28 | 708 | **2,93%** | 123 | **0,51%** | 154 | **0,64%** | 24146 |
| 36 | (La) pollution causée (par) (l') automobile | 31 | 298 | **1,22%** | 159 | **0,65%** | 60 | **0,25%** | 24342 |
| 37 | Vitesse (sur) (les) autoroutes (en) Suisse | 11 | 327 | **1,49%** | 162 | **0,74%** | 164 | **0,75%** | 21936 |
| 38 | (L') homosexualité (et) (la) loi | 40 | 432 | **1,07%** | 751 | **1,87%** | 65 | **0,16%** | 40202 |
| 39 | Processus (de) paix (au) Moyen-Orient | 28 | 479 | **2,70%** | 16 | **0,09%** | 0 | **0,00%** | 17720 |
| 40 | Effets (du) chocolat (sur) (la) santé | 47 | 580 | **2,74%** | 40 | **0,19%** | 44 | **0,21%** | 21198 |
| 41 | Ouvriers étrangers (en) Europe | 8 | 600 | **0,93%** | 163 | **0,25%** | 162 | **0,25%** | 64650 |
| 42 | Conséquences (de) (la) réunification allemande | 28 | 548 | **0,85%** | 714 | **1,11%** | 276 | **0,43%** | 64224 |
| 43 | Conversion (de) (la) dette (pour) (la) Pologne | 14 | 839 | **1,28%** | 295 | **0,45%** | 137 | **0,21%** | 65595 |
| 44 | Statut militaire (de) (l') Allemagne unifiée | 21 | 1102 | **1,09%** | 1537 | **1,52%** | 103 | **0,10%** | 100971 |
| 45 | Lutte (contre) (la) corruption | 38 | 437 | **1,97%** | 40 | **0,18%** | 0 | **0,00%** | 22171 |
| 46 | Protection (de) (l') environnement (au) sein (des) entreprises | 40 | 1288 | **1,97%** | 433 | **0,66%** | 113 | **0,17%** | 65543 |
| 47 | Maintien (de) (la) paix (par) (l') OUA | 42 | 1138 | **2,75%** | 121 | **0,29%** | 0 | **0,00%** | 41442 |
| 48 | (L') industrie européenne (du) film | 34 | 410 | **1,58%** | 167 | **0,65%** | 351 | **1,36%** | 25882 |
| 49 | Normes (de) protection professionnelle | 10 | 332 | **0,78%** | 147 | **0,35%** | 187 | **0,44%** | 42296 |
| 50 | Traitement (des) déchets nucléaires | 32 | 903 | **2,46%** | 95 | **0,26%** | 207 | **0,56%** | 36720 |
| | **Totaux** | | **26393** | **1,26%** | **14558** | **0,70%** | **6271** | **0,30%** | **2090747** |

Les résultats globaux des données analysées reflètent une tendance intéressante. Celle-ci caractérise l'ensemble des variables étudiées : *termes de la requête, variantes dérivées et variantes fléchies*. En ce sens, les résultats obtenus (Tableau 4.3) montrent que la fréquence des termes de la requête dans les documents rapportés est significative ; en ce sens, le nombre de termes de la requête atteint 26393, ce qui représente 1,26 % en termes de pourcentage. Ce dernier est calculé comme suit : le nombre de termes (26393 termes) divisé par la longueur des documents (2090747 mots) multiplié par 100. Les valeurs associées à la variable *termes de la requête* semblent indiquer que la présence des termes de la requête dans les documents retournés est significative. Mais cela ne veut pas dire que la fréquence de ces termes est synonyme de pertinence des documents rapportés. Si la fréquence des termes de la requête est importante, cette fréquence l'est moins pour les variantes dérivées et celles fléchies. À cet égard, les deux variantes (dérivées et fléchies) enregistrent des valeurs moins importantes que celles obtenues par les termes de la requête ; ainsi, en termes de pourcentage, la variable *variantes dérivées* obtient 0,70 %, ce qui correspond à 14558 variantes dérivées, c'est une valeur significative par rapport au pourcentage obtenu par la variable *variantes fléchies* qui, lui, atteint seulement 0,30 % (6271 variantes fléchies). Alors, est-ce que la tendance globale enregistrée reflète réellement les tendances individuelles ?

En effet, nous remarquons une grande disparité entre les valeurs enregistrées, et cela selon les requêtes et également selon les variables, c'est-à-dire qu'il n'y a pas de lien entre les termes et les variantes. Par exemple, pour les termes de la requête, le meilleur pourcentage enregistré correspond à 2,93%, il est associé à la requête 35 (*transport public local*) qui a score de pertinence de 28 (au-dessous de la moyenne), alors que le pire score obtenu est 0,10%, il est lié à la requête 31(*séparatistes catalans (et) galiciens*) qui obtient un score de pertinence de 13. Pour les variantes dérivées, le meilleur score obtenu atteint 1,87%, ce score est lié à la requête 38 (*(l') homosexualité (et) (la) loi*) qui obtient un score de pertinence de 40, alors que le pire score enregistré atteint 0,08%, ce pourcentage est lié à la requête 33 (*réfugiés (de) Bosnie*) qui enregistre un score de pertinence positif (31). Pour les variantes fléchies, le meilleur score obtenu atteint 1,73%, ce score est associé à la requête 17 (*Destruction (de) (la) forêt tropicale (en) Amérique (du) Sud*) qui enregistre un score de pertinence positif (31), alors que le pire score, en termes de fréquence, est 0,00%, ce score est enregistré

dans 5 requêtes (2, 13, 18, 39, 45 et 47) ; parmi ces six requêtes deux seulement enregistrent des scores de pertinence positifs, il s'agit des requêtes 45 et 47 avec des scores de 38 et 42 respectivement. Ce qui est intéressant de souligner c'est le fait que les meilleurs scores des termes et des variantes ne correspondent pas toujours à des bons scores de pertinence des requêtes. De même, les mauvais scores en termes de fréquence des variables peuvent correspondre à de bons scores de pertinence des requêtes.

Le Tableau 4.4, ci-dessous, illustre cette variation caractérisant les requêtes et les variables.

**Tableau 4.4 : Valeurs extrêmes des termes et des variantes**

| | Les 5 meilleures | | VS | | Les 5 plus mauvaises | |
|---|---|---|---|---|---|---|
| requête | **termes %** | score p. | | requête | **termes %** | score p. |
| 35 | **2,93%** | 28 (-) | | 31 | **0,10%** | 13 (-) |
| 47 | **2,75%** | **42 (+)** | | 2 | **0,47%** | 28 (-) |
| 40 | **2,74%** | **47 (+)** | | 4 | **0,50%** | 28 (-) |
| 45 | **2,71%** | **38 (+)** | | 6 | **0,64%** | **40 (+)** |
| 39 | **2,70%** | 28 (-) | | 21 | **0,70%** | 24 (-) |
| | | | | | | |
| requête | **dérivés %** | score p. | VS | requête | **dérivés %** | score p. |
| 38 | **1,87%** | **40 (+)** | | 33 | **0,08%** | **31 (+)** |
| 4 | **1,55%** | 28 (-) | | 39 | **0,09%** | 28 (-) |
| 44 | **1,50%** | 21 (-) | | 30 | **0,18%** | **35 (+)** |
| 6 | **1,32%** | **40 (+)** | | 40 | **0,19%** | **47 (+)** |
| 9 | **1,29%** | **34 (+)** | | 47 | **0,23%** | **42 (+)** |
| | | | | | | |
| requête | **fléchis %** | score p. | VS | requête | **fléchis %** | score p. |
| 17 | **1,73%** | **31 (+)** | | 2 | **0,00%** | 28 (-) |
| 23 | **0,88%** | 28 (-) | | 13 | **0,00%** | 22 (-) |
| 28 | **0,83%** | **38 (+)** | | 18 | **0,00%** | 29 (-) |
| 37 | **0,75%** | 11 (-) | | 39 | **0,00%** | 28 (-) |
| 16 | **0,73%** | 11 (-) | | 45 | **0,00%** | **38 (+)** |

Ce tableau présente les 5 meilleurs et les 5 mauvais scores pour chacune des variables (termes de la requête, variantes dérivées et variantes fléchies). Nous remarquons qu'il y a des variations selon les requêtes et aussi selon les variables. Ainsi, pour les termes de la requête, les pourcentages de fréquence varient d'une requête à l'autre ; de même, parmi les 5 meilleurs scores enregistrés en termes de fréquence, 3 requêtes sur 5 ont

obtiennent des scores de pertinence positifs, alors que 2 requêtes enregistrent des scores négatifs. Il est à noter que le meilleur score obtenu en termes de fréquence est associé à la requête 35, mais celle-ci n'obtient pas un score de pertinence positif. Pour les 5 mauvais scores obtenus en termes de fréquence, nous constatons également des variations de pourcentages entre les 5 requêtes, alors qu'au niveau des scores de pertinence, 4 requêtes enregistrent des scores négatifs et une seule requête réalise un score positif, il s'agit de la requête 6. Concernant les variantes dérivées, les 5 meilleurs pourcentages en termes de fréquence présentent des variations entre les différentes requêtes, ces variations caractérisent également les scores de pertinence qui leur sont associés ; en ce sens, 3 requêtes sur 5 obtiennent des scores positifs en termes de pertinence, alors que 2 requêtes enregistrent des scores négatifs. Pour les 5 mauvais pourcentages en termes de fréquence, nous relevons des variations entre les différentes requêtes ; toutefois, au niveau des scores de pertinence, les résultats sont intéressants car 4 requêtes sur 5 obtiennent des scores positifs. À propos des variantes fléchies, comme c'était le cas pour les autres variables, les 5 meilleurs pourcentages en termes de fréquence présentent des variations selon les requêtes ; ces variations caractérisent également les scores de pertinence associés à chacune des requêtes ; à cet égard, 2 requêtes obtiennent des scores positifs et 3 autres des scores négatifs. Pour les 5 mauvais pourcentages en termes de fréquence, nous constatons que la fréquence est nulle dans les 5 requêtes ; toutefois, au niveau des scores de pertinence, nous relevons que 4 requêtes enregistrent des scores négatifs et une requête obtient un score positif. Ainsi, les exemples cités au Tableau 4.4 révèlent une grande disparité selon les requêtes et selon les variables ; de même, cette variation ne permet pas de dégager une tendance confirmant le fait qu'il y a une corrélation entre la fréquence des variables et le score de pertinence des requêtes.

Une autre lecture des résultats consiste, en nous basant sur le Tableau 4.2, à regrouper les valeurs obtenues en quatre blocs, selon les scores de pertinence obtenus ; ces 4 blocs regroupent les scores allant de 7 à 18, de 20 à 27, de 28 à 34 et de 35 à 47. Nous présentons ces variations au Tableau 4.5 ci-dessous.

**Tableau 4.5 : Tendances selon 4 blocs de pertinence**

| Blocs | Nb requêtes | Termes | Dérivés | Fléchis |
|---|---|---|---|---|
| [7-18] | 9 | **0,99%** | **0,42%** | **0,31%** |
| [20-27] | 11 | **1,26%** | **0,66%** | **0,25%** |
| [28-34] | 15 | **1,61%** | **0,68%** | **0,44%** |
| [35-47] | 15 | **1,63%** | **0,77%** | **0,26%** |

Ce tableau (Tableau 4.5) présente les tendances en termes de fréquence des trois variables (les termes, les dérivés et les fléchis), et cela selon 4 blocs de pertinence (Tableau 4.2). Le bloc [7-18] comprend 9 requêtes, le bloc [20-27] contient 11 requêtes, le bloc [28-34] contient 15 requêtes et le bloc [35-47] comprend également 15 requêtes. Pour chaque bloc, nous avons établi la moyenne des pourcentages (en termes de fréquence) associée à chacune des variables. Si nous regardons les pourcentages enregistrés dans chacun des 4 blocs, nous constatons que globalement il semble y avoir une corrélation entre la pertinence et la fréquence des termes et des dérivés ; ce qui n'est pas le cas des variantes fléchies. Par exemple, le bloc [7-18], qui est le plus mauvais en termes de pertinence, obtient les moyennes les plus basses en termes de fréquence, notamment au niveau des termes (0,99%) et des dérivés (0,42%). Cette tendance varie d'une variable à l'autre, et cela d'une façon descendante, c'est-à-dire que le pourcentage le plus élevé est enregistré par les termes, suivis des dérivés et des fléchis (cette tendance caractérise les 4 blocs). De même, une tendance ascendante caractérise les 4 blocs au niveau des pourcentages des termes et des dérivés, en allant du bloc le plus mauvais au meilleur bloc. Toutefois, cette tendance globale ne reflète pas forcément les caractéristiques de chaque requête ; par exemple, dans ce même bloc [7-18], la requête 5 contient un pourcentage intéressant de termes (1,63%) ; de même, la requête 37 contient un pourcentage important de dérivés (0,74%) et de fléchis (0,75). Un autre exemple lié au meilleur bloc montre également cette disparité, il s'agit de la requête 6 qui enregistre peu de termes (0,64) et les requêtes 30 et 40 qui enregistrent très peu de dérivés, 0,18% et 0,19% respectivement. Pour les fléchis, les pourcentages sont moins intéressants et varient d'un bloc à l'autre. Nous remarquons que les mauvais pourcentages enregistrés (0,00%) au niveau de cette variable font partie des deux derniers blocs, c'est-à-dire les meilleurs [28-34] et [35-47], il s'agit

des requêtes 2, 39, 18, 45 et 47. Dans la même optique, nous présentons au Tableau 4.6, section suivante, un autre regroupement des valeurs. Il est question d'un regroupement par niveau de pertinence.

### 4.1.3 Tendances par niveau de pertinence

Nous proposons dans cette section un résumé des résultats bruts. Il s'agit d'une lecture globale prenant en considération le pourcentage obtenu, toutes requêtes confondues, en fonction du niveau de pertinence associé à chacune des variables (*termes de la requête, variantes dérivées et variantes fléchies*) ; ce pourcentage est établi en tenant compte de la longueur du document. Ainsi, nous avons procédé de la façon suivante :

- pour les 50 requêtes, nous avons établi la somme des fréquences liée à chaque variable, et cela à chaque niveau de pertinence (0, 1, 2 et 3) ;
- nous avons calculé, ensuite, le pourcentage associé à chacune des variables en prenant en compte la longueur du texte.

**Tableau 4.6 : Tendances par niveau de pertinence**

| Perti. | Nb Docs. | Termes | % | V. dérivées | % | V. fléchies | % | L. texte |
|---|---|---|---|---|---|---|---|---|
| **0** | 264 | 4392 | **0,80%** | 2632 | **0,48%** | 1093 | **0,20%** | 550645 |
| **1** | 265 | 5273 | **1,19%** | 2928 | **0,66%** | 1445 | **0,33%** | 443470 |
| **2** | 260 | 6997 | **1,43%** | 3529 | **0,72%** | 2149 | **0,44%** | 490915 |
| **3** | 211 | 9731 | **1,61%** | 5469 | **0,90%** | 1584 | **0,26%** | 605717 |
| **Total** | **1000** | **26393** | **1,26%** | **14558** | **0,70%** | **6271** | **0,30%** | **2090747** |

Le Tableau 4.6 ci-dessus met en valeur la pertinence des documents indépendamment des requêtes. Les données exposées au tableau représentent diverses variables : niveaux de pertinence, nombre de documents, termes de la requête, variantes dérivées, variantes fléchies et longueur du texte. Les valeurs des variables *termes de la requête, variantes dérivées et variantes fléchies* figurent dans le tableau sous forme de nombre et de pourcentage. Figure également dans le tableau le total associé à chaque variable, tous degrés de pertinence confondus. Les résultats présentés au Tableau 4.6 nous

révèlent une tendance intéressante entre les différents niveaux de pertinence. En ce sens, le pourcentage a tendance à augmenter en passant d'un niveau de pertinence à l'autre, c'est-à-dire une direction ascendante. Hormis la variable *variantes fléchies*, cette tendance ascendante entre les 4 niveaux de pertinence caractérise notamment les variables *termes de la requête* et *variantes dérivées*. Ainsi, en termes de pourcentage, le score obtenu pour la variable *termes de la requête* atteint 0,80 % au niveau 0 de la pertinence, 1,19 % au niveau 1, 1,43 % au niveau 2 et 1,61 % au niveau 3 de la pertinence. Cette corrélation est également positive concernant la variable *variantes dérivées* ; celle-ci enregistre 0,48 % au niveau 0 de la pertinence, 0,66% au niveau 1, 0,72 % au niveau 2 et 0,90 % au niveau 3 de la pertinence. Cette tendance indique que plus la fréquence des termes de la requête et des variantes dérivées est importante plus le degré de pertinence des documents est élevé. Toutefois, la variable *variantes fléchies*, avec des scores de 0,20 % au niveau 0 de pertinence, 0,33 % au niveau 1, 0,44 au niveau 2 et 0,26 au niveau 3, enregistre une tendance moins marquée et régulière que les autres variables surtout au niveau de pertinence 3, c'est-à-dire qu'à ce niveau de pertinence le pourcentage baisse au lieu d'augmenter comme c'est le cas pour les autres variables. Notons que les tendances relevées au Tableau 4.6 reflètent globalement les mêmes tendances constatées au Tableau 4.5. Ainsi, les résultats de ces deux tableaux semblent confirmer l'hypothèse de départ.

Le score des variantes fléchies, qui est le plus bas par rapport aux autres variables, confirme l'idée selon laquelle les variations morphologiques flexionnelles ont un impact minime sur l'amélioration de la qualité des documents rapportés. Cette affirmation rejoint les conclusions des travaux portant sur ce sujet (Chieze, 2000), et qui confirment le fait que l'impact de la morphologie flexionnelle sur le repérage d'information reste minime.

Les variations caractérisant les différentes variables sont illustrées par le graphique ci-dessous (Figure 4.2).

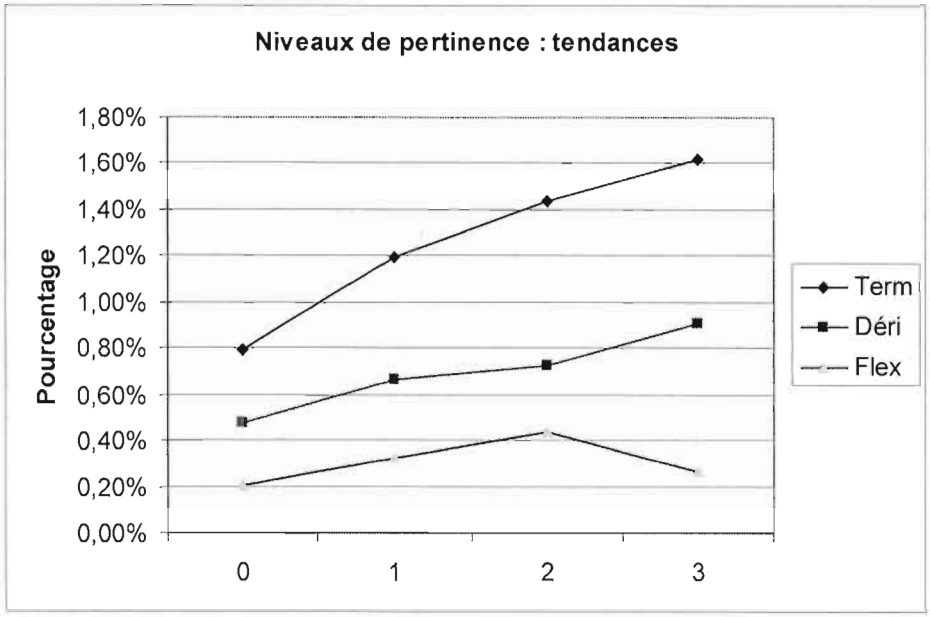



**Figure 4.2 : Corrélation entre niveau de pertinence et fréquence des variables**

La Figure 4.2 est une illustration, sous forme de courbes, des données présentées au Tableau 4.6. Ce graphique montre, d'une façon évidente, la tendance associée aux variables à chaque niveau de pertinence (0, 1, 2 et 3). Ainsi, suivant les tendances figurant dans le graphique ci-dessus, et excepté la tendance caractérisant la variable *variantes fléchies* au niveau de la pertinence 3, la fréquence des variables dans les documents serait en corrélation avec le niveau de pertinence de ces mêmes documents. À cet égard, nous observons une fréquence assez marquée des termes de la requête dans les documents rapportés, ce qui est, à notre avis, logique, c'est-à-dire que dans un processus de recherche d'information, les documents retournés sont, généralement, ceux qui contiennent le plus de termes de la requête originale (notion de correspondance entre termes de la requête et documents), et, selon les modèles, la pertinence est évaluée en fonction de la fréquence des termes de la requête dans les documents. Mais malgré cette tendance globale des disparités sont présentes selon les requêtes et selon les variables.

À ce stade d'analyse nous avons examiné, notamment, la pertinence des documents, autrement dit la corrélation entre les niveaux de pertinence et la fréquence des variables dans les documents rapportés. La section suivante portera plutôt sur la

pertinence des requêtes, c'est-à-dire vérifier les corrélations entre la pertinence et les différentes variables impliquées (*termes de la requête, variantes dérivées, variantes fléchies et longueur du* texte). Le calcul de ces corrélations se fera par le biais du test statistique Jonckheere-Terpstra. Le choix de ce test et ses caractéristiques sont présentés au chapitre III qui porte sur la méthodologie.

### 4.1.4 Test de Jonckheere-Terpstra

Nous présentons, au Tableau 4.7, les résultats portant sur les corrélations entre les degrés de pertinence et chacune des variables. L'application du test statistique de Jonckheere-Terpstra à ces résultats va nous permettre de vérifier l'intérêt des scores obtenus et leur impact par rapport aux différentes variables impliquées (*termes de la requête, variantes dérivées, variantes fléchies et longueur du texte*). Il est à signaler que la variable *longueur du texte* n'intervient pas dans les corrélations avec les autres variables, en l'occurrence les termes de la requête, les variantes dérivées et les variantes fléchies. Par ailleurs, lors de la présentation des données brutes (Tableau 4.3), nous avons utilisé la variable *longueur du texte* pour calculer les pourcentages correspondant aux valeurs obtenues par chaque variable retenue dans l'étude, mais ce n'est pas le cas avec l'application du présent test statistique. Le test statistique adopté ici établit seulement une corrélation entre le score de pertinence et la fréquence de chacune des variables y compris la variable *longueur du texte*. Certes, la prise en compte de cette dernière variable dans les corrélations avec les autres variables (c'est-à-dire une corrélation entre trois variables, par exemple, *degré de pertinence, variantes dérivées* et *longueur du texte*), donnerait, probablement, plus de poids et de pertinence aux résultats. Mais faute de test statistique adéquat pour mettre en œuvre une telle démarche, nous nous contentons des résultats issus du présent test statistique (Tableau 4.7).

**Tableau 4.7 : Résultats du test de Jonckheere-Terpstra**

| | % | % | % | % |
|---|---|---|---|---|
| Requête | Termes | Dérivation | Flexion | Texte |
| 1 | -- | 0.0033*** | 0.4212 | 0.0070*** |
| 2 | -- | -- | 0.5500 | 0.0108** |
| 3 | 0.0009*** | -- | 0.0635* | -- |
| 4 | 0.3164 | 0.2260 | 0.4266 | -- |
| 5 | 0.0208** | 0.0022*** | -- | -- |
| 6 | -- | 0.0514* | -- | 0.2265 |
| 7 | 0.0099*** | 0.1164 | -- | 0.0567* |
| 8 | 0.1154 | -- | -- | 0.0593* |
| 9 | -- | 0.0165** | -- | 0.0018*** |
| 10 | 0.0354** | 0.0434** | -- | 0.1079 |
| 11 | -- | -- | -- | 0.0002*** |
| 12 | 0.3694 | 0.0513* | 0.2790 | 0.2825 |
| 13 | 0.1720 | 0.0370** | -- | -- |
| 14 | 0.4200 | -- | 0.4732 | 0.2090 |
| 15 | 0.0847* | 0.0072*** | 0.2886 | -- |
| 16 | 0.1809 | 0.1666 | 0.0967* | 0.0619* |
| 17 | 0.1902 | 0.0076*** | -- | 0.1901 |
| 18 | -- | 0.1056 | | 0.2493 |
| 19 | 0.4725 | 0.1081 | 0.3038 | 0.0056*** |
| 20 | 0.5000 | 0.0635* | -- | 0.0175** |
| 21 | 0.0110** | 0.2180 | 0.0213** | -- |
| 22 | -- | 0.0044*** | -- | 0.0009*** |
| 23 | 0.2367 | 0.2543 | 0.1628 | -- |
| 24 | 0.0986* | 0.1115 | 0.1558 | 0.0309** |
| 25 | 0.3428 | 0.0038*** | 0.3774 | 0.0160** |
| 26 | 0.3411 | 0.0921* | 0.3019 | 0.0643* |
| 27 | -- | 0.0003*** | 0.0335** | 0.0003*** |
| 28 | 0.4596 | 0.0616* | 0.1166 | 0.3300 |
| 29 | 0.2473 | -- | 0.3126 | -- |
| 30 | 0.4468 | 0.1666 | 0.1912 | 0.0033*** |
| 31 | -- | 0.0020*** | 0.0108** | 0.0478** |
| 32 | 0.3695 | -- | -- | 0.0119** |
| 33 | 0.3699 | 0.1176 | -- | 0.0341** |
| 34 | 0.0675* | 0.2725 | 0.0137** | -- |
| 35 | 0.1894 | 0.3986 | -- | 0.0002*** |
| 36 | 0.0766* | 0.0569* | 0.3278 | 0.2721 |
| 37 | -- | 0.0691* | -- | 0.2320 |
| 38 | 0.4458 | 0.0021*** | 0.0271** | 0.0643* |
| 39 | 0.0644* | 0.0155** | | 0.1722 |
| 40 | 0.2597 | -- | 0.4126 | -- |
| 41 | 0.1403 | 0.0696* | 0.0126** | -- |
| 42 | -- | -- | 0.0709* | 0.0212** |
| 43 | 0.0048*** | 0.0244** | 0.0812* | 0.1538 |
| 44 | -- | -- | -- | 0.0072*** |
| 45 | 0.0006*** | 0.0350** | | 0.3181 |
| 46 | 0.0103** | 0.3018 | 0.4191 | 0.4192 |
| 47 | -- | 0.0797* | | 0.0182** |
| 48 | 0.2295 | 0.3286 | 0.0029*** | 0.2295 |

| | | | | |
|---|---|---|---|---|
| 49 | -- | 0.3823 | -- | 0.0256** |
| 50 | 0.0383** | 0.0970* | 0.3102 | 0.2731 |

Avant de commenter les résultats exposés au Tableau 4.7, nous présentons, d'abord, quelques points liés à ce test (Jonckheere-Terpstra) ; une explication plus détaillée du test en question est présentée au chapitre méthodologie (chapitre III). Le premier point concerne la valeur des astérisques ; en effet, chaque fois qu'une valeur obtenue est accompagnée d'un astérisque ou plus, cela signifie que la valeur en question est positive. À ce sujet, nous distinguons trois valeurs positives et chaque valeur est symbolisée par des astérisques allant de un à trois ; la valeur symbolisée par un seul astérisque (*) signifie que $p < 0.10$, c'est-à-dire que la valeur obtenue est significative et où le degré de probabilité atteint 10 %. La valeur symbolisée par deux astérisques (**) signifie que c'est une valeur significative dans la mesure où le degré de probabilité atteint 5 % ($p < 0.05$). La valeur symbolisée par trois astérisques (***) représente une valeur significative où le degré de probabilité atteint 1 % ($p < 0.01$). Il est à souligner que plus le degré de probabilité est faible plus le cœfficient est fort ; ainsi, par exemple, le degré de probabilité 1 % ($p < 0.01$) représente le degré le plus faible et par conséquent le coefficient le plus fort. Par ailleurs, le symbole (--) signifie que la statistique du test est inférieure à celle du test de Jonckheere-Terpstra sous hypothèse nulle ($\mathbf{H_0}$) où les échantillons sont uniformes et proviennent de la même population. Nous signalons, également, qu'il n'y a pas eu de mesure pour la variable *Flexion* (*variantes fléchies*), et cela pour les requêtes 18, 39, 45 et 47. Celles-ci n'ont enregistré, tout simplement, aucune variante fléchie.

Le Tableau 4.7 présente les résultats du test Jonckeere-Terpstra. Il s'agit des corrélations entre les niveaux de pertinence et la fréquence des variables : *termes de la requête, variantes dérivées, variantes fléchies et longueur du texte.* Les résultats obtenus se caractérisent par des variations entre les différentes variables. Ainsi, les corrélations entre le degré de pertinence et la fréquence des variables impliquées dans les tests présentent des tendances intéressantes. À cet égard, l'examen des valeurs obtenues montre que l'impact de la corrélation entre la pertinence et la fréquence des variables varie d'une variable à l'autre et aussi d'une requête à l'autre. Par exemple, la variable *variantes dérivées* obtient des valeurs positives dans 25 requêtes, ce qui est, à notre sens, significatif ; cela veut dire que plus la fréquence des variantes dérivées est

importante dans les documents rapportés, plus ces documents sont pertinents ; ce score est suivi de près par la variable *longueur du texte* où 24 requêtes obtiennent des valeurs positives ; cela signifie que plus un document est long plus il y a de forte chance que ce dernier soit pertinent. Cette corrélation est beaucoup moins significative pour les variables *termes de la requête et variantes fléchies*. Ces dernières enregistrent des résultats positifs dans seulement 14 et 11 requêtes respectivement. Ce qui est intéressant ici c'est le score enregistré par la variable *termes de la requête*. Ce score est intéressant parce qu'il est faible et, par le fait même, il diffère largement, certes dans un contexte d'analyse différent, des scores obtenus par cette même variable lors de la présentation des résultats bruts (Tableau 4.3).

De même, en se limitant à la comparaison des résultats associés aux trois variables (*termes de la requête, variantes dérivées* et *variantes fléchies*), nous constatons que la variable *variantes dérivées*, avec 25 requêtes où la corrélation est positive, est celle qui obtient le score le plus significatif. Cependant, les scores obtenus par le biais du présent test diffèrent de ceux des données brutes où la variable *termes de la requête* obtient des scores meilleurs que ceux enregistrés par la variable *variantes dérivées* (Tableau 4.3). Cette différence reviendrait, entre autres, au fait qu'il ne s'agit pas de la même démarche ; par exemple, les résultats du présent test statistique ne prennent pas en compte la longueur du document dans les corrélations, et nous sommes plus dans un contexte qui met en exergue la pertinence des requêtes que celle des documents. Par ailleurs, la variable *variantes fléchies* enregistre les scores les plus bas de toutes les variables à l'étude, c'est-à-dire que la corrélation est positive dans seulement 11 requêtes, ce qui confirme la tendance associée aux scores de cette variable enregistrés précédemment (Tableau 4.3).

Le Tableau 4.8 ci-dessous présente le résumé des résultats du test de Jonckheere-Terpstra :

**Tableau 4.8 : Résumé des résultats du test Jonckheere-Terpstra**

| | **P < 0.01 *** ** | **P < 0.05 ** ** | **P < 0.10 * ** | **Total** |
|---|---|---|---|---|
| **% termes** | 45, 3, 43, 7<br>**(4)** | 46, 21, 5, 10, 50<br>**(5)** | 34, 39,15, 24, 35<br>**(5)** | **14** |
| **% dérivation** | 27, 5, 1, 31, 38, 25, 22, 15, 17<br>**(9)** | 39, 9, 13, 10, 43, 45<br>**(6)** | 12, 6, 20, 36, 28, 37, 41, 47, 26, 50<br>**(10)** | **25** |
| **% flexion** | 48<br>**(1)** | 31, 41, 34, 21, 27, 38<br>**(6)** | 3, 42, 43,16<br>**(4)** | **11** |
| **% long. texte** | 11, 35, 27, 22, 9, 30, 19, 1, 44<br>**(9)** | 2, 20, 24, 31, 32, 25, 47, 42, 49, 33<br>**(10)** | 7, 8, 16, 26, 38<br>**(5)** | **24** |

Le Tableau 4.8 regroupe l'ensemble des requêtes ayant obtenu des valeurs positives à l'issue du test statistique. Ainsi, à chaque variable correspond le nombre de requêtes où la corrélation entre la fréquence de la variable et la pertinence est positive ; ces requêtes sont réparties en fonction de leur degré de probabilité (1 %, 5 % et 10 %). Globalement, la dérivation semble être la variable la plus pertinente dans la mesure où elle concerne la moitié des requêtes du corpus. Par exemple, selon le degré de probabilité, nous observons qu'au niveau du degré de probabilité 1 %, qui est le niveau le plus significatif, les variables *variantes dérivées* et *longueur du texte* enregistrent les meilleurs scores ; chacune des deux variables contient 9 requêtes appartenant au premier degré de probabilité. Les deux autres variables, en l'occurrence *termes de la requête* et *variantes fléchies*, obtiennent, au même niveau de probabilité, un score de 4 et 1 respectivement, ce qui n'est pas beaucoup notamment pour la variable *variantes fléchies*. À propos du deuxième degré de probabilité 5 %, les scores obtenus sont plus ou moins équilibrés surtout pour les trois premières variables, car la variable *termes de la requête* obtient le score de 5 requêtes et les deux variables obtiennent le score de 6 requêtes chacune. Par ailleurs, la variable *longueur du texte* obtient un score plus élevé que les autres en obtenant un total de 10 requêtes.

Finalement, pour le degré de probabilité 10 %, qui est le moins significatif, nous constatons que c'est la variable *variantes dérivées* qui contient le plus de requêtes appartenant à ce degré de probabilité, ce nombre atteint 10 requêtes. Les autres requêtes enregistrent des scores quasi-similaires, 5 requêtes pour la variable *termes de la requête*, 4 requêtes pour la variable *variantes fléchies* et 5 requêtes pour la variable *longueur du texte*. Toutefois, ce qui nous intéresse dans ce résumé c'est la somme de requêtes obtenant ces niveaux de probabilité qui sont, somme toute, significatifs.

Pour illustrer la variation caractérisant les scores obtenus par les diverses variables, nous proposons, ci-dessous (Figure 4.3), une présentation en pourcentage sous forme graphique.

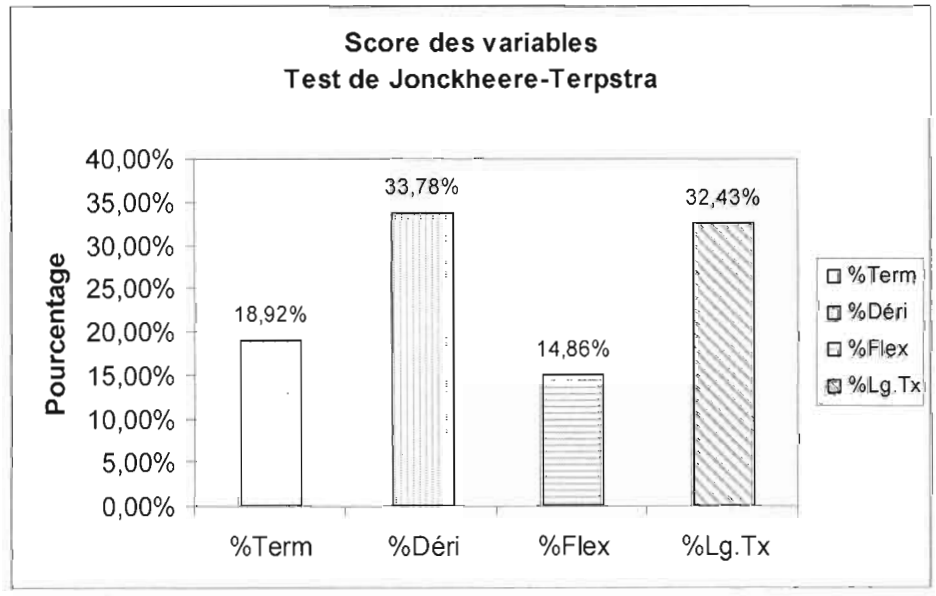



**Figure 4.3 : Résultats du test Jonckheere-Terpstra**

En comparant les scores des 4 variables issus du test non-paramétrique de Jonckheere-Terpstra, nous remarquons une certaine variation entre chacune des variables. Ainsi, la variable *variantes dérivées* obtient 33,78 %, le score le plus important en termes de pourcentage. Ce score est suivi de près par la variable *longueur du texte* qui atteint 32,43 %. Avec un pourcentage de 18,92 % la variable *termes de la requête* réalise un score moins significatif. Finalement, la variable *variantes fléchies*, avec un pourcentage de 14,86 %, enregistre le score le moins élevé parmi les variables étudiées, ce qui corrobore les résultats obtenus précédemment associés à cette dernière

variable. C'est-à-dire que l'impact des variantes fléchies sur les documents retournés n'est pas très important.

Une autre approche pour interpréter les résultats du test statistique consiste à établir une corrélation entre ces résultats (Tableau 4.7) et ceux liés aux scores de pertinence des différentes requêtes (Tableau 4.2). La tendance concernant les scores des différentes variables reste globalement invariable, et cela après l'exercice consistant à faire ressortir les requêtes qui obtiennent des scores positifs à la fois au test statistique (Tableau 4.7) et celui des scores de pertinence (Tableau 4.2). Ainsi, pour la variable *termes de la requête*, les requêtes ayant obtenu un score positif dans les deux tests sont au nombre de 4 (les requêtes 45, 46, 50 et 24), ce qui représente 5,41 % en termes de pourcentage. Pour la variable *variantes dérivées*, le nombre de requêtes ayant enregistré un score positif est 14 (les requêtes 27, 38, 25, 22, 17, 9, 45, 6, 20, 36, 28, 47, 26 et 50), ce qui équivaut à un pourcentage de 18,92 %. Pour la variable *variantes fléchies*, le score enregistré est celui de 3 requêtes, ce qui correspond à 4,05 % en termes de pourcentage. Finalement, pour la variable *longueur du texte*, le score obtenu est celui de 13 requêtes (les requêtes 11, 27, 22, 9, 30, 20, 24, 25, 47, 33, 8, 26 et 38), ce qui représente 17,57 %, en termes de pourcentage. Il est à noter que, concernant le test statistique (Tableau 4.7), une seule requête peut faire partie des résultats de plus d'une variable ; c'est le cas, par exemple, de la requête 35 qui figure aussi bien dans les résultats de la variable *termes de la requête* que celle de la *longueur du texte*.

La Figure 4.4, ci-dessous, est une illustration qui résume la tendance globale des différentes variables et cela après la mise en corrélation des résultats issus du test statistique de Jonckheere-Terpstra (Tableau 4.7) et ceux associés aux scores de pertinence des requêtes (Tableau 4.2).

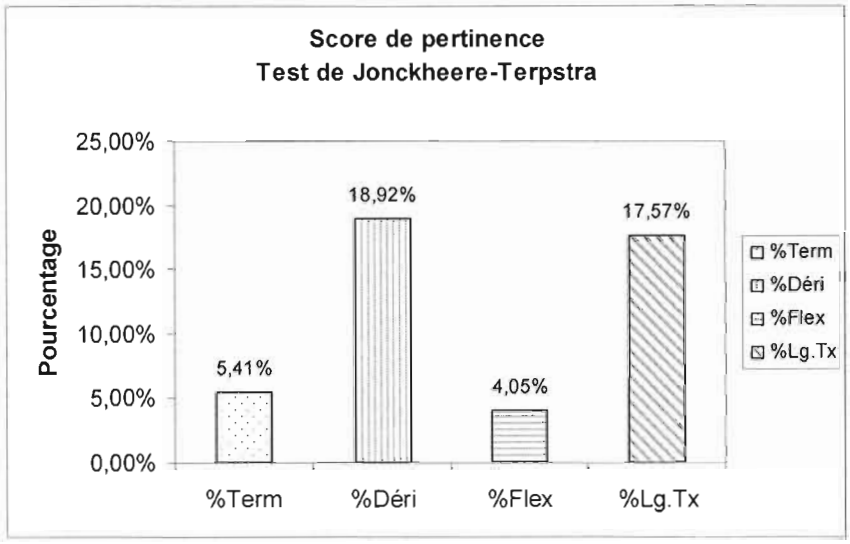



**Figure 4.4 : Scores de pertinence et test de Jonckheere-Terpstra**

En nous basant sur le même regroupement en 4 blocs de pertinence (Tableau 4.2), nous allons pouvoir vérifier s'il y a des corrélations entre les résultats du test statistique (Tableau 4.7) pour chacune des variables et les résultats des scores de pertinence des requêtes. Nous présentons au Tableau 4.9 ci-dessous les résultats obtenus en fonction des 4 blocs de pertinence.

**Tableau 4.9 : Test statistique, regroupement en 4 blocs de pertinence**

| Blocs | Termes/14 | Dérivés/25 | Fléchis/11 | Lg. Texte/24 |
|---|---|---|---|---|
| [7-18] | **4** (7, 15,43, 5) | **6** (15, 41, 37, 31, 43, 5) | **4** (41, 16, 31, 43) | **4** (7, 49, 16, 31) |
| [20-27] | **3** (21, 10, 3) | **4** (13, 1, 10, 12) | **2** (21, 3) | **3** (44, 1, 32) |
| [28-34] | **4** (50, 35, 39, 34) | **6** (17, 22, 9, 39, 36, 50) | **3** (42, 34, 48) | **4** (2, 35, 22, 33) |
| [35-47] | **3** (45, 46, 24) | **9** (28, 45, 25, 27, 6, 38, 20, 26, 47) | **2** (27, 38) | **10** (30, 11, 25, 27, 38, 20, 26, 47, 24, 8) |

Les résultats du Tableau 4.9 montrent qu'il y a une disparité dans chacun des 4 blocs et dans chacune des variables. Par exemple, au niveau des termes, 4 requêtes sur un

total de 14 appartiennent à la catégorie la plus mauvaise en termes de pertinence, c'est-à-dire le bloc [7-18] ; alors que 3 requêtes font partie du meilleur bloc [35-47]. Au niveau des catégories intermédiaires, les blocs [20-27] et [28-34], le nombre de requêtes enregistré atteint 3 et 4 respectivement. Au niveau des dérivés, nous obtenons 6 requêtes dans la catégorie la moins pertinente (le bloc [7-18]) et 9 requêtes dans la catégorie la plus pertinente (bloc [35-47], ce qui est intéressant. Concernant les catégories intermédiaires (blocs [20-27] et [28-34]), nous obtenons 4 et 6 requêtes respectivement. Pour les fléchis, nous enregistrons 4 requêtes dans la catégorie la moins pertinente et 2 requêtes dans la catégorie la plus pertinente. Au niveau des catégories intermédiaires, nous enregistrons 2 requêtes dans le bloc [20-27] et 3 requêtes dans le bloc [28-34]. Finalement, pour la longueur du texte, nous obtenons 4 requêtes dans la catégorie la moins pertinente (bloc [7-18]) et 10 requêtes dans la catégorie la plus pertinente (bloc [35-47]), ce qui est significatif. Au niveau des catégories intermédiaires, nous retrouvons 3 requêtes dans le bloc [20-27] et 4 requêtes dans le bloc [28-34]. Ce qui est intéressant de noter c'est le fait qu'au niveau des termes nous avons obtenu peu de requêtes appartenant à la catégorie la plus pertinente, ce qui diffère des résultats bruts. Par ailleurs, pour les dérivés, nous obtenons un nombre important de requêtes (9) appartenant à la catégorie la plus pertinente. De même, la variable longueur du texte enregistre un nombre important de requêtes (10) appartenant à la catégorie la plus pertinente. La tendance des fléchis n'a pas changé, c'est-à-dire que le nombre de requêtes appartenant à la catégorie la plus pertinente est 2, ce qui n'est pas significatif. Cependant, en faisant référence aux résultats bruts (Tableau 4.3), nous remarquons qu'il y a des disparités selon les requêtes et les variables. Par exemple, au niveau des termes, la requête 5 qui appartient à la catégorie la moins pertinente obtient un pourcentage significatif en termes de fréquence des termes (1,63%). Au niveau des dérivés, la requête 47 qui fait partie de la catégorie la plus pertinente enregistre un pourcentage non significatif en termes de fréquence des dérivés (0,29%).

Une autre lecture des résultats du test statistique (Tableau 4.7) consiste à établir des comparaisons entre les valeurs les plus significatives avec celles qui le sont moins, en termes de degré de probabilité, et cela pour chaque variable. Nous nous limitons à comparer 5 requêtes considérées comme les plus significatives avec 5 requêtes les moins significatives, et cela pour chaque variable. Ainsi, pour la variable *termes de la*

*requête,* 5 requêtes enregistrent un degré de probabilité le plus significatif ; les requêtes retenues sont : 3 (*exportation (de) médicaments dangereux*), 7 (*production minimale (au) Japon*), 43 (*conversion (de) (la) dette (pour) (la) Pologne*), 45 (*lutte contre (la) corruption*) et 46 (*protection (de) (l') environnement (au) sein (des) entreprises*). Parmi ces cinq requêtes seulement deux, en l'occurrence 45 et 46, obtiennent un score de pertinence positif (38 et 40). De même, en comparant les données de ces 5 requêtes présentées au Tableau 4.3, nous remarquons que les requêtes 45 et 46, avec des pourcentages de 2,71 % et 1,97 %, enregistrent de très bons résultats, au niveau de la fréquence des termes de la requête. Le troisième meilleur pourcentage parmi les 5 requêtes est celui obtenu par la requête 43, en l'occurrence 1,28 %. Par ailleurs, pour la même variable, les 5 requêtes ayant obtenu les valeurs les moins significatives au test statistique sont : 12 (*violence (des) adolescents*), 14 ((*la) politique économique (de) (la) Slovénie*), 20 (*limitations (des) importations (de) (l') UE*), 28 (*exploitation économique (du) fond marin*) et 38 ((*l') homosexualité (et) (la) loi*). Ce qui est étonnant c'est que, parmi ces 5 requêtes qui sont considérées par le test statistique comme non pertinentes, 3 requêtes ont un score de pertinence positif ; il s'agit des requêtes 20, 28 et 38. Celles-ci enregistrent des scores de 41, 38 et 40 respectivement. En faisant référence aux données brutes établies au Tableau 4.3, nous relevons que seulement 2 de ces requêtes enregistrent des pourcentages significatifs au niveau de la variable *termes de la requête* ; il s'agit des requêtes 12 et 14 avec un pourcentage de 1,61 % et 1,67 % respectivement.

Pour la variable *variantes dérivées* nous retenons les 5 requêtes les plus significatives par opposition à celles qui ne le sont pas, selon le test statistique. Les requêtes les plus significatives sont : 1 (*(la) culture écologique*), 5 (*politique extérieure (de) (l') Autriche*), 27 (*maltraitance (des) enfants*), 31 (*séparatistes catalans (et) galiciens*) et 38 (*(l') homosexualité (et) (la) loi*). Parmi les 5 requêtes seulement 2 enregistrent un score de pertinence positif ; il est question des requêtes 27 et 38. Ces dernières obtiennent un score de 39 et 40 respectivement. Par rapport aux résultats bruts du Tableau 4.3, seulement 2 des 5 requêtes enregistrent des bons pourcentages, en termes de fréquence des dérivés. Les deux requêtes sont 27 et 38 qui enregistrent des pourcentages de 0,90 % et 1,87 % respectivement. Par ailleurs, les 5 requêtes ayant les valeurs les moins significatives sont : 4 (*criminalité féminine*), 21 (*unité franco-allemande*), 23 (*(les) accidents (de) (la) route*), 48 (*(l') industrie européenne (du) film*)

et 49 (*normes (de) protection professionnelle*). Parmi les 5 dernières requêtes, une seule, en l'occurrence la requête 48, obtient un score de pertinence positif (34). L'examen des données brutes associées à ces requêtes montre que seulement la requête 4 obtient un pourcentage intéressant en termes de fréquence des variantes dérivées ; ce pourcentage est de l'ordre de 1,55 %, suivi de loin par la requête 48 avec un pourcentage de 0,65 %.

Pour la variable variantes fléchies, les 5 requêtes considérées comme les plus significatives sont : 31 (séparatistes catalans (et) galiciens), 34 (accidents (dans) (l') industrie minière), 38 ((l') homosexualité (et) (la) loi) 41 (ouvriers étrangers (en) Europe) et 48 ((l') industrie européenne (du) film). Parmi les 5 requêtes seulement 2 enregistrent un score de pertinence positif, il s'agit des requêtes 38 et 48 qui obtiennent un score de 40 et 34 respectivement. L'analyse des données brutes liées à ces requêtes indique qu'il y a une grande variation au niveau de la fréquence des variantes fléchies ; à ce sujet, seulement une requête, en l'occurrence 48, avec un pourcentage de 1,36 %, enregistre un score assez significatif. Les autres requêtes enregistrent des scores relativement bas. Par ailleurs, les 5 requêtes où la corrélation entre la variable et la pertinence est la moins significative sont : 1 *((l a) culture écologique*), 2 (*anti-sémitisme (en) Allemagne (après) 1945*), 4 (*criminalité féminine*), 14 *((la) politique économique (de) (la) Slovénie*) et 46 (*protection (de) (l') environnement (au) sein (des) entreprises*). En examinant les 5 requêtes par rapport à leur score de pertinence, nous remarquons qu'une seule requête, parmi les 5, enregistre un score de pertinence positif ; il s'agit de la requête 46, avec un score de 40. L'examen des données brutes liées à ces requêtes indique qu'au niveau des variantes fléchies seulement une requête enregistre un pourcentage significatif ; il s'agit de la requête 14 avec un score de 0,41 %.

Finalement, pour la variable *longueur de texte,* les 5 requêtes considérées parmi les plus significatives, en termes de corrélation, sont : 9 (*jeunesse (et) politique*), 11 (*charte sociale européenne*), 27 (*maltraitance (des) enfants*), 30 ((*les) débris spatiaux*) et 35 (*transport public local*). À propos du score de pertinence, 4 de ces 5 requêtes obtiennent un score positif, ce qui est intéressant dans la mesure où il y a une sorte de corrélation positive entre les requêtes les plus significatives (Tableau 4.7) et les scores de pertinence correspondants (Tableau 4.2). Cette tendance est quasiment absente dans

les autres variables examinées où diverses variations ont été constatées. Les 5 requêtes, dont les valeurs sont les moins significatives liées à cette variable (*longueur de texte*), sont : 12 (*violence (des) adolescents*), 36 *((la) pollution causée (par) (l') automobile*), 45 (*lutte contre (la) pollution*), 46 (*protection (de) (l') environnement (au) sein (des) entreprises*) et 50 (*traitement (des) déchets nucléaires*). Ce qui important ici, c'est qu'au niveau de cette variable (*longueur du texte)*, nous constatons que 4 requêtes, dont les valeurs sont moins significatives au test statistique, obtiennent un score de pertinence positif au Tableau 4.2. À la différence des autres variables, la comparaison des résultats du test statistique avec les résultats bruts (Tableau 4.3) associés à la variable *longueur du texte* n'a pas été prise en compte. La raison c'est qu'au niveau des données brutes cette variable est utilisée pour calculer les pourcentages de fréquence des trois autres variables, en l'occurrence *termes de la requête, variantes dérivées et variantes fléchies.*

Pour résumer, nous pouvons dire que les résultats du test statistique sont caractérisés par d'importantes variations. Alors, si le nombre de requêtes ayant obtenu des valeurs significatives est assez important au niveau des variables *variantes dérivées* et *longueur du texte*, ce n'est pas le cas pour les variables *termes de la requête* et *variantes fléchies*, qui, elles, obtiennent beaucoup moins de requêtes dont les valeurs sont significatives, en termes de corrélation entre la fréquence et la pertinence. Ce qui est intéressant de noter ici c'est le score faible enregistré par la variable *termes de la requête*, c'est-à-dire que le nombre de requêtes, dont la corrélation entre la fréquence et la pertinence est significative, est faible. Cela contraste avec ce qui est enregistré dans les résultats bruts. En effet, la comparaison des résultats du test statistique avec les données brutes a donné lieu à diverses variations et contrastes. Par exemple, la variable *termes de la requête* réalise les meilleurs scores aux résultats bruts ; situation qui a changé avec le test statistique. Cette variation apparaît également lors de la comparaison des résultats du test statistique avec les scores de pertinence des requêtes établis dans les données brutes. À cet égard, nous avons relevé que plusieurs requêtes, considérées comme significatives selon le test statistique, n'ont pas obtenu un score de pertinence positif, en faisant référence au Tableau 4.2. De même, de nombreuses requêtes ayant enregistré des valeurs significatives au test statistique, toutes variables confondues, enregistrent, dans les données brutes, des pourcentages faibles en termes de fréquence des variables. Dans le même ordre d'idée, des requêtes avec des valeurs

non significatives enregistrent des pourcentages significatifs au niveau de la fréquence des variables. Globalement, les résultats obtenus varient d'une requête à l'autre et d'une variable à l'autre.

La section suivante reprendra l'examen des résultats obtenus et en établira une analyse linguistique. Un intérêt particulier sera porté sur la structure syntaxique des requêtes ainsi que sur la nature et la qualité des variantes dérivées.

## 4.2 Analyse linguistique

La présente section se propose d'établir une analyse linguistique des résultats obtenus après l'application du test statistique de Jonckheere-Terpstra. L'objectif de cette démarche consiste à examiner l'impact de telle ou telle variable sur les résultats obtenus. À cet égard, nous allons analyser le contexte syntaxique ainsi que les variations morphologiques associés aux requêtes, et cela pour chacune des variables impliquées dans l'étude ; il s'agit des variables *termes de la requête, variantes dérivées et variantes fléchies*.

### 4.2.1 Structure syntaxique des termes

La variable *Termes de la requête* représente les termes de la requête de base tels qu'exprimés par l'utilisateur. Ces termes, comme nous l'avons indiqué au chapitre méthodologie, sont extraits des documents retournés suite à l'interrogation du moteur de recherche. Le but consiste à établir la fréquence de ces termes et leur impact sur la pertinence des documents. En nous basant sur les résultats obtenus, nous observons que la présence de ces termes dans les documents rapportés est significative, en termes de fréquence, et que la corrélation de cette variable avec les niveaux de pertinence donne des résultats qui peuvent être qualifiés de positifs, surtout au niveau des résultats bruts (Tableaux 4.3). Nous considérons que cette tendance positive est normale dans la mesure où le système de recherche d'information est basé généralement sur un processus d'appariement entre les mots de la requête exprimés par l'utilisateur et ceux représentant le contenu de chaque document. Toutefois, si cette corrélation paraît, globalement, évidente dans les données brutes, elle l'est moins après l'application du test statistique. Autrement dit, les résultats enregistrés par cette variable à ce test diffèrent de ceux obtenus aux résultats bruts.

À présent, nous nous proposons d'établir une description de la structure syntaxique des termes associés à l'ensemble des requêtes de notre corpus. Les structures syntaxiques établies sont :

- [N1-N2-ADJ], cette structure représente, par exemple, la requête 3 (*exportation (de) médicaments dangereux)*. N1 est la catégorie syntaxique associée au Mot1 (*exportation*), N2 est la catégorie associée à Mot2 (*médicaments*) et ADJ est la catégorie représentant le Mot3 (*dangereux*). Les mots de cette requête ne forment pas de liens syntaxiques. De même, la présence de ces termes dans les documents rapportés, comme expression, est très rare ; dans les 20 documents retournés, les termes de la requête apparaissent comme expression une seule fois. C'est le cas de l'exemple extrait du document 10 : «*la première dénonciation pénale internationale contre le trafic d'organes et contre l'exportation de médicaments dangereux dans les pays en voie de développement*». De même pour la requête 34 (*accidents (dans) (l') industrie minière*). Trois termes composent cette requête ; le terme *accidents* (N1) constitue l'élément central de la structure dans la mesure où il régit les deux autres termes, *industrie* (N2) et son modifieur *minière* (ADJ). Cette requête obtient un bon score de pertinence, mais la présence des termes de la requête dans les documents, comme étant une expression, est nulle. Par ailleurs, le groupe [N2-ADJ], de cette même requête, constitue un terme structuré ; cette paire de termes apparaît fréquemment en collocation dans les documents rapportés. D'autres requêtes de notre corpus ont cette même structure syntaxique ([N1-N2-ADJ]), il s'agit des requêtes 16 (*(les) communistes (au) Parlement européen*), 34 (*accidents (dans) (l') industrie minière*), 42 (*conséquences (de) (la) réunification allemande*), 49 (*normes (de) protection professionnelle*) et 50 (*traitement (des) déchets nucléaires*). Par ailleurs, en se référant aux scores de pertinence (Tableau 4.2), nous remarquons que deux requêtes ayant cette structure syntaxique appartiennent à la catégorie de pertinence la moins significative (bloc [7-18]), il s'agit des requêtes 16 et 49. Au niveau de la catégorie intermédiaire, nous enregistrons la présence de 4 requêtes, la requête 3 qui fait partie du bloc [20-27] et les requêtes 34, 42 et 50 qui font partie du bloc [28-34]. Cependant, aucune requête ne fait partie de la catégorie de pertinence la plus significative (bloc [35-47]). Par ailleurs, en nous basant sur les résultats du test statistique, nous relevons également des variations. Par exemple, les requêtes ayant la structure syntaxique [N1-N2-ADJ] entraînent des résultats positifs qui sont répartis selon les différentes

variables. Ainsi, trois requêtes sont présentes dans la variable *termes* (requêtes 3, 34 et 50), une requête est présente dans la variable *dérivation* (requête 50), 4 requêtes sont présentes dans la variable *flexion* (requêtes 3, 16, 34 et 42) et 3 requêtes intègrent la variable *longueur du texte* (requêtes 16, 42 et 49).

- [N1-ADJ-N2], représente la structure de la requête 5 (*politique extérieure (de) (l') Autriche)*. N1 est associé au M1 (*politique*), ADJ représente M2 (*extérieure*) et N2 représente M3 (*Autriche*). La requête 7 (*production minimale au Japon*) possède la même structure syntaxique. Dans les deux requêtes, les liens syntaxiques entre les trois mots ou catégories sont minimes ; ces liens sont plus présents entre les deux premiers mots de la requête N1-ADJ. Les termes des deux requêtes n'apparaissent ensemble, comme expressions, que dans de rares fois : 3 fois (documents 2, 3 et 19) pour la requête 5 (*politique extérieure de l'Autriche*) et une fois (document 1) pour la requête 7 (*production minimale au Japon*). Font partie aussi de cette catégorie les requêtes 14 *((la) politique économique (de) (la) Slovénie*), 32 (*coopération internationale (des) entreprises*), 41 (*ouvriers étrangers (en) Europe*) et 48 *((l') industrie européenne (du) film*).

En fonction des 4 blocs de pertinence établis précédemment, nous notons que les requêtes ayant cette structure syntaxique [N1-ADJ-N2] sont présentes dans la catégorie de pertinence la moins significative, c'est le cas des requêtes 5, 7 et 41 et dans les deux catégories intermédiaires (les requêtes 14 et 32 dans le bloc [20-27] et la requête 48 dans le bloc [28-34]). Aucune requête ne fait partie de la catégorie la plus pertinente (bloc 35-47). Par ailleurs, les résultats positifs caractérisant cette structure au niveau du test statistique sont variables, c'est-à-dire qu'ils varient d'une requête à l'autre et d'une variable à l'autre. Au niveau des termes, deux requêtes obtiennent des scores positifs (requêtes 5 et 7) ; au niveau des dérivés, deux requêtes enregistrent des scores positifs (requêtes 5 et 41) ; au niveau des fléchis, deux requêtes obtiennent des scores positifs (requêtes 48 et 41) ; finalement, au niveau de la longueur du texte, deux requêtes également obtiennent des scores positifs (requêtes 7 et 32).

- [N1-N2], il s'agit de la structure la plus fréquente du corpus ; elle est présente dans 13 requêtes. Il s'agit des requêtes 10 (*criminalité (des) adolescents*). 2 (*anti-sémitisme (en) Allemagne après 1945*), 9 (*jeunesse (et) politique*), 12 (*violence (des) adolescents*), 18 (*famine (au) Soudan*), 19 (*protection (des) animaux*), 22

(*immigration (et) racisme*), 23 (*(les) accidents (de) (la) route*), 25 (*(les) effets (de) (la) déforestation*), 27 (*maltraitance (des) enfants*), 33 (*réfugiés (de) Bosnie*), 38 (*(l') homosexualité (et) loi*) et 45 (*lutte (contre) (la) corruption*). Les termes de ces requêtes ne forment pas une expression. Autrement dit, ils n'apparaissent que rarement en cooccurrence dans les documents rapportés. Toutefois, les termes de la requête 45 apparaissent d'une façon fréquente en cooccurrence dans les documents rapportés. En terme de pertinence, nous remarquons que les requêtes ayant cette structure [N1-N2] ont généralement un score de pertinence positif, c'est-à-dire que, selon les 4 blocs de pertinence, nous retrouvons 9 requêtes (10, 12, 19, 2, 9, 18, 22, 23 et 33) dans la catégorie intermédiaire de pertinence et 3 requêtes (25, 27 et 45) dans la catégorie la plus pertinente. Selon les résultats du test statistique cette structure obtient des valeurs positives dans diverses variables. Par exemple, pour la variable *termes*, nous notons la présence d'une seule requête (10) ; pour la variable *dérivation*, nous constatons la présence de nombreuses requêtes (9, 10, 12, 22, 25, 27, 38 et 45) ; pour la variable *flexion*, nous enregistrons 2 requêtes (27 et 38), et pour la variable *longueur du texte*, nous enregistrons également de nombreuses requêtes ayant obtenus des scores positifs (requêtes 9, 19, 22, 25, 27, 33 et 38).

- [N1-N2-N3], cette structure syntaxique est la deuxième la plus fréquente dans le corpus ; elle est associée à 11 requêtes. Il s'agit des requêtes 15 (*déportation (des) étrangers (en) Autriche*), 39 (*processus (de) paix (au) Moyen-Orient*) 43 (*conversion (de) (la) dette (pour) (la) Pologne*), 6 (*conditions (de) vie (des) immigrants*), 8 (*extrémisme (de) droite (et) racisme*), 13 (*législation (sur) (la) protection (de) (la) nature*), 20 (*limitations (des) importations (de) (l') UE*), 29 (*exportation (d') armes (à) (la) Turquie*), 37 (*vitesse (sur) (les) autoroutes (en) Suisse*), 40 (*effets (du) chocolat (sur) (la) santé*) et 47 (*maintien (de) (la) paix (par) (l') OUA*). Parmi ces requêtes, c'est la requête 39 dont les termes apparaissent, d'une façon fréquente, en cooccurrence dans les documents rapportés. En termes du score de pertinence, les résultats sont variables. Ainsi, les requêtes appartenant à la catégorie la moins pertinente (bloc [7-18]) sont au nombre de trois (15, 37 et 43), les requêtes appartenant à la catégorie intermédiaire sont au nombre de trois (13, 29 et 39) et les requêtes appartenant à la catégorie la plus pertinente sont au nombre de 5 (6, 8, 20, 40 et 47). Par rapport au test statistique, les résultats positifs liés à cette structure varient en fonction des variables. Ainsi, nous relevons la présence de trois requêtes au niveau de

la variable *termes* (requêtes 15, 39 et 43), une présence significative de requêtes au niveau de la variable *dérivation*, c'est-à-dire que les requêtes ayant des valeurs positives atteint le nombre de 8 (6, 13, 15, 20, 37, 39, 43 et 47), ce qui est intéressant. Au niveau de la variable *flexion,* nous relevons la présence d'une seule requête (43), et au niveau de la variable *longueur du texte*, nous relevons la présence de trois requêtes (8, 20 et 47).

- [N1-ADJ], cette structure syntaxique caractérise les requêtes 1 (*(la) culture écologique*), 4 (*criminalité féminine*), 21 (*unité franco-allemande*), 24 (*éducation sexuelle*), 26 (*terrorisme international*) et 30 (*(les) débris spatiaux*). Le plus souvent, la structure syntaxique formée d'un substantif suivi d'un adjectif constitue un terme, au sens d'expression. Alors, est-ce que cette structure est un gage de pertinence ? Ce n'est pas toujours le cas ; par exemple, le score de pertinence de la requête 21 n'est pas très élevé ; de même, les termes de cette requête sont rarement présents comme expression dans les documents retournés. Par contre, la requête 24, qui a la même structure syntaxique, obtient un score de pertinence élevé et les termes de la requête sont, fréquemment, en collocation dans les documents rapportés. Nous pensons que les contre-performances de la requête 21, même si elle est syntaxiquement un terme structuré, seraient liées à l'ambiguïté sémantique du terme *unité* qui constitue la 'tête' régissant le modifieur *franco-allemande*. Par ailleurs, en nous basant sur les 4 blocs de pertinence, nous notons des résultats intéressants, c'est-à-dire que parmi les 6 requêtes ayant la structure syntaxique [N1-ADJ], 3 requêtes (1, 21 et 4) appartiennent à la catégorie intermédiaire de la pertinence, alors que 3 autres requêtes (26, 30 et 24) font partie de la catégorie la plus pertinente. Par rapport au test statistique, les résultats positifs varient en fonction des variables. Ainsi, deux requêtes ayant obtenu des valeurs positives sont présentes au niveau de la variable *termes* (requêtes 21 et 24), deux requêtes sont présentes au niveau de la variable *dérivation* (requêtes 1 et 26), une requête est présente au niveau de la variable *flexion* (requête 21) et 4 requêtes sont présentes au niveau de la variable *longueur du texte* (requêtes 1, 24, 26 et 30).

- [N1-ADJ1-ADJ2], cette structure représente, par exemple, la requête 35 (*transport public local*). Trois éléments composent cette requête, N1 (*transport*) qui représente la 'tête' du groupe, ADJ1 (*public*) premier modifieur et ADJ2 deuxième modifieur. Cette requête obtient un score moyen de pertinence. De même, la fréquence d'apparition des

termes de la requête, comme expression, est importante. Deux autres requêtes du corpus ont la même structure syntaxique, il s'agit des requêtes 11 (*charte sociale européenne*) et 31 (*Séparatistes catalans (et) galiciens*). Notons que les termes de la requête 11 apparaissent comme expression d'une façon fréquente dans les documents rapportés ; ce qui n'est pas le cas pour les termes de la requête 31, et cela malgré la présence de la paire [N1-ADJ1] dans la structure syntaxique qui est censée favoriser la cooccurrence des termes. En termes de pertinence, nous notons des variations selon les blocs de pertinence ; ainsi, la requête 31 appartient à la catégorie de pertinence la moins significative, la requête 35 appartient à la catégorie de pertinence intermédiaire et la requête 11 fait partie de la catégorie de pertinence la plus significative. Par rapport au test statistique, la présence de ces trois requêtes varie en fonction des variables. Ainsi, la requête 35 enregistre des valeurs positives au niveau des variables *termes* et *longueur du texte*, la requête 31 obtient des valeurs positives au niveau des variables *dérivation* et *flexion* et la requête 11 est positive au niveau de la variable *longueur du texte*.

-[N1-N2-N3-N4], cette structure syntaxique caractérise la requête 46 (*protection (de) (l') environnement (au) sein (des) entreprises*). Quatre éléments composent cette requête, N1 (*protection*), N2 (*environnement*), N3 (*sein*) et N4 (*entreprises*). Le score de pertinence obtenu par cette requête est significatif, c'est-à-dire que cette requête appartient à la catégorie de pertinence la plus significative (bloc [35-47]. Par rapport au test statistique, cette requête est présente (positive) au niveau d'une seule variable, *termes*. Par ailleurs, la présence des termes de la requête 46, comme expression, dans les documents retournés est nulle. Toutefois, nous notons une fréquence élevée où les deux premiers termes (*protection (de) (l') environnement*) apparaissent en cooccurrence.

- [N1-N2-ADJ-N3-N4], cette structure représente la requête 17 (*Destruction (de) (la) forêt tropicale (en) Amérique (du) Sud*). Cette requête est constituée de 5 mots, la plus longue requête de notre corpus. Les mots de la requête ne forment pas ici un terme ou une expression. La structure syntaxique complexe, c'est-à-dire la longueur de la requête, empêche que les termes de la requête apparaissent en cooccurrence dans les documents rapportés. La preuve c'est que la paire de mots *forêt tropicale,* qui forment un nom suivi d'un adjectif, apparaît fréquemment en collocation. En termes de

pertinence, cette requête fait partie de la catégorie de pertinence intermédiaire (bloc [28-34]). Par rapport au test statistique, la requête 17 est positive uniquement au niveau de la variable *dérivation*.

- [N1-ADJ1-N2-ADJ2], cette structure syntaxique caractérise les requêtes 28 (*Exploitation économique (du) fond marin*) et 44 (*statut militaire (de) (l') Allemagne unifiée*). Quatre mots, appartenant aux catégories nominale et adjectivale, composent chacune des requêtes. Les quatre mots constituent deux blocs de termes qui peuvent apparaître en cooccurrence dans les documents rapportés, il s'agit des paires syntaxiques N1-ADJ1 et N2-ADJ2. En termes de pertinence, la requête 44 appartient à la catégorie de pertinence intermédiaire, tandis que la requête 28 fait partie de la catégorie de pertinence la plus significative. Par rapport au test statistique, nous relevons la même variation, c'est-à-dire que la requête 44 est positive au niveau de la variable *longueur du texte,* alors que la requête 28 est positive au niveau de la variable *dérivation*.

- [N1-V1-N2], cette structure syntaxique caractérise la requête 36 (*(La) pollution causée (par) (l') automobile*). Les trois mots sont de nature nominale pour *pollution* et *automobile*, et verbale pour le mot *causée*. Il est question ici d'une construction passive, c'est-à-dire que c'est le sujet *automobile* qui subit l'action. Les constituants de la requête sont liés syntaxiquement. Au niveau du score de pertinence, la requête 36 appartient à la catégorie de pertinence intermédiaire (bloc [28-34], alors qu'au niveau du test statistique, cette requête est positive seulement au niveau de la variable *dérivation*.

En somme, nous notons que la variation caractérise l'ensemble des structures syntaxiques des termes de requêtes. Cette variation est constatée aussi bien au niveau du score de pertinence (Tableau 4.2) qu'au niveau des résultats du test statistique (Tableau 4.7). Toutefois, nous remarquons que les deux structures les plus fréquentes de notre corpus, en l'occurrence [N1-N2] et [N1-N2-N3], enregistrent, globalement, des résultats significatifs, et cela autant au niveau du score de pertinence qu'au niveau du test statistique. Cela veut dire que la présence de ces deux contextes syntaxiques a un impact sur la pertinence des documents rapportés ; autrement dit, il semble avoir des liens entre les structures syntaxiques des requêtes, notamment les structures [N1-N2] et [N1-N2-N3], et la pertinence des documents récupérés.

### 4.2.2 Nature et qualité des dérivés

Les variantes morphologiques, notamment dérivées, permettent généralement d'améliorer le rappel. Ces variantes peuvent affecter les différents termes de la requête de base. Nous examinons, dans cette section, la nature et la qualité des variantes dérivées. Rappelons que notre démarche pour extraire les termes et les variantes morphologiques est basée sur une morphologie dite aléatoire (non contrôlée) ; cette démarche consiste à extraire l'ensemble des formes associées à un terme à partir d'une racine commune (forme tronquée). Vu la grande quantité de données, nous nous limitons à examiner seulement quelques exemples, cela pour montrer que nous pouvons trouver de tout ; c'est-à-dire le cas de requêtes où il y a des déviations et les autres où il y a moins ou pas de déviations.

Globalement, les variations morphologiques (dérivationnelles) affectant les mots de la requête de base peuvent prendre diverses formes : nominale, verbale, adjectivale et adverbiale. Par exemple, la requête 3 (*exportation (de) médicaments dangereux)* contient des variantes dérivées représentant différentes catégories : verbe, nom et adjectif. Pour le Mot1 (premier mot de la requête) *exportation*, un nombre important de dérivés a été rapatrié ; ces dérivés sont d'ordre nominal comme *exportateur, export, transport, importateur, ports, etc.* D'autres sont d'ordre verbal, c'est le cas de *exporte, importe, transporte, porte, comporte, etc.* Ces derniers apparaissent d'une façon fréquente dans les documents. Pour le Mot2 (deuxième mot de la requête) *médicaments,* nous relevons surtout des dérivés d'ordre adjectival, c'est l'exemple de *médicaux, médical, biomédical, etc.* Pour le Mot3 *dangereux,* les dérivés se font rares ; la seule variante dérivée enregistrée est de nature nominale (*danger*). Nous remarquons que les dérivés associés notamment au Mot1 [exportation] (*transport, importateur, importe, porte, comporte, etc.*) diffèrent des mots de base ; il s'agit des déviations qui peuvent entraîner des glissements sémantiques.

La requête 7 (*production minimale (au) Japon*) est un autre exemple qui montre la variation des dérivés caractérisant les termes de base et où nous constatons quelques déviations. Cette requête, qui obtient le score de pertinence le plus bas de toutes les requêtes, n'enregistre pas un pourcentage élevé de variantes dérivées (0,59 %). La nature de ces dérivés diffère en fonction des termes de base. Ainsi, pour le Mot1

*production*, elle est nominale, comme c'est le cas de *produit, producteur, productivité, reproduction, etc.*, verbale, comme c'est le cas de *produit, produisent, produira, etc.*, et adjectivale, c'est l'exemple de *productif.* Ce dernier dérivé apparaît rarement dans les documents rapportés. Toutefois, les variantes d'ordre nominal et verbal apparaissent d'une façon fréquente dans les documents. Pour le Mot2 *minimale*, nous relevons deux types de dérivés, nominal et verbal. C'est l'exemple de *minimum* et *minimalistes* pour le nominal, et *minimisés, minimisent, etc.* pour le verbal. Finalement, pour le Mot3 (troisième mot de la requête) *Japon*, une seule catégorie de dérivés est relevée, il s'agit de la catégorie adjectivale. C'est l'exemple de l'adjectif *japonais* (*e*) (*es*). Ce dérivé apparaît fréquemment dans les documents rapportés. Les déviations sémantiques relevées sont associées notamment au Mot2 [*minimale*] (*minimisés*, *minimalistes*, *etc.*). Il est à noter que la non pertinence des documents rapportés est liée notamment aux contraintes imposées par l'énoncé narratif de la requête. En effet, la partie narrative associée à cette requête (7) est ambiguë et imprécise. Autrement dit, le thème sur lequel porte l'énoncé narratif n'est pas défini clairement, car nous ne savons pas de quoi il est question exactement.

Un autre exemple de déviation sémantique est lié à la requête 27 (*maltraitance (des) enfants*). Celle-ci enregistre un score de pertinence positif (bloc [35-47]) et un pourcentage significatif en termes de fréquence des dérivés (0,90 %). Nous relevons des variantes dérivées de nature variée. Pour le Mot1 (maltraitance), les dérivés sont d'ordre nominal (*traitement, traité, etc.*), adjectival (*maltraité, traitant, maltraitant, etc.*) et verbal (*maltraiter, traiter, etc.*). Pour le Mot2 (enfants), les dérivés se limitent à un seul qui est de nature nominale (*enfance*). Il est à noter que les dérivés associés au Mot1 sont plus nombreux et variés que ceux liés au Mot2. Notons que les déviations sémantiques constatées relèvent surtout du Mot1 [maltraitance] ; il s'agit de variantes dérivées *traité*, *traitant* et *maltraitant* (verbe ou substantif). Cela indique que la possibilité d'interchanger les dérivés avec les termes de la requête est dans ce cas une démarche risquée, car les glissements sémantiques seraient inévitables.

Par ailleurs, nous citons un autre exemple où les déviations sémantiques sont minimes, voire absentes. Il s'agit par exemple de la requête 15 (*déportation (des) étrangers (en) Autriche*). Cette requête comprend trois mots appartenant à la même catégorie syntaxique, le nom [N1-N2-N3]. Les variantes dérivées issues des termes de base ne

sont pas fréquentes, car elles atteignent seulement 0,34 % en termes de pourcentage. La nature des dérivés diffère d'un terme à l'autre. Pour le Mot1 les dérivés relevés sont d'ordre nominal (*déporté, déportés*) et verbal (*déporter, déportés, etc.*). Ces dérivés sont sémantiquement liés au terme de base (*déportation*) ; à ce niveau, le risque d'un glissement sémantique est minime. Pour le Mot2 *étrangers,* aucune variante dérivée n'a été relevée. Pour le Mot3 *Autriche,* les dérivés relevés sont surtout d'ordre adjectival (*autrichien, autrichienne, etc.*). Ainsi, cette requête, en plus de ne pas avoir un score de pertinence positif, n'est pas riche en termes de dérivés, ce qui explique peut-être l'absence des déviations. Une autre requête où nous n'avons pas noté de déviations sémantiques est la requête 40 (*effets (du) chocolat (sur) (la) santé*) ; cela est lié, notamment, à la structure de la requête qui forme une expression et aussi à la nature des mots de base. De même, les variantes dérivées associées aux mots de base de cette requête n'enregistrent pas un pourcentage intéressant en termes de fréquence (0,19%).

En somme, nous avons cité les quelques exemples ci-dessus afin de montrer la variabilité caractérisant les variantes dérivées. Cette variabilité porte en particulier sur la nature et la qualité des dérivés. À ce sujet, nous observons que la nature et la qualité des variantes dérivées dépendent notamment de la nature et de la qualité des mots de la requête de base ainsi que de sa structure syntaxique.

Nous allons examiner maintenant les dérivés obtenus pour les requêtes les plus pertinentes (Tableau 4.2). Nous nous limitons aux 5 requêtes qui obtiennent les meilleurs scores de pertinence ; il s'agit des requêtes 26, 47, 24, 8 et 40. Ainsi, la requête 26 (*terrorisme international*) enregistre un pourcentage de 0,99% en termes de fréquence des dérivés, ce qui est significatif. La majorité des dérivés sont de type adjectival (*terroriste, antiterroriste, etc.*). Aussi, les dérivés rapportés sont globalement proches sémantiquement des termes de la requête, c'est-à-dire qu'il n'y a pas de déviations sémantiques. Pour la requête 47 (*maintien (de) (la) paix (par) (l') OUA*), qui enregistre un pourcentage non significatif en termes de fréquence (0,29%), nous relevons des dérivés de type divers : verbal (*maintenir*), adjectival (*pacifique*) et adverbial (*pacifiquement*). Notons que cette requête présente des déviations sémantiques importantes, surtout au niveau du Mot3 (OUA) ; il s'agit de variantes comme *[Oua]gadougou, M[oua]mmar* ; ces derniers dérivés montrent les risques

auxquels peut être confronté le choix d'un morphologie non contrôlée. La requête 24 (*(l') éducation sexuelle*), qui fait partie de la catégorie la plus pertinente, enregistre un pourcentage significatif en termes de fréquence des dérivés (0,96%). Les dérivés sont de nature variée : nominale (*éducateur, éducatrice, sexualité, sexologue, homosexualité, etc.*), verbale (*éduquer, éduqués, sexualisée, etc.*) et adjectivale (*éducative, éducatif, sexué, etc.*) et adverbiale (*sexuellement*). Notons que les dérivés d'ordre nominal sont beaucoup plus fréquents que les autres catégories de dérivés. De même, cette requête ne comprend pas de déviations sémantiques. La requête 8 (*extrémisme (de) droite (et) racisme*), qui fait partie de la catégorie la plus pertinente, enregistre un pourcentage proche de la moyenne en termes de fréquence, 0,69% plus exactement (la moyenne est 0,70%). Les dérivés associés aux termes de cette requête sont de nature diverse : nominale (*raciste, extrémiste, antiracisme, etc.*), adjectivale (*raciste, extrémiste, etc.*) et adverbiale (*extrêmement*). Notons que les dérivés sont proches sémantiquement des termes de la requête, c'est-à-dire que les déviations ne sont pas présentes. Finalement, la requête 40 (*effets (du) chocolat (sur) (la) santé*) est une requête qui obtient le plus haut score de pertinence. Toutefois, elle enregistre un pourcentage de 0,19% en termes de fréquence des dérivés, ce qui est très peu. Les dérivés relevés sont de nature diverse : nominale (*chocolaterie, chocolatier, etc.*), adjectivale (*chocolatée, sanitaire, etc.*). De même, les dérivés associés aux termes de cette requête ne présentent pas de déviations sémantiques majeures.

En somme, malgré l'appartenance des 5 requêtes à la catégorie la plus pertinente, nous remarquons qu'il y a des variations selon les requêtes, selon la fréquence et selon la nature et la qualité des dérivés.

Désormais, nous examinons les variantes dérivées pour les 5 requêtes les moins pertinentes ; il s'agit des requêtes 7, 15, 41, 49 et 16. Ainsi, la requête 7 (*production minimale (au) Japon*) enregistre un pourcentage non significatif en termes de fréquence des dérivés (0,59%). Ces dérivés sont de nature variée : nominale (*minimum, productivité, etc.*), verbale (*minimiser, produire*) et adjectivale (*productif, japonais, etc.*). Par ailleurs, nous relevons des déviations sémantiques associées notamment au Mot2 de la requête (*minimale*) ; il s'agit, par exemple, des dérivés *minimiser et minimalistes.* La requête 15 (*déportation (des) étrangers (en) Autriche*) enregistre un pourcentage de 0,34% en termes de fréquence des dérivés, ce qui est non

significatif. Les dérivés relevés sont de nature variée : nominale (*déporté, déportés*), verbale (*déporter*) et adjectivale (*autrichien, autrichienne*). Au niveau de la qualité des dérivés, nous ne relevons pas de déviations sémantiques. Pour la requête 41 (*ouvriers étrangers (en) Europe*), nous notons le faible pourcentage obtenu en termes de fréquence des dérivés (0,25%). Les dérivés relevés varient d'un mot de la requête à l'autre et ils sont de nature diverse : nominale (*ouvrage, européiste*), verbale (*ouvrir, découvrir*) et adjectivale (*européen, européenne*). Toutefois, des déviations sémantiques sont bien présentes, surtout au niveau du Mot1 (*ouvriers*). Pour la requête 49 (*normes (de) protection professionnelle*), nous notons un pourcentage non significatif en termes de fréquence des dérivés (0,35%). La nature des dérivés associés aux termes de la requête est variée : nominale (*normalisation, profession, professeur, etc.*), verbale (*protéger*), adjectivale (*interprofessionnel, normatives, etc.*) et adverbiale (*normalement*). Au niveau de la qualité des dérivés, nous constatons que globalement les dérivés récupérés sont proches sémantiquement des termes de base. Toutefois, nous relevons une déviation associée au Mot3 [professionnelle], il s'agit de *profil.* Finalement, la requête 16 (*(les) communistes (au) parlement européen*), comme c'était le cas des autres requêtes les moins pertinentes, enregistre un faible pourcentage en termes de fréquence des dérivés (0,47%). La nature des dérivés récupérés est variée, nominale (*parlementarisme, Europe, communauté, communisme, etc.*), verbale (*parler*) et adjectivale (*parlementaire, commun, communautaire, etc.*). Notons que les dérivés récupérés sont globalement proches des termes de base, c'est-à-dire qu'il n'y a pas de déviations sémantiques majeures.

En somme, indépendamment de leur degré de pertinence (Tableau 4.2), les requêtes citées ci-dessus montrent que la nature et la qualité des variantes dérivées varient en fonction de la nature et la qualité des mots de la requête de base.

Nous procédons, désormais, à examiner la nature et la qualité des dérivés de 5 requêtes ayant obtenu les valeurs les plus significatives au test statistique (Tableau 4.7). Il s'agit des requêtes 1, 5, 27, 31 et 38. Ainsi, la requête 1 (*(la) culture écologique*) n'enregistre pas un pourcentage significatif en termes de fréquence des dérivés. Ces derniers, qui atteignent 0,57%, diffèrent d'un mot de la requête à l'autre et ils sont de nature diverse : nominale (*agriculture, agriculteur, horticulture, agro-écologie, écologie, etc.*), verbale (*cultiver*) et adjectivale (*culturel, culturelle, etc.*). En

termes de qualité, les dérivés liés aux mots de la requête sont proches sémantiquement ; en ce sens, nous ne relevons pas de déviations sémantiques. La requête 5 (*politique extérieure (de) (l') Autriche*) enregistre 0,31% comme pourcentage de la fréquence des dérivés, ce qui est non significatif. Les dérivés liés aux termes de cette requête sont de nature variée : nominale (*géopolitique, politicien, etc.*), adjectivale (*politico-culturel, autrichien, intérieure, etc.*) et adverbiale (*politiquement*). Notons que ces dérivés appartiennent au même champ sémantique que les termes de la requête, c'est-à-dire que nous ne relevons pas de déviations sémantiques majeures. La requête 27 (*maltraitance (des) enfants*), à la différence des requêtes précédentes, enregistre un pourcentage significatif en termes de fréquence des dérivés, 0,90% plus exactement. Les dérivés relevés sont de type varié : nominal (*traitement, enfance, etc.*), verbal (*maltraiter, traiter*) et adjectival (*traitant, maltraitant, etc.*). À part l'ambiguïté d'ordre polysémique liée à un dérivé comme *traité* (verbe ; nom), les dérivés ont globalement des liens sémantiques avec les termes de la requête ; autrement dit, nous ne relevons pas de déviations sémantiques. La requête 31 (*séparatistes catalans (et) galiciens*) enregistre un taux de dérivés très bas en termes de fréquence ; ce taux atteint le pourcentage de 0,26%. Les dérivés récupérés sont de nature variée : nominale (*séparatisme, séparation, Galicie, Galice, Catalogne, etc.*), verbale (*séparer*) et adverbiale (*séparément*). Notons que nous ne relevons pas de déviations sémantiques, c'est-à-dire que les variantes dérivées récupérées sont sémantiquement proches des termes de la requête. Finalement, pour la requête 38 (*(l') homosexualité (et) (la) loi*), nous notons qu'elle obtient un pourcentage de dérivés le plus élevé de toutes les requêtes, en termes de fréquence (1,87%). Les dérivés récupérés sont de différents types : nominal (*sexualité, hétérosexualité, sexe, etc.*) et adjectival (*sexué, sexuel, homosexuel, etc.*). Au niveau de la qualité des dérivés, nous constatons que globalement les variantes récupérés sont proches sémantiquement des termes de base ; toutefois, nous relevons quelques déviations liées au Mot2 [loi] comme c'est le cas de *loisirs, emploi, etc.* En somme, les 5 requêtes les plus significatives par rapport au test statistique présentent des variations au niveau de la fréquence des dérivés, de leur type et de leur qualité. Notons que seulement deux requêtes parmi les 5 obtiennent des pourcentages significatifs en termes de fréquence des dérivés, il s'agit des requêtes 27 et 38.

En somme, par le biais de ces exemples, nous constatons que la fréquence des dérivés, leur nature et leur qualité varient d'une requête à l'autre et d'un mot de la requête à l'autre ; cela indépendamment des résultats des requêtes par rapport à leur score au test statistique.

Nous examinons, désormais, les 5 requêtes où la dérivation obtient les valeurs les moins significatives au test statistique, il s'agit des requêtes 4, 21, 23, 48 et 49. Ainsi, la requête 4 (*criminalité féminine*), même si elle fait partie des requêtes les moins significatives, enregistre un pourcentage élevé en termes de fréquence des dérivés, 1,55% plus exactement. Au niveau de la nature des dérivés récupérés, nous relevons des dérivés de type nominal (*féministe, féminisme, féminité, crime, criminologie, etc.*), verbal (*criminaliser*) et adjectival (*criminel, criminologique, etc.*). Notons qu'en termes de qualité, ces dérivés sont globalement proches des termes de la requête de base ; en ce sens, nous ne relevons pas de déviations sémantiques. La requête 21 (*unité franco-allemande*) obtient un pourcentage de 0,28% en termes de fréquence des dérivés, ce qui est non significatif. Les dérivés récupérés varient d'un terme à l'autre et ils sont de nature variée : nominale (*union, unification, France-Allemagne, etc.*), verbale (*unir, réunir, réunifier*), adjectivale (*uni, unique, franco-lituano-allemande, etc.*) et adverbiale (*uniquement*). Au niveau de la qualité des dérivés, à part quelques déviations comme *unique, uniquement et franco-lituano-allemande*, les liens sémantiques entre les termes de la requête et les dérivés sont globalement proches. La requête 23 (*(les) accidents (de) la route*) n'enregistre pas un pourcentage significatif en termes de fréquence des dérivés ; ce pourcentage atteint seulement 0,40%. Les dérivés récupérés sont de nature variée : nominale (*accidentologie, accidentologue, autoroute, etc.*) et adjectivale (*accidentel, routier, accidenté, etc.*). Au niveau de la qualité des dérivés, nous ne relevons pas de déviations sémantiques parmi les dérivés récupérés. Ces derniers sont en relation sémantique avec les termes de base. La requête 48 (*(l') industrie européenne (du) film*) obtient un pourcentage de 0,65% en termes de fréquence des dérivés ; il s'agit d'un pourcentage légèrement bas par rapport à la moyenne qui est de 0,70%. Concernant les dérivés récupérés, ils sont de nature variée : nominale (*désindustrialisation, Europe, téléfilm, etc.*) verbale (*industrialiser*) et adjectivale (*industriel, paneuropéens, transeuropéens, etc.*). Au niveau de la qualité des dérivés, nous ne relevons pas de déviations sémantiques majeures, c'est-à-dire que globalement les dérivés récupérés sont en relation sémantique avec les termes de base.

Finalement, la requête 49 (*normes (de) protection professionnelle*) est une requête qui enregistre une fréquence de dérivés non significative, c'est-à-dire que le pourcentage obtenu atteint seulement 0,35%. La nature de ces dérivés varie d'un mot de la requête à l'autre : nominale (*normalisation, profession, professeur, professionnalisme, etc.*), verbale (*protéger*) adjectivale (*normal, anormaux, protégé, interprofessionnel, etc.*) et adverbiale (*normalement*). Globalement, les dérivés relevés sont proches sémantiquement des termes de base. Toutefois, nous relevons une déviation dans le cas du dérivé *profil.*

En somme, par le biais de ces exemples, nous observons que la fréquence des dérivés, leur nature et leur qualité varient d'une requête à l'autre et d'un mot de la requête à l'autre ; cela indépendamment des résultats des requêtes par rapport au test statistique (Tableau 4.7).

### 4.2.3 Variantes fléchies

Nous examinons, dans cette section, les variantes fléchies associées aux termes de requêtes de base. Rappelons que la flexion ne permet pas de générer de nouveaux mots, mais plutôt des variantes liées, par exemple, au genre et au nombre des mots de base. Notons, également, que la variable *variantes fléchies* est celle qui enregistre les résultats les moins significatifs, et cela aussi bien au niveau des résultats bruts qu'au niveau des résultats du test statistique. Ainsi, nous allons examiner l'impact de cette variable (*flexion*) en prenant comme exemple, d'une part, 5 requêtes les plus pertinentes et, d'autre part, 5 requêtes les moins pertinentes (Tableau 4.2). Par ailleurs, dans le même type d'exercices, nous allons examiner, d'abord, la flexion dans 5 requêtes les plus significatives et ensuite dans 5 requêtes les moins significatives, et cela par rapport au test statistique (Tableau 4.7).

Les 5 requêtes les plus pertinentes sont 26, 47, 24, 8 et 40. La requête 26 (*terrorisme international*) est une requête qui enregistre un pourcentage de fléchis de 0,23%, c'est un pourcentage faible, il est au-dessous de la moyenne (0,30%). Nous relevons que l'ensemble des fléchis récupérés sont associés au Mot2 de la requête (*international*), il s'agit de flexions en genre et en nombre (*internationale, internationaux, etc.*). La requête 47 (*maintien (de) (la) paix (par) (l') OUA*) obtient un pourcentage nul en termes de fréquence de variantes fléchies (0,00%). La requête 24 (*(l') éducation*

*sexuelle*), à la différence de la requête précédente, enregistre un pourcentage de 0,27% en termes de fréquence de variantes fléchies, ce qui est proche de la moyenne (0,30%). Ces variantes, en genre et en nombre, sont associées uniquement au Mot2 de la requête (*sexuelle*) ; c'est le cas par exemple de *sexuel, sexuels, sexuelle, etc.* La requête 8 (*extrémisme (de) droite (et) racisme*) obtient un pourcentage faible (0,20%) en termes de fréquence des variantes fléchies. Notons que ces variantes sont liées majoritairement au Mot2 de la requête (*droite*) ; il s'agit de flexion en genre et en nombre, comme c'est le cas de *droit, droits.* Une seule flexion est relevée par rapport au Mot1 (*extrémisme*), il s'agit de la flexion en nombre *extrémismes*. Finalement, la requête 40 (*effets (du) chocolat (sur) (la) santé*) obtient un pourcentage faible en termes de fréquence des fléchis, 0,21% plus exactement. Ces variantes fléchies (en nombre) sont associées au Mot1 (*effets*), c'est le cas de la variante *effet* et au Mot2 (*chocolat*), c'est le cas de *chocolats*.

Désormais, nous examinons les 5 requêtes les moins pertinentes, en l'occurrence les requêtes 7, 15, 41, 49 et 16. La requête 7 (*production minimale (au) Japon*) obtient un pourcentage de 0,08% en termes de fréquence des variantes fléchies, ce qui est insignifiant. Les flexions relevées, en genre et en nombre, sont liées au Mot1 (*production*) et au Mot2 (*minimale*). Il est question des fléchis *productions* et *minimal*. La requête 15 (*déportation (des) étrangers (en) Autriche*) enregistre un pourcentage faible en termes de fréquence des fléchis ; ce pourcentage atteint seulement 0,12%. Les flexions, en genre et en nombre, sont associées aux deux premiers mots de la requête, *déportation* et *étrangers*. C'est le cas des variantes comme *déportations, étranger, étrangère, étrangères.* La requête 41 (*ouvriers étrangers (en) Europe*) atteint un pourcentage de 0,25% en termes de fréquence des variantes fléchies, ce qui est non significatif. Ces variantes, en genre et en nombre, sont associées au Mot1 (*ouvriers*) et au Mot2 (*étrangers*) ; c'est l'exemple des variantes comme *ouvrier, ouvrière, ouvrières, étranger, etc.* La requête 49 (*normes (de) protection professionnelle*), à la différence des requêtes précédentes, enregistre un pourcentage significatif en termes de fréquence des variantes fléchies, 0,44% plus précisément. Ces variantes, en genre et en nombre, résultent des trois mots de la requête. C'est le cas, par exemple, de *norme, protections, professionnel, professionnels, etc.* Finalement, la requête 16 (*(les) communistes (au) parlement européen*) enregistre 0,73% comme pourcentage de fréquence des variantes fléchies, ce qui est très significatif. Ces variantes, en genre et

en nombre, sont associées au Mot1 (*communistes*) et au Mot3 (*européen*) ; c'est le cas des flexions comme *communiste, européens, européenne, européennes*.

En somme, du point de vue de la pertinence des requêtes, les deux types de requêtes, c'est-à-dire les 5 requêtes les plus pertinentes et les 5 requêtes les moins pertinentes, enregistrent, pour la majorité, des pourcentages non significatifs en termes de fréquence des variantes fléchies. La fréquence des flexions varie d'une requête à l'autre et également d'un mot à l'autre. Deux requêtes seulement obtiennent des pourcentages significatifs en termes de fréquence, il est question des requêtes 49 et 16 qui font partie des requêtes les moins pertinentes.

Nous passons à l'examen des variantes fléchies dans les 5 requêtes les plus significatives (variable flexion) par rapport au test statistique (Tableau 4.7). Il s'agit des requêtes 48, 31, 41, 34 et 21. La requête 48 (*(l') industrie européenne (du) film*) enregistre un pourcentage très significatif en termes de fréquence des variantes fléchies ; le score obtenu atteint 1,36%, c'est l'un des scores les plus élevés. Les flexions relevées, en genre et en nombre, sont liées aux trois mots de la requête. C'est le cas, par exemple, de *industries, européen, européens, européenne, films, etc.* La requête 31 (*séparatistes catalans (et) galiciens*) obtient un pourcentage de 0,11% en termes de fréquence des fléchis, ce qui est non significatif. Les variantes fléchies relevées varient en genre et en nombre et sont associées aux trois mots de la requête. C'est l'exemple de *séparatiste, catalan, catalane, galicien, galicienne, etc.* La requête 41 (*ouvriers étrangers (en) Europe*) enregistre un pourcentage de 0,25% en termes de fréquence des variantes fléchies ; c'est un score qui est au dessous de la moyenne (0,30%). Les variantes relevées varient en genre et en nombre et concernent surtout les deux premiers mots de la requête, *ouvriers* et *étrangers*. C'est le cas des flexions comme *ouvrier, ouvrière, étranger, étrangères, etc.* La requête 34 (*accidents (dans) (l') industrie minière*) enregistre un pourcentage de 0,44% en termes de fréquence des variantes fléchies. Il s'agit d'un score significatif. Ces variantes liées aux trois mots de la requête varient, selon les mots, en genre et en nombre ; c'est le cas, par exemple, de *accident, industries, minier, miniers, minière, etc.* La requête 21 (*unité franco-allemande*) enregistre un pourcentage de fréquence des fléchis légèrement au-dessous de la moyenne. Le pourcentage en question atteint 0,29%. Les variantes fléchies relevées varient en genre et en nombre en fonction des mots de la requête. C'est

l'exemple des variantes comme *unités, franco-allemand, franco-allemands, franco-allemandes*.

Concernant les requêtes les moins significatives par rapport au test statistique, nous examinons la flexion dans les 5 requêtes suivantes : 2, 14, 1, 46 et 40. La requête 2 (*anti-sémitisme (en) Allemagne (après) 1945*) enregistre un pourcentage nul, c'est-à-dire 0,00% en termes de fréquence des variantes fléchies. La requête 14 (*(la) politique économique (de) (la) Slovénie*), à la différence de la requête précédente, obtient un pourcentage de 0,41% en termes de fréquence des variantes fléchies, ce qui est significatif. Les flexions relevées, en nombre, sont associées aux deux premiers mots de la requête, c'est le cas des fléchis *politiques* et *économiques*. La requête 1 (*(la) culture écologique*) enregistre un pourcentage non significatif en termes de fréquence des variantes fléchies ; ce pourcentage atteint seulement 0,12%. Les flexions récupérées sont associées aux deux mots de la requête. Il s'agit uniquement de la flexion en nombre, c'est l'exemple de *cultures* et *écologiques*. La requête 46 (*protection (de) (l') environnement (au) sein (des) entreprises*) obtient un pourcentage non significatif en termes de fréquence des variantes fléchies ; ce pourcentage atteint 0,17% plus précisément. Les variantes récupérées sont associées notamment au dernier mot de la requête (*entreprises*). Une seule flexion, en nombre, est relevée, c'est le cas de la variante *entreprise*. Finalement, la requête 40 (*effets (du) chocolat (sur) (la) santé*) enregistre un pourcentage de variantes fléchies de 0,21% en termes de fréquence, ce qui est faible. Il s'agit de flexions en nombre qui sont associées notamment aux deux premiers mots de la requête. Ces variantes fléchies sont *effet* et *chocolats*.

Ainsi, l'examen des variantes fléchies des 5 requêtes les plus significatives et des 5 requêtes les moins significatives, par rapport au test statistique, montre que la fréquence des fléchis est faible dans la majorité des requêtes et varie d'une requête à l'autre et d'un mot à l'autre. En ce sens, seulement deux requêtes parmi les plus significatives (48 et 34) et une requête parmi les moins significatives obtiennent des scores de fréquence positifs. De même, les flexions varient en genre et en nombre selon les mots de chaque requête. Notons également qu'en termes de qualité aucune déviation associée aux flexions n'a été relevée.

En somme, les variations caractérisent aussi bien les requêtes ayant obtenu des résultats significatifs au test statistique que celles qui n'ont pas enregistré de tels résultats. Ces variations sont présentes indépendamment des facteurs liés à la fréquence des variables (termes et variantes), à la structure syntaxique des termes de requêtes, à la nature et à la qualité des variantes dérivées. Les exemples cités précédemment ont montré que la pertinence des requêtes (et/ou documents) n'est pas toujours en corrélation avec la fréquence d'une variable ou une autre. C'est-à-dire que nous pouvons avoir des requêtes ayant un bon score de pertinence, mais sans que la fréquence des termes, des dérivés ou des fléchis ne soit importante. De même, nous pouvons avoir des requêtes où la fréquence des variables (termes ou variantes) est significative, mais sans que le score de pertinence ne soit positif. En examinant d'autres paramètres d'ordre linguistique et leur possible impact sur la pertinence des documents, nous sommes arrivé à la même remarque, c'est-à-dire que la variation reste présente entre les différentes requêtes. Par exemple, deux requêtes ayant une même structure syntaxique peuvent avoir des résultats complètement différents au niveau de leur score de pertinence. De même, la nature et la qualité des dérivés sont des paramètres qui varient d'une requête à l'autre, voire d'un mot à l'autre. À ce sujet, si les variantes dérivées sont généralement proches d'un point de vue sémantique des termes de base, la présence des déviations sémantiques et leur fréquence dépend des requêtes, des mots et aussi de la démarche morphologique utilisée (formes tronquées) pour l'extraction des termes et des variantes.

## 4.3 Conclusion

Ce chapitre présente les résultats obtenus après l'analyse des différentes requêtes. Il s'agit de deux types de résultats : bruts et statistiques. L'objectif de cette analyse consiste à vérifier les questions de recherche posées dans la problématique ; il est question de savoir si la fréquence des variantes morphologiques, notamment dérivées, est en corrélation avec la pertinence des documents ; en d'autres termes, examiner l'impact des variantes morphologiques, en termes de fréquence, sur la pertinence des documents rapportés. Ces variantes, qui sont extraites sur la base d'une morphologie aléatoire, sont analysées dans une perspective de la reformulation de requêtes, et cela dans le but d'aider l'utilisateur à mieux exprimer ses besoins d'information par la reformulation des requêtes.

Alors, si les données brutes laissent croire que globalement la corrélation entre la fréquence des variables (notamment les termes de la requête et les dérivés) et la pertinence des documents est un fait, cela semble moins évident en examinant les résultats individuellement ainsi que ceux liés au test statistique. En ce sens, les tendances varient d'une requête à l'autre, d'un mot à l'autre et d'une variable à l'autre.

Outre la notion de fréquence, nous avons examiné d'autres paramètres en vue de vérifier s'il y a d'autres facteurs qui auraient un impact sur la pertinence des documents. À cet égard, nous avons traité des aspects linguistiques liés à la morphologie, syntaxe et sémantique. Le choix d'une morphologie basée sur une démarche aléatoire nous a permis de récupérer dans les documents rapportés eux-mêmes non seulement des variantes proches sémantiquement des termes de base, mais également des variantes complètement différentes des termes initiaux (déviations). À vrai dire, le choix de la morphologie aléatoire s'avère intéressant dans la mesure où il permet d'avoir une large couverture des mots de la langue, cependant cette démarche doit être soumise, en aval, à un processus de contrôle qui aura la tâche de limiter, voire d'empêcher, éventuellement, les déviations sémantiques. Notons que la fréquence de ces déviations (surgénération) varie d'une requête à l'autre et d'un mot à l'autre. De même, nous avons examiné la structure syntaxique des termes de la requête ; certes, la structure syntaxique est à prendre en compte, c'est-à-dire que plus une requête contient plusieurs termes, plus elle est descriptive et plus elle est susceptible de contenir des termes structurés, mais il s'est avéré que ce critère n'est pas toujours synonyme de pertinence.

# CONCLUSION

Un système de recherche d'information est un outil logiciel qui a pour tâche de retrouver et de rapporter des informations en réponse à la requête de l'utilisateur. Cette requête est exprimée, généralement, en langage naturel, notamment pour la recherche d'information de nature textuelle. Toutefois, la difficulté majeure à laquelle un SRI est confronté est de pouvoir satisfaire le besoin d'information (BI) de l'utilisateur. Cette difficulté s'accentue principalement dans le contexte de la recherche d'information (RI) sur le Web. Ce fut le sujet du premier chapitre de notre mémoire où il était question de discuter de la problématique liée à la RI sur le Web par rapport à la RI classique. À ce sujet, nous avons mis en exergue le fait que la RI classique, par son caractère homogène, ses sources structurées et normalisées, la base de données qui est fixe, etc., pose moins de problèmes que la RI sur le Web. Ce dernier est considéré comme un espace hétérogène, les informations sont non structurées, la base de données en évolution constante, la variété des formats et des langues, etc. Cette hétérogénéité rend l'accès à l'information répondant au BI de l'utilisateur plus difficile. Aussi, nous avons mis l'accent, dans le premier chapitre, sur les problèmes liés à la complexité des langues et leur ambiguïté. Parmi ces problèmes linguistiques les plus récurrents, nous avons relevé ceux liés à l'homographie (homonymie), synonymie, polysémie, etc. En plus de ces difficultés linguistiques, un SRI fait face à d'autres problèmes d'ordre cognitif ; ceux-ci concernent les interactions entre l'utilisateur et le système. L'examen des aspects cognitifs n'est pas pris en compte dans la présente étude.

Ainsi, pour faire face aux problèmes linguistiques, comme ceux évoqués ci-dessus, un SRI se doit d'intégrer des connaissances linguistiques dans le processus d'analyse, ce qui permettra une meilleure compréhension des contenus et une meilleure performance au niveau de l'appariement entre requêtes et documents. En effet, l'intégration des connaissance linguistiques par le biais des outils du traitement automatique des langues (TAL) est bien présent en RI et son impact sur l'amélioration des SRI est considérable. Notre étude a abordé le même axe de recherche, c'est-à-dire l'apport des connaissances linguistiques à la RI sur le Web ; toutefois, nous nous somme limité à étudier uniquement la question des variations morphologiques et leur apport à la RI. Les questions de recherche que nous avons posées portaient sur

l'impact de la fréquence des variantes morphologiques sur la pertinence des documents rapportés.

Le deuxième chapitre a porté sur le cadre théorique de notre travail de recherche. Après un bref historique lié à la RI notamment dans le contexte du Web, nous avons abordé deux notions fondamentales en RI, la notion de modèle et celle de pertinence. Ensuite, nous avons soulevé la question des connaissances linguistiques et leur apport à la RI. Un intérêt particulier a porté sur la variation morphologique, notre objectif de recherche. Ainsi, le cadre théorique nous a permis de mettre en lumière des méthodes et des techniques utilisées pour traiter des phénomènes linguistiques, de même que le progrès réalisé au niveau de l'intégration des connaissances linguistiques aux SRI (par le biais des techniques du TAL), notamment les connaissances morphologiques.

Le troisième chapitre, qui a porté sur la méthodologie, a étayé l'ensemble des outils utilisés dans l'expérimentation, il s'agit, entre autres, du choix du corpus, du choix du moteur de recherche Google, etc. Le choix du corpus s'est basé sur des critères liés à la langue et à l'autorité du corpus (TREC) dans le domaine de la RI. En effet, il s'agit d'un corpus en français et non pas un corpus traduit ; c'est un corpus extrait de TREC qui est une référence dans le domaine. Par ailleurs, le choix du moteur de recherche Google est lié à des considérations techniques et fonctionnelles, c'est-à-dire que ce moteur de recherche est considéré comme le plus généraliste et le plus performant parmi les moteurs existants. D'autres outils ont été utilisés comme c'est le cas du logiciel le Fréquencier et le logiciel d'analyse statistique SAS (Statistical Analysis System). De même, diverses étapes ont été suivies pour mettre en œuvre le processus expérimental ; l'analyse des documents rapportés, c'est-à-dire l'extraction des termes, des variantes morphologiques, etc., constitue une étape importante dans ce processus. Ainsi, pour extraire les termes et les variantes des documents rapportés, nous avons adopté une morphologie aléatoire ; en d'autres termes, une morphologie basée sur une procédure de troncation (possibilité de regrouper l'ensemble des variantes sous une forme (racine) commune. Une procédure qui a montré ses limites dans la mesure où elle engendre des déviations donnant lieu à des glissements sémantiques. Cela rend l'intégration de cette démarche morphologique à un processus de reformulation automatique de requêtes difficilement applicable.

Le quatrième chapitre a été consacré à la présentation des résultats de l'analyse. Ces résultats nous ont permis de répondre à nos questions de recherche et de vérifier leur validité. Au niveau de la première question portant sur l'impact (en termes de fréquence) des variations morphologiques, notamment les dérivés, sur la pertinence des documents, les résultats obtenus sont variables. À cet égard, nous avons distingué deux types de résultats, bruts et statistiques. Les résultats bruts, qui prennent en compte la variable *longueur du texte* dans le calcul des fréquences de chaque variable, montrent que globalement il y a une corrélation entre la fréquence des termes et des variantes et la pertinence des documents ; autrement dit, plus la fréquence des termes ou des variantes est élevée plus le document a de la chance d'être pertinent. Cependant, les résultats individuels ne confirment pas cette tendance, c'est-à-dire qu'il y a des variations selon les requêtes, selon les mots et selon les variables. Par ailleurs, les résultats d'ordre statistique ne sont pas concluants ; en ce sens, les valeurs obtenues varient d'une requête à l'autre, c'est-à-dire que nous n'avons pas obtenu une tendance claire confirmant une corrélation entre la fréquence des termes et des variantes et la pertinence des documents. Cette variation affecte également chacune des variables. Alors, si les résultats bruts révèlent que les *termes de la requête* enregistrent des scores supérieurs aux autres variables (variantes dérivées et fléchies), cette tendance a changé avec le test statistique où les variantes dérivées enregistrent des valeurs plus significatives que celles de la variable *termes de la requête* ; tandis que le score des variantes fléchies est resté bas. La variation caractérisant les résultats du test statistique serait liée au fait que ce test a pris en compte des paramètres différents de ceux utilisés dans les résultats bruts. Par exemple, pour calculer la fréquence des termes et des variantes, le test statistique n'intègre pas la variable *longueur du texte*, etc. En somme, il semble qu'il y a des liens entre la fréquence des variantes et la pertinence des documents sauf que cela varie et dépend des requêtes, des mots de la requête et des variantes.

De même, au sujet de la nature et de la qualité des dérivés, l'analyse d'un ensemble de requêtes a montré que ces deux paramètres se caractérisent par une grande variabilité, et cela indépendamment des résultats des requêtes en termes du test de pertinence et du test statistique. Alors, si la nature et la qualité des dérivés semblent avoir un impact sur la pertinence des documents rapportés, cela dépend notamment de la nature et de la qualité des mots de la requête de base ainsi qu'à leur structure syntaxique.

Par ailleurs, nous pouvons affirmer que la valeur scientifique de toute recherche dépend, en partie, de sa validité. Celle-ci constitue l'idéal du chercheur dans la mesure où elle lui permet d'avoir des réponses valables aux questions et hypothèses formulées (Robert, 1988). Néanmoins, la particularité liée à chaque recherche et les conditions dans lesquelles se déroulent les démarches empiriques rendent l'objectif d'avoir une recherche parfaite difficilement atteignable. C'est le cas de notre recherche qui comprend des points forts et des limites. L'un des points forts est lié au fait que notre analyse ne s'est pas limitée seulement aux données brutes, c'est-à-dire qu'un test statistique a été adopté pour l'analyse des données recueillies. Par ailleurs, les limites de notre recherche peuvent être liées, par exemple, au choix du corpus et sa représentativité et surtout à l'évaluation de la pertinence des documents rapportés. À cet égard, une part de subjectivité du chercheur n'est pas à écarter dans ce genre d'exercice ; l'idéal aurait été de confier le processus d'évaluation de la pertinence à un sujet externe.

Pour conclure, dans une perspective de la reformulation de requêtes, la prise en compte des variantes morphologiques, notamment dérivées, extraites des documents rapportés est une démarche qui peut poser problème au niveau des résultats rapportés, en termes de pertinence (précision) ; surtout si ce processus se fait automatiquement et sans contrôle. En effet, la morphologie aléatoire adoptée pour extraire les termes et les variantes génère des déviations (surgénération) qui peuvent être à l'origine d'importants glissements d'ordre sémantique et thématique. Ces glissements peuvent exister également entre les termes de base et les variantes qui sont supposés être proches sémantiquement. L'idéal serait l'adoption d'une morphologie contrôlée pour réduire le nombre de déviations ; ce contrôle doit aussi être appliqué pour empêcher, lors de la reformulation de requêtes, de générer des requêtes non structurées et non cohérentes. Le but consiste à éviter, par exemple, des combinaisons telles *criminologue féminité*, *crime féminisme*, etc. Dans cette optique, une voie de recherche consisterait à appliquer, en amont, une morphologie aléatoire pour pouvoir extraire le maximum de variantes morphologiques et à intégrer, en aval, une morphologie contrôlée pour empêcher, dans le processus de reformulation, la prise en compte des déviations sémantiques.

## ANNEXE A

### Descriptif et narratif des 50 requêtes TREC

| N° | Requête | Descriptif | Narratif |
|---|---|---|---|
| 1 | La culture écologique | Est-ce que l'agriculture écologique influe sur le commerce international ? | Un document pertinent examinera la demande de produits agricoles écologiques sur les marchés internationaux et l'existence de tarifs qui faciliteraient l'échange de ces produits. |
| 2 | Anti-sémitisme en Allemagne après 1945 | Quels sont les rapports, affaires, recherches empiriques et études disponibles concernant l'anti-sémitisme en Allemagne après la Deuxième Guerre Mondiale ? | Les documents concernés traitent de toutes les formes de l'anti-sémitisme en Allemagne (République Fédérale et République Démocratique Allemande) après 1945. Les actes et les mouvements d'anti-sémitisme ainsi que les aspects idéologiques de l'anti-sémitisme sont aussi pris en compte. Les études historiques (avant 1945) ne sont pas prises en compte. |
| 3 | Exportation de médicaments dangereux | On recherche des documents concernant l'exportation par l'industrie pharmaceutique suisse de médicaments de qualité suspecte dans le Tiers-Monde. | Selon une étude scientifique publiée il y a quelques années, près de la moitié des médicaments de l'industrie pharmaceutique suisse vendus au Tiers-Monde ne sont pas conformes aux exigences pharmacologiques. Les documents pertinents doivent contenir tous les renseignements concernant cette affaire (quels médicaments, quels dangers, quelles entreprises) et en général, l'exportation de médicaments de qualité suspecte dans le Tiers-Monde. |

| | | | |
|---|---|---|---|
| 4 | Criminalité féminine | Quels sont les rapports, affaires, recherches empiriques et études disponibles au sujet de la criminalité et de la délinquance des femmes ? | Les documents concernés abordent les problèmes particuliers de la criminalité des femmes, y compris les problèmes de la réinsertion sociale et de l'emprisonnement des femmes. Les études historiques (avant 1945), les statistiques générales, les réflexions sur la philosophie du droit et le terrorisme ne sont pas pris en compte. |
| 5 | Politique extérieure de l'Autriche | Quels sont les rapports et études disponibles concernant la politique extérieure de l'Autriche ? | Les documents pertinents contiennent les discussions et études sur la politique étrangère de l'Autriche. Les questions concernant la coopération européenne, l'adhésion à la Communauté européenne, les taxes de transit des véhicules et des poids lourds traversant l'Autriche offrent aussi un intérêt. Les études historiques (avant 1945) ne sont pas prises en compte. |
| 6 | Conditions de vie des immigrants | Quels sont les rapports et études disponibles concernant les conditions de vie des immigrants ? | Les documents pertinents contiennent les rapports et études concernant les conditions de vie et la situation sociale des immigrants. Il n'est pas tenu compte des études historiques (avant 1945) et des problèmes généraux des migrations, ni de la politique de migration et de la politique d'asile. |

| | | | |
|---|---|---|---|
| 7 | Production minimale au Japon | Quels sont les rapports, recherches empiriques et études disponibles concernant l'introduction et l'application du concept de production minimale au Japon ? | Les documents pertinents traitent des questions particulières de la production minimale au Japon. Ces documents peuvent inclure la production minimale présentée comme concept de gestion et offrant des exemples concrets d'applications et d'études au Japon. Les considérations générales sur la production minimale et les études spécifiques à d'autres pays ne sont pas prises en compte. |
| 8 | Extrémisme de droite et racisme | Quels sont les rapports et études traitant des liens entre l'extrémisme de droite et le racisme ? | Les documents pertinents traitent des liens et de l'interdépendance entre l'extrémisme de droite et le racisme. Les rapports et analyses sur l'idéologie et les rapports sur les actions concrètes sont aussi applicables. |
| 9 | Jeunesse et politique | Quels sont les rapports et études qui traitent des relations entre la jeunesse et la politique ? | Les documents pertinents traitent des relations entre la jeunesse et la politique et de la disponibilité des adolescents à s'engager en politique. Les études historiques (avant 1945) ne sont pas prises en compte. |
| 10 | Criminalité des adolescents | Quels sont les rapports et études traitant de la criminalité et de la délinquance des adolescents? | Les documents pertinents traitent des causes et des conséquences de la criminalité des adolescents. Les rapports sur la situation sociale et sur la réinsertion sociale sont aussi pris en compte. |
| 11 | Charte sociale européenne | Quels sont les rapports qui traitent de la Charte sociale européenne ? | Les documents pertinents traitent de la Charte sociale européenne et de son application. |

| | | | |
|---|---|---|---|
| 12 | Violence des adolescents | Quels sont les rapports, recherches empiriques et études disponibles sur la violence des adolescents ? | Les documents pertinents traitent de la violence et des actes de violence des adolescents. Les rapports sur les adolescents victimes de violences, sur la violence et la sexualité, et sur les problèmes généraux de la violence ne sont pas pris en compte. |
| 13 | Législation sur la protection de la nature | Quels sont les rapports disponibles concernant la législation sur la protection de la nature ? | Les documents pertinents traitent de la législation sur la protection de la nature. Les documents sur la protection de la nature en général ne sont pas pris en compte. |
| 14 | La politique économique de la Slovénie | On recherche des documents concernant le développement et les changements de la politique économique de la Slovénie. | Les documents pertinents concernent la politique économique slovène dans le cadre de la transformation de la société, y compris la privatisation d'industries. Seront rejetées les déclarations d'ordre général sur la situation économique dans lesquelles la politique économique n'est pas mentionnée explicitement. |
| 15 | Déportation des étrangers en Autriche | Quels sont les rapports disponibles sur la déportation des étrangers en Autriche ? | Les documents pertinents traitent de la déportation des étrangers en Autriche. |
| 16 | Les communistes au Parlement Européen | La proportion et le nombre de communistes au Parlement Européen sont-ils en progrès ? | Les documents pertinents doivent indiquer quelle est la proportion de communistes élus au Parlement Européen. Cette information peut être plus ou moins précise (des pourcentages exacts aux simples comparaisons comme "plus qu'avant" ou "moins que les partis écologistes"). Les documents ne contenant que des résultats de sondages ne seront pas retenus. |

| | | | |
|---|---|---|---|
| 17 | Destruction de la forêt tropicale en Amérique du Sud | Quels sont les rapports et études qui traitent des raisons et des conséquences de la destruction de la forêt tropicale en Amérique du Sud ? | Les documents pertinents expliquent les raisons et les conséquences de la destruction de la forêt tropicale en Amérique du Sud. Les rapports sur les mesures pour la reconstitution et la préservation de la forêt tropicale sont aussi pertinents. Les rapports sur la forêt tropicale dans d'autres parties du monde ne sont pas pertinents. |
| 18 | Famine au Soudan | Quels sont les rapports et études disponibles concernant la famine au Soudan ? | Les documents à rechercher décrivent la famine au Soudan. Les documents qui concernent de manière générale la guerre civile au Soudan, les conséquences des inondations ou l'invasion des sauterelles ne sont pas pris en compte. |
| 19 | Protection des animaux | On recherche des documents concernant des associations et individus du monde entier qui s'engagent pour sauver des espèces protégées. | Les animaux qui appartiennent à des espèces protégées sont souvent mis en danger ou blessés à cause des hommes. Les documents pertinents parlent d'animaux d'espèces protégées sauvés par des associations ou par des personnes impliquées dans la sauvegarde des animaux et de leurs droits. |
| 20 | Limitations des importations de l'UE | Quels sont les rapports et études qui traitent des limitations de l'importation et des barrières tarifaires de l'Union Européenne (UE) ? | Les documents disponibles traitent des limitations de l'importation, des barrières tarifaires et des droits de douane. Les rapports sur les taxes d'importation et les taxes de pénalisation peuvent aussi être consultés. Les documents qui présentent de manière générale les obligations ou le GATT ne sont pas pris en compte. |

| | | | |
|---|---|---|---|
| 21 | Unité Franco-Allemande | Trouver les documents qui traitent des efforts conjoints pour l'unité entre la France et l'Allemagne de l'Ouest | Les documents pertinents montrent comment la coopération et les initiatives communes renforcent l'union Franco-Allemande. Les documents pertinents présentent par exemple les faits suivants : la création d'un conseil commun de défense, la création d'un partenariat économique pour favoriser la compétitivité sur le marché mondial, les ambassades communes, un corps militaire commun pour améliorer la défense, et la création d'une société mixte de production d'hélicoptères. |
| 22 | Immigration et racisme | Quelles sont les mesures prises par L'Union Européenne ou les gouvernements de chaque pays de l'Union en faveur de l'intégration des immigrants ? | L'augmentation de l'immigration en provenance des pays du tiers-monde vers les pays développés est la cause d'un ressentiment croissant dans certains groupes sociaux. Les documents pertinents présentent les propositions, les mesures légales et les lois qui ont pour but l'intégration des immigrants et d'éviter les actes de racisme et d'intolérance, en particulier à l'intérieur de l'Union Européenne vis-à-vis des immigrants de nationalité non européenne. Les documents traitant d'actes individuels d'intolérance et/ou de marches de protestation sont non pertinents. |
| 23 | les accidents de la route | Quelles sont les principales causes des accidents de la route ? | Les causes des accidents de la route varient mais sont principalement la vitesse, l'alcool, la drogue et les mauvaises conditions météorologiques, en particulier le brouillard. Les documents pertinents doivent explicitement mentionner les raisons de l'accident. |

| | | | |
|---|---|---|---|
| 24 | L'éducation sexuelle | Les articles qui discutent l'introduction de l'éducation sexuelle dans les écoles dans un effort de combattre l'escalade de sida. | L'éducation et l'information semblent être la clef afin de contrôler l'escalade de sida. Un document pertinent contiendra des informations sur les efforts de l'introduction de l'éducation sexuelle dans le programme afin de combattre la diffusion de sida dans une envergure toujours croissante. |
| 25 | Les effets de la déforestation | Quels sont les effets de la déforestation sur la désertification ? | Tous les documents qui donnent des analyses spécifiques sur les mesures des gouvernements locaux ou des agences internationales pour freiner la déforestation sont pertinents. Les articles qui contiennent des renseignements sur la désertification et ses effets secondaires comme les changements de climat, l'épuisement de la terre, les inondations et les ouragans sont également applicables. |
| 26 | Terrorisme international | Quelles sont les mesures prises pour combattre le terrorisme international ? | Les documents utiles citent les pays qui fomentent le terrorisme ou qui informent sur les mesures prises par les gouvernements ou organisations internationales pour le combattre. |
| 27 | Maltraitance des enfants | Quelles sont les mesures pour combattre le phénomène croissant de la maltraitance des enfants dans le monde ? | Les documents utiles donnent des statistiques sur la maltraitance des enfants dans le monde et informent sur les mesures pour y mettre fin. Les mentions de cas particuliers d'abus sont sans rapport. |
| 28 | Exploitation économique du fond marin | On recherche des documents concernant l'exploitation économique du fond marin et de la plateforme continentale. | L'exploitation économique du fond marin comprend également l'exploitation de la plateforme continentale, en particulier les ressources minérales telles que le pétrole, le gaz naturel et les métaux (manganèse, etc.)que l'on peut trouver dans la mer. On retiendra également les documents concernant l'utilisation de plateformes d'extraction. |

| | | | |
|---|---|---|---|
| 29 | Exportation d'armes à la Turquie | Quelles nations européennes fournissent encore des armes à la Turquie ? | L'Europe a dénoncé la répression cruelle des Kurdes par l'armée turque. Les documents pertinents concernent les nations européennes qui n'ont pas encore suspendu la livraison en armement à la Turquie, les protestations contre les gouvernements de ces pays ou les discussions internes concernant cette question. |
| 30 | Les débris spatiaux | Quels dangers représentent les débris spatiaux produits par l'homme ? | On s'intéressera avant tout aux débris spatiaux produits par l'homme et des dangers provoqués. Les documents qui traitent d'objets naturels comme les comètes ne seront pas retenus. |
| 31 | Séparatistes catalans et galiciens | A part l'ETA, quels mouvements séparatistes sont actifs en Espagne? | Le pays basque n'est pas la seule province espagnole où des groupes séparatistes s'opposent violemment à l'autorité nationale. La Catalogne et la Galice ont également leurs mouvements indépendantistes. Les documents pertinents doivent relater les activités et revendications de groupes séparatistes dans ces deux provinces. |
| 32 | Coopération internationale des entreprises | Quels sont les rapports et études qui traitent de la coopération internationale des entreprises ? | Les documents pertinents traitent de la coopération internationale des entreprises interdépendantes, des usines et des syndicats. Les documents concernant les compagnies multinationales et la gestion du personnel international ne sont pas pris en compte. |
| 33 | Réfugiés de Bosnie | Quels sont les rapports disponibles sur les réfugiés et les personnes cherchant asile depuis la Bosnie et l'Herzégovine ? | Les documents applicables traitent des problèmes spécifiques des réfugiés et des personnes cherchant asile depuis la Bosnie et l'Herzégovine. Ils comprennent les problèmes des demandes d'asile et de permis de séjour posés par ces personnes. Les problèmes généraux des réfugiés et de l'asile, de même que les informations sur la guerre civile en Bosnie-Herzégovine ne sont pas pris en compte. |

| | | | |
|---|---|---|---|
| 34 | Accidents dans l'industrie minière | Quels sont les rapports disponibles sur les accidents et catastrophes survenus dans l'industrie minière ? | Les documents applicables traitent des accidents et catastrophes dans l'industrie minière. Ils comprennent les problèmes de la sécurité industrielle dans l'industrie minière et les catastrophes survenues dans les mines. |
| 35 | Transport public local | Quels sont les rapports disponibles concernant le transport public local ? | Les documents pertinents traitent du transport public local, en particulier des concepts de trafic local. |
| 36 | La pollution causée par l'automobile | Des documents analysant la pollution de l'atmosphère provoquée par l'automobile. | Quels sont les dangers causés par les émissions de l'automobile qui mènent à la dégradation de l'environnement? Quels sont les pays les plus affectés par les effects néfastes de la pollution sur la santé comme, par exemple, les maladies respiratoires pour lesquelles les agences de la protection de l'environnement identifient la pollution comme étant directement responsable ? |
| 37 | Vitesse sur les autoroutes en Suisse | Les articles qui parlent du rejet de l'initiative en faveur de l'augmentation de vitesse sur les autoroutes en Suisse. | Un article pertinent donnera des opinions sur la modification de la limite de la vitesse qui etait en vigueur pendant une période de trois ans. |
| 38 | L'homosexualité et la loi | Quels sont les droits légaux des individus ou couples homosexuels ? | Les documents pertinents doivent préciser les lois concernant les droits des personnes homosexuelles, y compris le droit au mariage ou à l'adoption d'enfants. |
| 39 | Processus de paix au Moyen-Orient | Quelle est l'attitude des pays arabes à l'égard du processus de paix au Moyen-Orient ? | Les documents pertinents fournissent des renseignements sur l'attitude des différents pays arabes à l'égard des initiatives prises par les Etats-Unis ou sur les efforts d'unifier leur position vis-à-vis du processus de paix Moyen-Orient. |

| | | | |
|---|---|---|---|
| 40 | Effets du chocolat sur la santé | Le chocolat a-t-il un effet quelconque sur la santé ? | Les documents valables donnent des informations sur les recherches médicales et leur résultat à propos des effets du chocolat sur la santé. |
| 41 | Ouvriers étrangers en Europe | La demande de main d'oeuvres étrangers en Allemagne et en Suisse a-t-elle changée ou est-ce qu'elle est restée pareille après la chute du mur de Berlin ? | Un document pertinent contiendra des informations sur les effects sur la demande d'emploi, les conditions de travail et les attitudes envers les ouvriers étrangers après la chute du mur de Berlin qui a amené un flot de main d'oeuvre bon marché de l'Europe de l'est. |
| 42 | Conséquences de la réunification allemande | Quelles sont les conséquences de la réunification allemande ? | Les documents pertinents expliquent les conséquences principales de la réunification des deux Allemagnes. |
| 43 | Conversion de la dette pour la Pologne | Quels sont les rapports et études disponibles concernant la conversion de la dette de la Pologne ? | Les documents à rechercher présentent les négociations et leurs résultats concernant la conversion de la dette Polonaise. Les rapports sur l'endettement de la Pologne et les actions visant à la conversion de la dette dans le contexte des accords bilatéraux et internationaux sont aussi pris en compte. |
| 44 | Statut militaire de l'Allemagne unifiée | Quelles sont les difficultés soulevées par le statut militaire de l'Allemagne unifiée ? | Le statut militaire de l'Allemagne unifiée est le principal obstacle à un consensus sur la réunification. Pour les Etats-Unis, l'Allemagne devrait être membre de l'OTAN. L'URSS souhaite qu'elle soit à la fois membre de l'Alliance atlantique et du Pacte de Varsovie. Quant à l'Allemagne de l'est, elle est pour une Allemagne unie et militairement neutre, thèse jugée inacceptable par les occidentaux. Les documents concernés décrivent ce que les différents partenaires politiques prévoient pour régler le nouveau statut militaire de l'Allemagne unifiée et montrent les nouvelles orientations politiques des alliances traditionnelles. |

| | | | |
|---|---|---|---|
| 45 | Lutte contre la corruption | Quels sont les rapports et études qui traitent de la lutte contre la corruption ? | Les documents concernés décrivent les causes et les conséquences de la corruption et de la lutte contre la corruption. |
| 46 | Protection de l'environnement au sein des entreprises | Quels sont les rapports existants concernant la protection de l'environnement à l'intérieur des entreprises ? | Les documents applicables traitent de la protection de l'environnement à l'intérieur des entreprises. |
| 47 | Maintien de la paix par l'OUA | On recherche des documents concernant la politique de maintien de la paix de l'Organisation de l'Unité Africaine (OUA). | Les documents pertinents traitent de la politique de maintien de la paix et de règlement des conflits de l'Organisation de l'Unité Africaine (OUA). Ils traitent également des interventions militaires des troupes de l'OUA dans les régions en conflit ou dans les guerres entre états membres. |
| 48 | L'industrie européenne du film | On recherche des documents concernant la situation économique et artistique de l'industrie européenne du film. | Les documents pertinents traitent de la situation économique et des performances artistiques de l'industrie européenne du film. Les distributeurs de films seront également pris en considération. |
| 49 | Normes de protection professionnelle | On recherche des documents concernant les normes de protection professionnelle de l'OIT (Organisation Internationale du Travail) et leur mise en oeuvre. | Les documents pertinents traitent des Normes de protection professionnelle de l'OIT (Organisation Internationale du Travail) ainsi que de leur application dans les états membres. Le non-respect de ces normes est également à prendre en considération. |
| 50 | Traitement des déchets nucléaires | Quelles procédures sont actuellement utilisées ou envisagées pour un traitement des déchets nucléaires non nuisible à l'environnement ? | Un document pertinent doit exposer une méthode ou une procédure pour un traitement efficace des déchets nucléaires ou radioactifs et que l'on juge sûre pour l'environnement. |

## ANNEXE B

### Exemple d'évaluation de la pertinence multivaluée

Degré 3 de la pertinence : implique que le document évoque d'une façon approfondie le BI ou couvre plusieurs facettes d'un BI.

Degré 2 de la pertinence : correspond à un document partiellement pertinent, c'est un document qui évoque seulement quelques facettes du BI.

Degré 1 de la pertinence : correspond à un document où le BI est seulement mentionné.

Degré 0 de la pertinence : représente un document non pertinent où le BI n'est pas mentionné.

```
<F-title> Criminalité féminine

<F-desc> Description :
Quels sont les rapports, affaires, recherches empiriques et
études disponibles au sujet de la criminalité et de la
délinquence des femmes ?

<F-narr> Narrative :
Les documents concernés abordent les problèmes particuliers de
la criminalité des femmes, y compris les problèmes de la
réinsertion sociale et de l'emprisonnement des femmes. Les
études historiques (avant 1945), les statistiques générales,
les réflexions sur la philosophie du droit et le terrorisme ne
sont pas pris en compte.
```

**Pertinence 3**
**Extrait du document rapporté (doc 3):**

[...] Dans son volume, Marie-Andrée Bertrand cherche à expliquer la hausse de la criminalité chez les femmes en abordant diverses hypothèses : accroissement réel des délits causé par leurs nouvelles conditions de vie, visibilité plus grande de leurs comportements délictueux ou encore changement d'attitude à leur égard de la part de l'appareil judiciaire.

Cette dernière hypothèse est la plus controversée puisque les données et les témoignages semblent se contredire. Une étude citée par la chercheuse tend à indiquer que le système judiciaire est plus clément envers les femmes. Par exemple, dans les cas où l'accusation ne concerne qu'une seule infraction, 20 % des hommes accusés de

voies de fait se sont vu imposer une peine d'emprisonnement, contre 8 % des femmes ; dans les cas de conduite avec facultés affaiblies, 33 % des hommes ont reçu une sentence d'emprisonnement contre 6 % des femmes. Non seulement les femmes sont moins souvent condamnées à la prison, mais leurs peines sont généralement plus courtes.

L'explication de leur présence grandissante dans les pénitenciers serait donc dans l'accroissement des «crimes contre la vie humaine», pour lesquels il n'y a pas de peine non carcérale.

Marie-Andrée Bertrand émet également l'hypothèse que le milieu judiciaire pourrait avoir changé d'attitude envers les femmes en considérant leur autonomie croissante et leur réussite professionnelle. Des témoignages entendus de la part de policiers vont en ce sens. «Des policiers affirment que l'autonomie des femmes les amène à les considérer comme des êtres responsables au même titre que les hommes», indique la criminologue. Ce qui voudrait dire qu'ils faisaient auparavant preuve de plus de mansuétude à leur égard [...].

**Pertinence 2**

**Extrait du document rapporté (doc 19) :**

[...] De mon expérience, quand un homme fait violence à une femme on en fait une cause d'État. Quand une mère et une grand-mère payent des tueurs à gages pour empêcher tout contact entre le père et son enfant ça passe après la page aux aubaines. Les histoires d'horreurs du côté sombre des femmes, j'en ai vu, mais c'est un sujet tabou qu'on ne veut voir. On préfère trouver ou inventer des excuses pour ces dames. Parlez-en à Latimer.

Quand une mère tue son enfant pour ne pas que le père le voit et quand elle se sauve sans laisser d'adresse, même réaction. Les seules subventions récurrentes qu'ont des OSBL pour hommes sont celles qui traitent de la violence, comme si le seul problème masculin était une violence innée qu'il faut soigner à tout prix. À comparer des millions pour les organismes pour femmes c'est de la petite bière. D'ailleurs le discours de ces organismes est assez paradoxal. En effet, quand ces OSBL féminins revendiquent plus de subventions ils déclarent que le gouvernement a amplement les moyens. Quand des OSBL d'hommes revendiquent des subventions bien à eux, pas celles des organismes pour femmes, des groupes de femmes crient à l'injustice, on ne doit en aucun cas diviser les maigres ressources de l'État. Cela affaiblirait «la cause» des femmes et ferait reculer leur acquis. Surtout, il y aurait quelques unes qui perdraient leur emploi... [...]

[...] L'ouvrage de la criminologue Marie-Andrée Bertrand dresse un portrait de la criminalité féminine. Elle lève le voile sur le tabou, j'imagine que plusieurs l'étiquetteront d'anti féministe, de masculiniste, de membre de la droite, mais c'est le risque à courir pour dévoiler à quelqu'un son " défaut ".

On excuse la violence des femmes. On se penche plus facilement sur ce qui peut excuser ou expliquer le geste dans leur cas, on parle de pauvreté, de père absent (alors que le système les exclu), etc. Ou on excuse la violence

féminine par le fait que ce doit être absolument en réplique d'un conjoint agresseur et que de toute façon, la force et les dommages sont moins grands lorsque cette violence est féminine. On ne réclame ni égalité ni parité de traitement dans ce secteur. Pourtant le traitement devrait être le même. [...]

**Pertinence 1**

**Extrait du document rapporté (doc 5) :**

[...] *Les femmes résistent au crime*
Robert Cario
Ed. L'Harmattan, 1999

Les femmes bénéficient d'une socialisation orientée vers l'altruisme, la douceur et le sacrifice ; celles qui deviennent criminelles présentent, en général, des défaillances psychoculturelles et sociales profondes...

Crimes de femmes
*"25 histoires vraies"*
*Anne-Sophie Martin, Brigitte Vital-Durand*
*Ed. Flammarion, 2004*

Grandes figures de la criminalité féminine de ces deux dernières décennies, leur renommée est comparable à celle des noires héroïnes du passé : Marie Besnard, Pauline Dubuisson, les s¿urs Papin et autres "diaboliques" ou "anges de la mort"... Qui sont-elles ? Qui sont leurs victimes ? Pourquoi tuent-elles ? A lire comme un roman... [...]

**Pertinence 0**

**Extrait du document rapporté (doc 6) :**

[...] La femme de Rafik Khalifa et deux de ses proches interpellés à Paris

PARIS (AP) - Nadia Amirouchane, l'épouse de Rafik Khalifa, et deux proches de l'ancien homme d'affaires algérien ont été interpellés mardi en milieu de matinée à Paris a-t-on appris de source policière.

Ces trois personnes étaient recherchées par la justice algérienne après leur condamnation pour "association de malfaiteurs et escroqueries".

Domiciliés dans le centre de la capitale, Nadia Amirouchane, 34 ans, Ghazi Kebbache, l'ancien gérant de la banque Khalifa et Mohammed Nanouche, son ancien directeur général, âgés tous deux de 58 ans, ont été arrêtés dans la rue dans les 7e, 5e et 15e arrondissements par les policiers de l'Office central de lutte contre le crime organisé. [...]

[...] A cause de son mariage forcé et de sa grossesse : Le Cem exclut une de ses brillantes élèves

Au Cem de Barkédji, les mariages précoces et forcés inquiètent beaucoup l'administration de l'établissement. Et selon le principal M N, si la loi était vigoureusement appliquée, le Cem de Barkédji pourrait être fermé, car la plupart des filles sont mariées à bas âge. Même les garçons ne sont pas en reste. Certaines filles sont même épousées avant de fréquenter le collège.

En fait, l'élève F B, de la classe de 5è, ayant une moyenne de 13,50 avec tableau d'honneur, a été exclue, pour avoir été épousée et engrossée durant l'année scolaire. La brillante F venant du village de Fouthity, à une trentaine de kilomètres de Barkédji, pour poursuivre ses études, est, malheureusement, rentrée dans son village natal avec la tête baissée. [...]

# ANNEXE C

## Résultats des corrélations entre pertinence et variables

| | Pertinence | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | | | 1 | | | 2 | | | 3 | | |
| | %TERME | | | %TERME | | | %TERME | | | %TERME | | |
| | N | MOY | É.T. | N | MOY | É.T. | N | MOY | É.T. | N | MOY | É.T. |
| Requête | | | | | | | | | | | | |
| 1 | 3 | 28.063 | 45.003 | 12 | 1.592 | 1.407 | 4 | 0.665 | 0.220 | 1 | 1.155 | . |
| 2 | 6 | 0.455 | 0.337 | 3 | 0.612 | 0.379 | 8 | 0.634 | 0.623 | 3 | 0.397 | 0.192 |
| 3 | 6 | 0.723 | 0.733 | 6 | 1.312 | 0.872 | 4 | 1.747 | 1.171 | 4 | 2.687 | 1.182 |
| 4 | 4 | 1.213 | 1.282 | 9 | 0.932 | 0.832 | 2 | 0.382 | 0.137 | 5 | 1.579 | 1.441 |
| 5 | 11 | 1.694 | 1.093 | 4 | 1.957 | 1.109 | 1 | 2.322 | . | 4 | 2.337 | 0.298 |
| 6 | 2 | 0.386 | 0.157 | 3 | 1.244 | 0.767 | 8 | 1.569 | 1.022 | 7 | 0.652 | 0.459 |
| 7 | 16 | 0.850 | 0.644 | 2 | 1.413 | 0.213 | 1 | 1.694 | . | 1 | 1.974 | . |
| 8 | . | . | . | 3 | 1.915 | 0.885 | 8 | 1.458 | 0.814 | 9 | 2.407 | 1.183 |
| 9 | 3 | 4.415 | 4.479 | 6 | 4.332 | 1.971 | 5 | 4.546 | 1.737 | 6 | 2.316 | 1.698 |
| 10 | 9 | 0.684 | 0.542 | 2 | 1.855 | 1.833 | 4 | 1.487 | 1.308 | 5 | 1.425 | 0.559 |
| 11 | 2 | 4.389 | 2.321 | 6 | 5.424 | 2.519 | 5 | 4.362 | 2.488 | 7 | 2.944 | 1.346 |
| 12 | 7 | 2.358 | 1.892 | 5 | 2.387 | 1.713 | 4 | 1.789 | 0.757 | 4 | 2.201 | 0.884 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 13 | 8 | 1.333 | 1.058 | 5 | 4.873 | 2.687 | 4 | 3.013 | 1.945 | 3 | 1.309 | 0.540 |
| 14 | 8 | 1.923 | 0.689 | 5 | 2.266 | 1.159 | 3 | 1.778 | 0.410 | 4 | 2.057 | 0.525 |
| 15 | 14 | 0.474 | 0.252 | 5 | 0.650 | 0.393 | . | . | . | 1 | 1.017 | . |
| 16 | 13 | 1.357 | 0.715 | 3 | 1.597 | 0.635 | 4 | 1.737 | 0.064 | . | . | . |
| 17 | 3 | 1.701 | 1.416 | 4 | 1.381 | 0.243 | 12 | 1.692 | 0.881 | 1 | 3.240 | . |
| 18 | 3 | 4.744 | 5.517 | 8 | 2.718 | 1.377 | 6 | 2.609 | 1.715 | 3 | 1.494 | 0.403 |
| 19 | 3 | 1.376 | 0.769 | 10 | 4.452 | 6.669 | 4 | 2.229 | 1.269 | 3 | 1.891 | 1.280 |
| 20 | 3 | 0.863 | 1.357 | 2 | 1.357 | 0.484 | 6 | 1.586 | 0.909 | 9 | 1.164 | 0.515 |
| 21 | 7 | 0.616 | 0.943 | 4 | 1.071 | 1.067 | 7 | 1.089 | 0.533 | 2 | 1.411 | 0.223 |
| 22 | 2 | 0.920 | 0.338 | 9 | 1.451 | 0.483 | 6 | 1.372 | 0.523 | 3 | 0.808 | 0.577 |
| 23 | 3 | 1.798 | 1.696 | 9 | 1.623 | 0.764 | 5 | 1.332 | 0.469 | 3 | 1.897 | 0.167 |
| 24 | 1 | 0.000 | . | 3 | 2.956 | 4.190 | 7 | 2.167 | 1.532 | 9 | 2.055 | 1.125 |
| 25 | 2 | 1.426 | 0.034 | 4 | 1.417 | 1.950 | 7 | 0.935 | 0.615 | 7 | 1.247 | 0.390 |
| 26 | 2 | 0.751 | 0.551 | 4 | 3.514 | 2.077 | 5 | 0.894 | 0.495 | 9 | 1.811 | 0.686 |
| 27 | 1 | 2.727 | . | 7 | 4.205 | 3.285 | 4 | 1.372 | 0.674 | 8 | 2.021 | 0.888 |
| 28 | 2 | 1.024 | 0.024 | 4 | 0.866 | 0.543 | 8 | 1.240 | 0.777 | 6 | 0.978 | 0.384 |
| 29 | 7 | 1.715 | 0.944 | 6 | 2.706 | 1.167 | 7 | 2.354 | 1.056 | . | . | . |
| 30 | 4 | 1.312 | 1.844 | 3 | 5.016 | 4.322 | 7 | 2.028 | 0.711 | 6 | 2.027 | 0.670 |
| 31 | 11 | 0.388 | 0.699 | 5 | 0.115 | 0.050 | 4 | 0.176 | 0.155 | . | . | . |
| 32 | 6 | 2.535 | 2.165 | 5 | 1.375 | 1.048 | 6 | 1.426 | 0.582 | 3 | 2.424 | 1.382 |
| 33 | 4 | 1.178 | 1.338 | 6 | 3.932 | 3.016 | 5 | 2.078 | 0.813 | 5 | 1.811 | 0.477 |
| 34 | 7 | 1.312 | 0.945 | 2 | 1.657 | 0.635 | 6 | 2.261 | 1.007 | 5 | 2.159 | 1.228 |
| 35 | 7 | 2.925 | 1.650 | 2 | 1.162 | 0.021 | 7 | 3.441 | 1.551 | 4 | 3.709 | 2.570 |
| 36 | 3 | 1.148 | 1.219 | 8 | 1.168 | 1.071 | 4 | 1.723 | 0.778 | 5 | 1.956 | 0.671 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 37 | 11 | 1.658 | 0.588 | 8 | 1.797 | 1.559 | . | . | . | 1 | 0.569 | . |
| 38 | 2 | 1.075 | 0.474 | 5 | 1.639 | 1.518 | 4 | 1.074 | 0.233 | 9 | 1.090 | 0.293 |
| 39 | 2 | 0.821 | 0.246 | 10 | 2.890 | 1.530 | 6 | 3.268 | 1.534 | 2 | 2.891 | 2.427 |
| 40 | 1 | 0.394 | . | . | . | . | 10 | 3.607 | 1.365 | 9 | 3.436 | 1.071 |
| 41 | 13 | 1.222 | 1.097 | 6 | 1.408 | 0.448 | 1 | 0.800 | . | . | . | . |
| 42 | 1 | 0.702 | . | 12 | 1.465 | 0.956 | 5 | 0.901 | 0.164 | 2 | 0.712 | 0.050 |
| 43 | 10 | 0.957 | 0.472 | 6 | 1.651 | 1.022 | 4 | 3.238 | 2.318 | . | . | . |
| 44 | 8 | 1.344 | 0.373 | 6 | 0.956 | 0.329 | 3 | 1.107 | 0.244 | 3 | 0.937 | 0.483 |
| 45 | 1 | 1.149 | . | 6 | 2.061 | 0.562 | 7 | 3.306 | 1.855 | 6 | 4.901 | 2.215 |
| 46 | 2 | 1.564 | 1.093 | 3 | 2.073 | 0.455 | 8 | 1.792 | 0.731 | 7 | 3.041 | 0.533 |
| 47 | 1 | 2.365 | . | 3 | 2.517 | 0.859 | 9 | 3.072 | 1.322 | 7 | 2.485 | 0.448 |
| 48 | 5 | 1.643 | 0.546 | 2 | 1.486 | 0.937 | 7 | 2.049 | 1.192 | 6 | 2.022 | 0.734 |
| 49 | 11 | 1.536 | 1.096 | 8 | 0.816 | 0.497 | 1 | 0.789 | . | . | . | . |
| 50 | 3 | 1.039 | 0.946 | 6 | 2.321 | 2.648 | 7 | 3.987 | 1.253 | 4 | 3.013 | 1.169 |

| | Pertinence | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | | | 1 | | | 2 | | | 3 | | |
| | %DERIVATION | | | %DERIVATION | | | %DERIVATION | | | %DERIVATION | | |
| | N | MOY | É.T. | N | MOY | É.T. | N | MOY | É.T. | N | MOY | É.T. |
| Requête | | | | | | | | | | | | |
| 1 | 3 | 0.103 | 0.178 | 12 | 0.470 | 0.375 | 4 | 0.707 | 0.258 | 1 | 1.403 | . |
| 2 | 6 | 0.848 | 0.759 | 3 | 0.570 | 0.522 | 8 | 0.830 | 0.461 | 3 | 0.456 | 0.256 |
| 3 | 6 | 1.372 | 1.209 | 6 | 1.032 | 0.912 | 4 | 0.860 | 0.739 | 4 | 0.507 | 0.185 |
| 4 | 4 | 1.845 | 1.144 | 9 | 2.249 | 1.538 | 2 | 1.415 | 0.038 | 5 | 2.989 | 2.200 |
| 5 | 11 | 0.249 | 0.220 | 4 | 0.405 | 0.306 | 1 | 0.332 | . | 4 | 0.728 | 0.082 |
| 6 | 2 | 0.090 | 0.127 | 3 | 1.310 | 0.835 | 8 | 1.043 | 0.770 | 7 | 1.715 | 1.392 |
| 7 | 16 | 0.643 | 0.773 | 2 | 0.728 | 0.277 | 1 | 0.496 | . | 1 | 1.666 | . |
| 8 | . | . | . | 3 | 0.439 | 0.761 | 8 | 0.997 | 0.738 | 9 | 0.673 | 0.424 |
| 9 | 3 | 0.030 | 0.051 | 6 | 1.243 | 1.014 | 5 | 1.889 | 1.476 | 6 | 1.514 | 0.725 |
| 10 | 9 | 0.277 | 0.469 | 2 | 0.435 | 0.088 | 4 | 0.524 | 0.443 | 5 | 0.706 | 0.651 |
| 11 | 2 | 0.000 | 0.000 | 6 | 1.323 | 1.415 | 5 | 1.142 | 0.661 | 7 | 0.507 | 0.280 |
| 12 | 7 | 0.269 | 0.293 | 5 | 0.539 | 0.888 | 4 | 0.756 | 0.353 | 4 | 0.543 | 0.240 |
| 13 | 8 | 0.291 | 0.163 | 5 | 0.270 | 0.447 | 4 | 0.559 | 0.400 | 3 | 1.184 | 0.958 |
| 14 | 8 | 0.709 | 1.037 | 5 | 0.365 | 0.469 | 3 | 0.495 | 0.468 | 4 | 0.321 | 0.254 |
| 15 | 14 | 0.263 | 0.348 | 5 | 0.674 | 0.400 | . | . | . | 1 | 0.949 | . |
| 16 | 13 | 0.436 | 0.555 | 3 | 0.895 | 0.518 | 4 | 0.640 | 0.749 | . | . | . |
| 17 | 3 | 0.229 | 0.110 | 4 | 0.553 | 0.296 | 12 | 0.835 | 0.653 | 1 | 0.935 | . |
| 18 | 3 | 0.051 | 0.089 | 8 | 0.530 | 0.500 | 6 | 0.253 | 0.299 | 3 | 0.385 | 0.113 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 19 | 3 | 0.071 | 0.123 | 10 | 0.410 | 0.765 | 4 | 0.304 | 0.412 | 3 | 0.237 | 0.212 |
| 20 | 3 | 0.230 | 0.200 | 2 | 1.074 | 0.083 | 6 | 1.476 | 0.695 | 9 | 1.309 | 0.293 |
| 21 | 7 | 0.217 | 0.267 | 4 | 0.196 | 0.228 | 7 | 0.249 | 0.234 | 2 | 0.400 | 0.418 |
| 22 | 2 | 0.227 | 0.321 | 9 | 0.423 | 0.386 | 6 | 0.674 | 0.377 | 3 | 1.417 | 0.748 |
| 23 | 3 | 0.051 | 0.089 | 9 | 0.985 | 1.068 | 5 | 0.301 | 0.267 | 3 | 0.513 | 0.376 |
| 24 | 1 | 1.653 | . | 3 | 0.611 | 0.540 | 7 | 0.941 | 0.406 | 9 | 1.398 | 0.768 |
| 25 | 2 | 0.830 | 0.148 | 4 | 0.411 | 0.285 | 7 | 0.770 | 0.483 | 7 | 1.800 | 0.794 |
| 26 | 2 | 1.331 | 1.882 | 4 | 0.668 | 0.915 | 5 | 0.772 | 0.510 | 9 | 1.181 | 0.471 |
| 27 | 1 | 0.000 | . | 7 | 0.199 | 0.272 | 4 | 1.352 | 1.138 | 8 | 1.061 | 0.576 |
| 28 | 2 | 0.134 | 0.116 | 4 | 0.574 | 0.485 | 8 | 0.658 | 0.441 | 6 | 0.775 | 0.341 |
| 29 | 7 | 1.129 | 0.487 | 6 | 1.406 | 1.014 | 7 | 0.875 | 0.532 | . | . | . |
| 30 | 4 | 0.332 | 0.618 | 3 | 0.441 | 0.576 | 7 | 0.365 | 0.394 | 6 | 0.426 | 0.269 |
| 31 | 11 | 0.073 | 0.098 | 5 | 0.495 | 0.759 | 4 | 0.860 | 0.926 | . | . | . |
| 32 | 6 | 0.410 | 0.383 | 5 | 0.238 | 0.220 | 6 | 0.692 | 1.086 | 3 | 0.299 | 0.215 |
| 33 | 4 | 0.000 | 0.000 | 6 | 0.038 | 0.060 | 5 | 0.122 | 0.123 | 5 | 0.087 | 0.195 |
| 34 | 7 | 0.974 | 0.824 | 2 | 1.492 | 0.047 | 6 | 1.240 | 0.645 | 5 | 1.166 | 0.484 |
| 35 | 7 | 0.841 | 0.958 | 2 | 0.053 | 0.076 | 7 | 0.567 | 0.528 | 4 | 0.650 | 0.505 |
| 36 | 3 | 1.004 | 1.350 | 8 | 0.425 | 0.387 | 4 | 0.509 | 0.459 | 5 | 1.165 | 0.591 |
| 37 | 11 | 0.765 | 0.842 | 8 | 1.278 | 0.791 | . | . | . | 1 | 0.711 | . |
| 38 | 2 | 0.360 | 0.509 | 5 | 1.047 | 0.786 | 4 | 2.059 | 2.506 | 9 | 2.228 | 0.686 |
| 39 | 2 | 0.029 | 0.042 | 10 | 0.040 | 0.095 | 6 | 0.129 | 0.085 | 2 | 0.138 | 0.057 |
| 40 | 1 | 0.023 | . | . | . | . | 10 | 0.236 | 0.394 | 9 | 0.167 | 0.256 |
| 41 | 13 | 0.252 | 0.323 | 6 | 0.410 | 0.374 | 1 | 0.297 | . | . | . | . |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 42 | 1 | 1.053 | . | 12 | 2.025 | 1.139 | 5 | 1.441 | 0.844 | 2 | 1.045 | 1.016 |
| 43 | 10 | 0.209 | 0.129 | 6 | 0.321 | 0.253 | 4 | 1.844 | 1.711 | . | . | . |
| 44 | 8 | 1.613 | 0.868 | 6 | 1.165 | 0.855 | 3 | 1.769 | 0.285 | 3 | 1.449 | 0.385 |
| 45 | 1 | 0.265 | . | 6 | 0.253 | 0.294 | 7 | 0.662 | 0.509 | 6 | 0.558 | 0.365 |
| 46 | 2 | 0.384 | 0.137 | 3 | 0.510 | 0.075 | 8 | 0.738 | 0.689 | 7 | 0.627 | 0.373 |
| 47 | 1 | 0.000 | . | 3 | 0.289 | 0.444 | 9 | 0.204 | 0.124 | 7 | 0.253 | 0.136 |
| 48 | 5 | 0.947 | 0.700 | 2 | 0.176 | 0.249 | 7 | 0.758 | 0.961 | 6 | 0.736 | 0.353 |
| 49 | 11 | 0.466 | 0.662 | 8 | 0.354 | 0.403 | 1 | 0.225 | . | . | . | . |
| 50 | 3 | 0.348 | 0.384 | 6 | 0.217 | 0.312 | 7 | 0.176 | 0.098 | 4 | 0.448 | 0.314 |

| | Pertinence | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | | | 1 | | | 2 | | | 3 | | |
| | %FLEXION | | | %FLEXION | | | %FLEXION | | | %FLEXION | | |
| | N | MOY | É.T. | N | MOY | É.T. | N | MOY | É.T. | N | MOY | É.T. |
| Requête | | | | | | | | | | | | |
| 1 | 3 | 0.103 | 0.178 | 12 | 0.229 | 0.489 | 4 | 0.154 | 0.194 | 1 | 0.000 | . |
| 2 | 6 | 0.000 | 0.000 | 3 | 0.000 | 0.000 | 8 | 0.009 | 0.025 | 3 | 0.000 | 0.000 |
| 3 | 6 | 0.137 | 0.145 | 6 | 0.362 | 0.504 | 4 | 0.177 | 0.090 | 4 | 0.267 | 0.086 |
| 4 | 4 | 0.109 | 0.092 | 9 | 0.104 | 0.209 | 2 | 0.056 | 0.011 | 5 | 0.133 | 0.143 |
| 5 | 11 | 0.172 | 0.162 | 4 | 0.081 | 0.063 | 1 | 0.166 | . | 4 | 0.142 | 0.082 |
| 6 | 2 | 0.360 | 0.464 | 3 | 0.142 | 0.162 | 8 | 0.169 | 0.262 | 7 | 0.036 | 0.041 |
| 7 | 16 | 0.100 | 0.245 | 2 | 0.000 | 0.000 | 1 | 0.000 | . | 1 | 0.000 | . |
| 8 | . | . | . | 3 | 0.860 | 1.057 | 8 | 0.118 | 0.078 | 9 | 0.213 | 0.168 |
| 9 | 3 | 0.240 | 0.334 | 6 | 0.527 | 0.772 | 5 | 0.499 | 0.420 | 6 | 0.069 | 0.070 |
| 10 | 9 | 0.222 | 0.333 | 2 | 0.140 | 0.198 | 4 | 0.157 | 0.237 | 5 | 0.350 | 0.609 |
| 11 | 2 | 0.938 | 0.616 | 6 | 1.180 | 1.507 | 5 | 1.073 | 0.897 | 7 | 0.595 | 0.347 |
| 12 | 7 | 0.138 | 0.139 | 5 | 0.183 | 0.247 | 4 | 0.770 | 0.712 | 4 | 0.122 | 0.086 |
| 13 | 8 | 0.151 | 0.426 | 5 | 0.000 | 0.000 | 4 | 0.000 | 0.000 | 3 | 0.000 | 0.000 |
| 14 | 8 | 0.537 | 0.541 | 5 | 0.447 | 0.411 | 3 | 0.177 | 0.176 | 4 | 0.560 | 0.515 |
| 15 | 14 | 0.134 | 0.220 | 5 | 0.191 | 0.203 | . | . | . | 1 | 0.136 | . |
| 16 | 13 | 0.649 | 0.472 | 3 | 0.620 | 0.791 | 4 | 1.360 | 0.438 | . | . | . |
| 17 | 3 | 1.543 | 1.520 | 4 | 2.156 | 1.624 | 12 | 1.630 | 0.997 | 1 | 0.125 | . |
| 18 | 3 | 0.000 | 0.000 | 8 | 0.000 | 0.000 | 6 | 0.000 | 0.000 | 3 | 0.000 | 0.000 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 19 | 3 | 0.397 | 0.223 | 10 | 0.512 | 0.630 | 4 | 0.472 | 0.156 | 3 | 0.447 | 0.248 |
| 20 | 3 | 0.432 | 0.494 | 2 | 0.664 | 0.138 | 6 | 0.238 | 0.200 | 9 | 0.204 | 0.165 |
| 21 | 7 | 0.169 | 0.279 | 4 | 2.031 | 3.712 | 7 | 0.258 | 0.118 | 2 | 0.819 | 0.418 |
| 22 | 2 | 0.000 | 0.000 | 9 | 0.036 | 0.109 | 6 | 0.006 | 0.015 | 3 | 0.000 | 0.000 |
| 23 | 3 | 0.101 | 0.107 | 9 | 0.865 | 0.578 | 5 | 1.468 | 1.051 | 3 | 0.432 | 0.374 |
| 24 | 1 | 1.653 | . | 3 | 0.149 | 0.206 | 7 | 0.105 | 0.215 | 9 | 0.365 | 0.294 |
| 25 | 2 | 0.047 | 0.066 | 4 | 0.118 | 0.103 | 7 | 0.512 | 0.318 | 7 | 0.142 | 0.112 |
| 26 | 2 | 0.491 | 0.157 | 4 | 0.391 | 0.317 | 5 | 0.071 | 0.057 | 9 | 0.402 | 0.238 |
| 27 | 1 | 0.000 | . | 7 | 0.594 | 0.871 | 4 | 1.198 | 0.567 | 8 | 1.145 | 0.805 |
| 28 | 2 | 0.386 | 0.472 | 4 | 0.835 | 0.637 | 8 | 1.814 | 2.050 | 6 | 1.023 | 0.538 |
| 29 | 7 | 0.345 | 0.311 | 6 | 0.115 | 0.083 | 7 | 0.536 | 0.516 | . | . | . |
| 30 | 4 | 0.296 | 0.356 | 3 | 0.926 | 0.943 | 7 | 0.553 | 0.430 | 6 | 0.638 | 0.440 |
| 31 | 11 | 0.073 | 0.088 | 5 | 0.109 | 0.087 | 4 | 0.404 | 0.362 | . | . | . |
| 32 | 6 | 0.665 | 0.360 | 5 | 0.383 | 0.360 | 6 | 0.421 | 0.161 | 3 | 0.455 | 0.395 |
| 33 | 4 | 0.154 | 0.209 | 6 | 0.000 | 0.000 | 5 | 0.069 | 0.154 | 5 | 0.015 | 0.033 |
| 34 | 7 | 0.609 | 1.187 | 2 | 0.417 | 0.180 | 6 | 0.670 | 0.622 | 5 | 0.740 | 0.285 |
| 35 | 7 | 0.385 | 0.482 | 2 | 1.903 | 1.393 | 7 | 0.728 | 0.505 | 4 | 0.324 | 0.371 |
| 36 | 3 | 0.210 | 0.189 | 8 | 0.272 | 0.214 | 4 | 0.291 | 0.262 | 5 | 0.319 | 0.294 |
| 37 | 11 | 1.377 | 1.430 | 8 | 0.505 | 0.413 | . | . | . | 1 | 0.711 | . |
| 38 | 2 | 0.015 | 0.021 | 5 | 0.064 | 0.092 | 4 | 0.045 | 0.058 | 9 | 0.158 | 0.137 |
| 39 | 2 | 0.000 | 0.000 | 10 | 0.000 | 0.000 | 6 | 0.000 | 0.000 | 2 | 0.000 | 0.000 |
| 40 | 1 | 0.023 | . | . | . | . | 10 | 0.246 | 0.233 | 9 | 0.336 | 0.393 |
| 41 | 13 | 0.130 | 0.178 | 6 | 0.592 | 0.585 | 1 | 0.149 | . | . | . | . |
| 42 | 1 | 0.000 | . | 12 | 0.364 | 0.435 | 5 | 0.399 | 0.297 | 2 | 0.443 | 0.430 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 43 | 10 | 0.082 | 0.087 | 6 | 0.532 | 0.738 | 4 | 0.302 | 0.234 | . | . | . |
| 44 | 8 | 0.182 | 0.172 | 6 | 0.049 | 0.070 | 3 | 0.143 | 0.042 | 3 | 0.147 | 0.119 |
| 45 | 1 | 0.000 | . | 6 | 0.000 | 0.000 | 7 | 0.000 | 0.000 | 6 | 0.000 | 0.000 |
| 46 | 2 | 0.157 | 0.166 | 3 | 0.340 | 0.480 | 8 | 0.512 | 0.401 | 7 | 0.420 | 0.527 |
| 47 | 1 | 0.000 | . | 3 | 0.000 | 0.000 | 9 | 0.000 | 0.000 | 7 | 0.000 | 0.000 |
| 48 | 5 | 0.329 | 0.382 | 2 | 1.388 | 0.799 | 7 | 1.530 | 0.617 | 6 | 1.886 | 1.130 |
| 49 | 11 | 0.841 | 0.812 | 8 | 0.482 | 0.536 | 1 | 0.188 | . | . | . | . |
| 50 | 3 | 0.534 | 0.374 | 6 | 0.593 | 0.861 | 7 | 0.922 | 0.618 | 4 | 0.540 | 0.136 |

| | Pertinence | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | | | 1 | | | 2 | | | 3 | | |
| | %TOTAL | | | %TOTAL | | | %TOTAL | | | %TOTAL | | |
| | N | MOY | É.T. | N | MOY | É.T. | N | MOY | É.T. | N | MOY | É.T. |
| Requête | | | | | | | | | | | | |
| 1 | 3 | 28.269 | 44.816 | 12 | 2.291 | 1.668 | 4 | 1.526 | 0.603 | 1 | 2.558 | . |
| 2 | 6 | 1.303 | 1.009 | 3 | 1.181 | 0.897 | 8 | 1.473 | 1.047 | 3 | 0.854 | 0.376 |
| 3 | 6 | 2.231 | 1.889 | 6 | 2.706 | 2.015 | 4 | 2.783 | 1.833 | 4 | 3.462 | 1.087 |
| 4 | 4 | 3.167 | 1.498 | 9 | 3.284 | 2.329 | 2 | 1.853 | 0.088 | 5 | 4.701 | 3.611 |
| 5 | 11 | 2.115 | 1.135 | 4 | 2.443 | 1.289 | 1 | 2.819 | . | 4 | 3.207 | 0.290 |
| 6 | 2 | 0.836 | 0.180 | 3 | 2.695 | 1.483 | 8 | 2.781 | 0.859 | 7 | 2.402 | 1.110 |
| 7 | 16 | 1.593 | 1.255 | 2 | 2.141 | 0.490 | 1 | 2.189 | . | 1 | 3.640 | . |
| 8 | . | . | . | 3 | 3.214 | 1.755 | 8 | 2.574 | 1.089 | 9 | 3.292 | 1.278 |
| 9 | 3 | 4.685 | 4.770 | 6 | 6.102 | 3.018 | 5 | 6.934 | 2.511 | 6 | 3.900 | 2.099 |
| 10 | 9 | 1.182 | 0.816 | 2 | 2.429 | 1.724 | 4 | 2.168 | 1.178 | 5 | 2.481 | 0.883 |
| 11 | 2 | 5.327 | 1.705 | 6 | 7.927 | 5.163 | 5 | 6.576 | 2.754 | 7 | 4.046 | 1.718 |
| 12 | 7 | 2.765 | 1.979 | 5 | 3.109 | 1.340 | 4 | 3.315 | 1.643 | 4 | 2.866 | 1.073 |
| 13 | 8 | 1.774 | 1.544 | 5 | 5.143 | 2.540 | 4 | 3.572 | 2.015 | 3 | 2.493 | 0.887 |
| 14 | 8 | 3.169 | 1.211 | 5 | 3.078 | 1.377 | 3 | 2.450 | 1.026 | 4 | 2.939 | 0.622 |
| 15 | 14 | 0.871 | 0.480 | 5 | 1.515 | 0.478 | . | . | . | 1 | 2.102 | . |
| 16 | 13 | 2.443 | 1.414 | 3 | 3.113 | 1.217 | 4 | 3.737 | 0.988 | . | . | . |
| 17 | 3 | 3.473 | 2.853 | 4 | 4.090 | 1.942 | 12 | 3.938 | 0.974 | 1 | 4.299 | . |
| 18 | 3 | 4.795 | 5.471 | 8 | 3.249 | 1.465 | 6 | 2.862 | 1.545 | 3 | 1.879 | 0.489 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 19 | 3 | 1.843 | 1.011 | 10 | 5.374 | 6.508 | 4 | 3.006 | 1.142 | 3 | 2.575 | 1.695 |
| 20 | 3 | 1.525 | 1.638 | 2 | 3.096 | 0.430 | 6 | 3.301 | 1.458 | 9 | 2.677 | 0.630 |
| 21 | 7 | 1.002 | 1.099 | 4 | 3.298 | 4.559 | 7 | 1.596 | 0.492 | 2 | 2.630 | 0.613 |
| 22 | 2 | 1.147 | 0.017 | 9 | 1.910 | 0.612 | 6 | 2.052 | 0.784 | 3 | 2.225 | 0.971 |
| 23 | 3 | 1.950 | 1.608 | 9 | 3.473 | 1.539 | 5 | 3.100 | 1.439 | 3 | 2.841 | 0.317 |
| 24 | 1 | 3.306 | . | 3 | 3.716 | 3.550 | 7 | 3.212 | 1.597 | 9 | 3.817 | 1.643 |
| 25 | 2 | 2.302 | 0.181 | 4 | 1.945 | 1.954 | 7 | 2.217 | 0.922 | 7 | 3.189 | 1.112 |
| 26 | 2 | 2.573 | 2.276 | 4 | 4.573 | 3.190 | 5 | 1.737 | 0.770 | 9 | 3.393 | 1.002 |
| 27 | 1 | 2.727 | . | 7 | 4.998 | 3.396 | 4 | 3.922 | 0.955 | 8 | 4.227 | 1.130 |
| 28 | 2 | 1.544 | 0.564 | 4 | 2.275 | 0.595 | 8 | 3.712 | 2.656 | 6 | 2.777 | 0.976 |
| 29 | 7 | 3.188 | 1.184 | 6 | 4.227 | 1.351 | 7 | 3.765 | 1.266 | . | . | . |
| 30 | 4 | 1.940 | 2.140 | 3 | 6.383 | 4.929 | 7 | 2.946 | 0.686 | 6 | 3.091 | 0.842 |
| 31 | 11 | 0.534 | 0.679 | 5 | 0.720 | 0.871 | 4 | 1.440 | 1.433 | . | . | . |
| 32 | 6 | 3.610 | 2.247 | 5 | 1.996 | 1.087 | 6 | 2.539 | 1.283 | 3 | 3.177 | 1.630 |
| 33 | 4 | 1.332 | 1.531 | 6 | 3.970 | 2.981 | 5 | 2.268 | 0.771 | 5 | 1.912 | 0.449 |
| 34 | 7 | 2.894 | 2.560 | 2 | 3.566 | 0.503 | 6 | 4.171 | 1.288 | 5 | 4.065 | 1.198 |
| 35 | 7 | 4.151 | 2.394 | 2 | 3.118 | 1.490 | 7 | 4.735 | 1.938 | 4 | 4.683 | 2.320 |
| 36 | 3 | 2.362 | 2.757 | 8 | 1.865 | 1.150 | 4 | 2.522 | 0.927 | 5 | 3.440 | 0.955 |
| 37 | 11 | 3.800 | 2.189 | 8 | 3.579 | 1.562 | . | . | . | 1 | 1.991 | . |
| 38 | 2 | 1.449 | 1.004 | 5 | 2.749 | 1.314 | 4 | 3.178 | 2.559 | 9 | 3.476 | 0.770 |
| 39 | 2 | 0.850 | 0.204 | 10 | 2.930 | 1.513 | 6 | 3.397 | 1.522 | 2 | 3.029 | 2.484 |
| 40 | 1 | 0.440 | . | . | . | . | 10 | 4.089 | 1.573 | 9 | 3.939 | 1.341 |
| 41 | 13 | 1.604 | 1.284 | 6 | 2.410 | 0.524 | 1 | 1.245 | . | . | . | . |
| 42 | 1 | 1.755 | . | 12 | 3.854 | 2.090 | 5 | 2.740 | 1.150 | 2 | 2.200 | 1.496 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 43 | 10 | 1.247 | 0.574 | 6 | 2.504 | 1.808 | 4 | 5.384 | 3.833 | . | . | . |
| 44 | 8 | 3.139 | 1.290 | 6 | 2.171 | 0.605 | 3 | 3.020 | 0.236 | 3 | 2.534 | 0.669 |
| 45 | 1 | 1.415 | . | 6 | 2.314 | 0.805 | 7 | 3.967 | 2.236 | 6 | 5.460 | 2.089 |
| 46 | 2 | 2.106 | 1.395 | 3 | 2.923 | 0.818 | 8 | 3.042 | 1.118 | 7 | 4.088 | 0.609 |
| 47 | 1 | 2.365 | . | 3 | 2.807 | 0.459 | 9 | 3.276 | 1.362 | 7 | 2.738 | 0.536 |
| 48 | 5 | 2.919 | 0.864 | 2 | 3.050 | 1.487 | 7 | 4.337 | 2.146 | 6 | 4.644 | 1.009 |
| 49 | 11 | 2.843 | 2.212 | 8 | 1.652 | 0.830 | 1 | 1.203 | . | . | . | . |
| 50 | 3 | 1.921 | 0.970 | 6 | 3.131 | 2.674 | 7 | 5.084 | 1.613 | 4 | 4.001 | 1.280 |

| | Pertinence | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | | | 1 | | | 2 | | | 3 | | |
| | TEXTE | | | TEXTE | | | TEXTE | | | TEXTE | | |
| | N | MOY | É.T. | N | MOY | É.T. | N | MOY | É.T. | N | MOY | É.T. |
| Requête | | | | | | | | | | | | |
| 1 | 3 | 137.667 | 166.133 | 12 | 2065.750 | 2655.277 | 4 | 2088.250 | 1398.278 | 1 | 2424.000 | . |
| 2 | 6 | 1538.833 | 835.219 | 3 | 1171.000 | 586.665 | 8 | 4000.625 | 4098.027 | 3 | 5927.333 | 5886.724 |
| 3 | 6 | 7381.000 | 8306.732 | 6 | 1897.167 | 1534.280 | 4 | 3108.500 | 3155.221 | 4 | 1551.000 | 1033.848 |
| 4 | 4 | 1631.500 | 832.988 | 9 | 3067.889 | 3906.624 | 2 | 4200.000 | 2985.405 | 5 | 1476.400 | 752.263 |
| 5 | 11 | 2103.545 | 2062.449 | 4 | 2397.250 | 2365.563 | 1 | 603.000 | . | 4 | 1901.500 | 1712.146 |
| 6 | 2 | 5091.000 | 6171.628 | 3 | 1539.333 | 628.397 | 8 | 2171.875 | 3229.487 | 7 | 2800.857 | 1320.371 |
| 7 | 16 | 1301.375 | 727.791 | 2 | 2208.500 | 1133.492 | 1 | 2421.000 | . | 1 | 1621.000 | . |
| 8 | . | . | . | 3 | 1098.000 | 702.368 | 8 | 1417.875 | 563.607 | 9 | 2197.667 | 1429.466 |
| 9 | 3 | 762.000 | 523.986 | 6 | 317.167 | 158.095 | 5 | 1351.200 | 1538.220 | 6 | 3881.500 | 3941.874 |
| 10 | 9 | 1336.444 | 901.643 | 2 | 838.500 | 333.047 | 4 | 1380.500 | 1101.665 | 5 | 3044.000 | 3430.068 |
| 11 | 2 | 381.000 | 24.042 | 6 | 256.167 | 104.463 | 5 | 791.200 | 517.812 | 7 | 3959.714 | 4652.844 |
| 12 | 7 | 2022.857 | 1872.716 | 5 | 2466.000 | 3716.901 | 4 | 971.500 | 642.313 | 4 | 16300.250 | 23172.939 |
| 13 | 8 | 3679.375 | 5471.748 | 5 | 487.600 | 465.240 | 4 | 751.000 | 82.942 | 3 | 1809.333 | 1360.614 |
| 14 | 8 | 1573.500 | 1188.583 | 5 | 1742.400 | 1574.460 | 3 | 2548.000 | 1428.094 | 4 | 2090.750 | 1963.186 |
| 15 | 14 | 2777.429 | 1480.578 | 5 | 2524.000 | 1665.890 | . | . | . | 1 | 1475.000 | . |
| 16 | 13 | 1114.769 | 812.384 | 3 | 937.000 | 309.543 | 4 | 1415.500 | 528.936 | . | . | . |
| 17 | 3 | 1535.667 | 455.755 | 4 | 2706.750 | 1512.170 | 12 | 3178.167 | 2087.524 | 1 | 1605.000 | . |
| 18 | 3 | 579.333 | 649.013 | 8 | 530.000 | 434.917 | 6 | 724.667 | 458.000 | 3 | 622.667 | 129.218 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 19 | 3 | 617.333 | 385.000 | 10 | 388.000 | 256.392 | 4 | 1347.750 | 1023.130 | 3 | 1369.000 | 767.586 |
| 20 | 3 | 1417.333 | 1386.218 | 2 | 373.500 | 28.991 | 6 | 1046.667 | 612.887 | 9 | 1267.111 | 316.583 |
| 21 | 7 | 1634.857 | 625.955 | 4 | 1175.000 | 1149.443 | 7 | 1105.286 | 762.050 | 2 | 1315.000 | 844.285 |
| 22 | 2 | 613.000 | 379.009 | 9 | 819.222 | 510.098 | 6 | 2016.833 | 733.206 | 3 | 2001.667 | 1361.489 |
| 23 | 3 | 2112.333 | 2708.710 | 9 | 713.222 | 328.993 | 5 | 1040.400 | 694.025 | 3 | 692.667 | 223.003 |
| 24 | 1 | 121.000 | . | 3 | 822.333 | 767.376 | 7 | 1984.000 | 2516.165 | 9 | 2046.556 | 1457.237 |
| 25 | 2 | 604.000 | 659.024 | 4 | 1959.750 | 2846.908 | 7 | 1355.143 | 725.064 | 7 | 3014.571 | 2436.259 |
| 26 | 2 | 546.500 | 400.930 | 4 | 575.750 | 707.091 | 5 | 2770.800 | 1730.457 | 9 | 2169.556 | 1645.775 |
| 27 | 1 | 110.000 | . | 7 | 698.571 | 571.364 | 4 | 1404.500 | 771.141 | 8 | 5353.000 | 8446.449 |
| 28 | 2 | 1656.000 | 376.181 | 4 | 1006.250 | 488.280 | 8 | 1239.625 | 447.927 | 6 | 2520.500 | 2802.007 |
| 29 | 7 | 1255.143 | 658.073 | 6 | 1309.000 | 824.092 | 7 | 1362.429 | 883.709 | . | . | . |
| 30 | 4 | 7285.500 | 14090.468 | 3 | 221.667 | 193.913 | 7 | 1145.143 | 1135.395 | 6 | 1657.500 | 881.686 |
| 31 | 11 | 3662.273 | 3863.423 | 5 | 6197.200 | 2551.482 | 4 | 11613.000 | 12833.670 | . | . | . |
| 32 | 6 | 483.667 | 284.262 | 5 | 1357.400 | 1873.780 | 6 | 2760.000 | 3730.907 | 3 | 2106.000 | 1966.127 |
| 33 | 4 | 546.000 | 325.224 | 6 | 832.333 | 741.571 | 5 | 1066.800 | 510.027 | 5 | 1146.000 | 500.271 |
| 34 | 7 | 4413.286 | 7107.452 | 2 | 8679.000 | 10326.587 | 6 | 1137.667 | 682.724 | 5 | 1510.800 | 871.750 |
| 35 | 7 | 636.571 | 486.993 | 2 | 903.500 | 44.548 | 7 | 1289.000 | 676.464 | 4 | 2215.000 | 713.455 |
| 36 | 3 | 1040.000 | 707.488 | 8 | 1151.125 | 886.914 | 4 | 1539.250 | 1044.876 | 5 | 1171.200 | 621.433 |
| 37 | 11 | 1084.909 | 938.494 | 8 | 1075.000 | 537.503 | . | . | . | 1 | 1406.000 | . |
| 38 | 2 | 1937.500 | 1974.949 | 5 | 1217.200 | 989.484 | 4 | 1066.250 | 409.607 | 9 | 2886.222 | 3535.573 |
| 39 | 2 | 1302.500 | 562.150 | 10 | 641.400 | 372.840 | 6 | 813.333 | 196.839 | 2 | 1910.500 | 1257.943 |
| 40 | 1 | 4319.000 | . | . | . | . | 10 | 847.600 | 230.802 | 9 | 933.667 | 413.982 |
| 41 | 13 | 3304.308 | 3662.594 | 6 | 2156.833 | 2202.320 | 1 | 8753.000 | . | . | . | . |
| 42 | 1 | 2279.000 | . | 12 | 1785.500 | 2729.514 | 5 | 2155.600 | 482.242 | 2 | 14870.500 | 13649.282 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 43 | 10 | 2553.800 | 1871.470 | 6 | 4546.667 | 2264.753 | 4 | 3194.250 | 2146.072 | . | . | . |
| 44 | 8 | 2602.625 | 2025.280 | 6 | 5563.500 | 2448.266 | 3 | 3406.000 | 2101.371 | 3 | 12183.667 | 10623.543 |
| 45 | 1 | 1131.000 | . | 6 | 706.167 | 246.237 | 7 | 1465.286 | 1693.297 | 6 | 1091.000 | 656.782 |
| 46 | 2 | 6981.500 | 7815.651 | 3 | 1502.000 | 190.197 | 8 | 1528.250 | 885.579 | 7 | 4978.286 | 9282.592 |
| 47 | 1 | 761.000 | . | 3 | 1858.333 | 2263.949 | 9 | 1721.556 | 741.218 | 7 | 2801.714 | 1957.948 |
| 48 | 5 | 1010.000 | 478.385 | 2 | 1106.500 | 840.750 | 7 | 1642.571 | 1614.826 | 6 | 1186.833 | 528.098 |
| 49 | 11 | 1152.364 | 850.232 | 8 | 3369.875 | 3391.001 | 1 | 2661.000 | . | . | . | . |
| 50 | 3 | 1674.333 | 1221.381 | 6 | 1873.333 | 1442.968 | 7 | 1343.286 | 765.439 | 4 | 2763.500 | 1703.423 |

# BIBLIOGRAPHIE


Assadi H. et Bourigault D. (1996). « Acquisition et modélisation des connaissances à partir de textes : outils informatiques et éléments méthodologiques ». Actes du 10ème congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA'96). Rennes, France

Bailey P., Craswel N. et Hawking D. (2003). « Engineering a multi-purpose test collection for web retrieval experiments ». Information Processing and Management, 39 : 853-871.

Blair D.C. (1990). *Language and Representation in Information Retrieval. Elsevier Science Publishers, Amsterdam.*

_____ . (2002). « Some thoughts on the reported results of TREC ». Information processing and Management, 38 : 445-451.

Bonnel N. et Moreau F. (2005). « Quel avenir pour les moteurs de recherche? ». Actes de la 3ème manifestation des jeunes chercheurs francophones dans les domaines des STIC (MAJECSTIC'05), p. 291-299, Rennes, France.

Buckley C. et Voorhees E. M. (2004). « Retrieval Evaluation with Incomplete Information ». Proceedings of ACM-SIGIR'04, Sheffield, UK, p. 25-32.

Bouillon P., Vandooren Fr, Da Sylva L., Jacqmin L., Lehmann S., Russell G. and Viegas E. (1998). « *Traitement automatique des langues naturelles* ». Duculot, Paris, Bruxelles.

Bouillon P., Fabre C., Sébillot P. et Jacqmin L. (2000). « Apprentissage de ressources lexicales pour l'extension de requêtes ». TAL 41-2 : 367-393.

Chieze E. (2000). « Prise en compte de la morphologie du français dans le repérage d'information sur le Web ». Mémoire de maîtrise, Montréal, Université du Québec à Montréal, 145 p.

_____ . (2006). « Reformulation automatique de requêtes par intégration d'éléments syntaxiques dans le cadre du repérage de l'information en français sur le Web ». Thèse de doctorat, Montréal, Université du Québec à Montréal, 249 p.

Chieze E., Bouchard L. et Emirkanian L. (2000). « Connaissance linguistique et recherche sur le Web ». Actes du Colloque international «Traduction humaine, Traduction automatique, Interprétation», 28-30 septembre 2000, CERES, Tunis, série linguistique n° 11, p. 21-37.

Chieze E., Emirkanian L. et Bouchard L. (2001). « Impact de la prise en compte de la morphologie sur le repérage d'information sur le Web ». Distances, vol. 5, n° 2, p. 177-194.

Chignell M., Gwizdka J. et Bodner R.C. (1999). « Discriminating meta-search: a framework for evaluation ». *Information Processing and Management,* 35: 337-362.

Chbeir R. (2001). « Modélisation de la description d'images : Application au domaine médical ». Thèse de doctorat, INSA (Institut National des Sciences Appliquées), Lyon.

Cleverdon C.W. (1967). « The Cranfield tests on index language devices ». *Aslib Proceedings*, 19-6 : 173-193.

Corbin D. (1991). « La morphologie lexicale : bilan et perspectives ». *Travaux de linguistique* 23 : 33-56.

Cormack G.V., Palmer C.R., Van Biesbrouck M. et Clarke C.L.A. (1998). « Deriving very short queries for high precision and recall (MultiText Experiments for TREC-7) ». Proceedings of TREC-7, Gaithersburg, Maryland, p. 121-132.

Daille B. et Morin E. (2000). « Reconnaissance automatique des noms propres de la langue écrite : les récentes réalisations ». TAL, 41-3 : 601-621.

Dias G., Guillore S., Bassano J.C. et Pereira-Lopes J.G. (2000). « Extraction automatique d'unités lexicales complexes : un enjeu fondamental pour la recherche documentaire ». TAL, 41-2 : 447-472.

Emirkanian L. et Chieze E. (2003). « Variations morphologiques, syntaxiques, sémantiques et repérage d'information sur le Web ». Revue Québécoise de Linguistique, 32-1 : 135-154.

Fradin B. (2003). *Nouvelles approches en morphologie.* PUF, Paris.

Gaussier E., Grefenstette G., Hull D. et Roux Cl. (2000). « Recherche d'information en français et traitement automatique des langues ». TAL 41-2 : 473-493.

Habert B. et Jacquemin C. (1993). « Noms composés, termes, dénominations complexes : problématiques linguistiques et traitements automatiques ». TAL, 34-2 : 5-41.

Haddad H. (2003). « French noun phrase indexing and mining for an information retrieval system ». In Nascimento, MA, De Moura E.S. et Oliveira A.L. (éd.), SPIRE 2003, LNCS 2857, Springer-Verlag, p. 277-286.

Hamon T. et Nazarenko A. (2001). « Detection of synonymy links between terms : experiment and results ». *Recent Advances in Computational Terminology*, edited by Bourigault D., Jacquemin C. and L'Homme M.-C., John Benjamins Publishing Company, Amesterdam, p. 185-208.

Harman D.K. (1992). « Overview of the Second Text REtrieval Conference (TREC-2) ». Proceedings of the Second Text REtrieval Conference, NIST Special Publication, 500-215, 1-20.

Hawking D., Craswel N., Thistlewaite P. et Harman D. (1999). « Results and Challenges in Web Search Evaluation ». *Computer Networks*, 31: 1321-1330.

Hollander M. et al. (1973). *Nonparametric Statistical Methods*. John Wiley & Sons, New York.

Jacquemin C. (1997). « Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes en corpus ». Mémoire d'Habilitation à Diriger des Recherches en informatique fondamentale, Université de Nantes, Nantes, France.

Jacquemin Ch. et Tzoukermann E. (1997). « Ana lyse automatique de la morphologie dérivationnelle et filtrage de mots possibles ». In *Proceedings, Forum de morphologie 1ères rencontres : Mots Possibles et Mots Existants*, Lille. SILEX, Université de Lille 3. Silexicales 1 :251-260.

Jacquemin Ch. et Tzoukermann E. (1999). « NLP for term variant extraction: Synergy of morphology, lexicon and syntax ». In Strzalkowski T. et al., p. 25-74.

Jacquemin, C. et Zweigenbaum, P. (2000). « Traitement automatique des langues pour l'accès au contenu des documents ». In J. Le Maitre, J. Charlet and C. Garbay, editors, *Le document Multimédia en Sciences du Traitement de l'Information*, p. 71-110. CÉPADUÈS-Éditions, Toulouse, France.

Jing H. et Tzoukermann E. (1999). « Information Retrieval Based on context Distance and Morphology ». SIGIR'99, Berkley, CA, p. 90-96.

Leighton H.V. et Srivastava J. (1999). « First 20 Precision among World Wide Web Search Services (search Engines) ». *Journal of the American Society for Information Science*, 50-10 : 870-881.

Lespinasse K., Kremer P., Schibler D. et Schmitt L. (1999). « Evaluation des outils d'accès à l'information textuelle, les expériences américaine (TREC) et française (AMARYLLIS) ». *Langues*. Vol. 2 : 2, p. 99-109.

Lewis D. et Sparck Jones K. (1996). « Natural language processing for information retrieval». *Communications of the ACM*, 39-1 : 92-101.

Loupy C. (2001). « L'apport de connaissances linguistiques en recherche documentaire ». TALN'2001, Tome 2, p. 129-143, Tours, France.

Loupy C., Combet V. et Crestan É. (2003). « Linguistic resources for Information Retrieval», paru dans les actes de ENABLER/ELSNET, atelier International Roadmap for Language Resources, 28-29 août 2003, Paris, France.

Martinet J., Chiaramella Y. et Mulhem P. (2002). « Un modèle vectoriel étendu de recherche d'information adapté aux images ». 20ème Congrès INFORSID'02, p. 337-348, 4-7 juin 2002, Nantes, France.

Moreau F. et Claveau V. (2006). « Extension de requêtes par relations morphologiques acquises automatiquement ». *Actes de la 3ème Conférence en Recherche d'Informations et Applications, (CORIA'06)*, p. 181-192, Mars 2006, Lyon, France.

Moreau F. et Sébillot P. (2005). « Contributions des techniques du traitement automatique des langues à la recherche d'information ». Rapport de Recherche IRISA, No 1690

Namer F. (2000). « FLEM : un analyseur flexionnel du français à base de règles ». TAL 41-2: 523-547.

Paice C.D. (1996). « Method for Evaluation of Stemming Algorithms Based on Error Counting ». JASIS, 47-8 : 632-649.

Porter M. (1980). « An algorithm for suffix stripping », *Program*, 14 (3), pp. 130-137.

Robert M. (1988). *Fondements et étapes de la recherche scientifique en psychologie.* St-Hyacinthe, Québec.

Salton G. et McGill M. (1983). *Introduction to Modern Information Retrieval.* New-York : McGraw-Hill.

Saracevic T. (1970). « The concept of relevance in information science : A historical review». In Saracevic T (Ed.), *Introduction to Information Science*, New-York : R.R. Bowker, 111-151.

Savoy J. (1993). « Stemming of French Words Based on Grammatical Categories ». *Journal of the American Society for Information Science,* 44-1 : 1-9.

_____ . (2003). «Cross-language information retrieval : experiments based on CLEF 2000 corpora». *Information Processing and Management,* 39 : 75-115.

Sormunen E. (2000). « Liberal Relevance Criteria of TREC – Counting on Negligible Documents ». Proceedings of SIGIR'02, Tampere, Finlande, 324-330.

Sparck-Jones K. (1999). « What is the role of NLP in Information Retrieval ? ». *Natural Language Information retrieval*, edited by Strzalkowski T., p. 1-24, Kluwer Academic Publishers, Dordrecht.

Strzalkowski T. (1995). « Natural language information retrieval ». *Information Processing and Management*, 31-3 : 397-417.

Strzalkowski T., Lin F., Wang J. et Perez-Carballo J. (1999). « Evaluating Natural Language Processing Techniques in Information Retrieval ». *Natural Language Information Retrieval,* edited by Strzalkowski T., Kluwer Academic Publishers, Dordrecht, p. 113-146,

Strzalkowski T. et al. (1999). *Natural Language Information Retrieval.* Dordrecht, Kluwer.

Voorhees E.M. (2000). « Variations in relevance judgments and the measurement of retrieval effectiveness ». *Information Processing and Management*, 36 : 697-716.

Voorhees E.M., Tice D.M. (2000). « The TREC-8 Question Answering Track evaluation ». *Proceedings of the 8th Text Retrieval Conference,* NIST.

Woods W.A., Bookman L.A., Houston A., Kuhns R.J., Martin P., Green S. et al. (2000). «Linguistic knowledge can improve information retrieval ». Processing of the 6th Applied Natural Language Processing Conference.

Zweigenbaum P., Grabar N. et Darmoni S. (2001). « L'apport de connaissances morphologiques pour la projection de requêtes sur une terminologie normalisée ». TALN 2001, p. 403-408, 2-5 juillet 2001, Tours, France.