

Étude sur l'influence du vocabulaire utilisé pour l'indexation de l'image numérique sur le repérage en contexte multilingue

Résumé : Cette communication examine le repérage de l'image, en contexte de repérage multilingue, lorsque que celle-ci est indexée selon l'une ou l'autre des deux approches suivantes : indexation en vocabulaire contrôlé ou indexation en vocabulaire libre. Cette communication présente le contexte et la problématique de la recherche, ainsi que l'expérience effectuée.

Abstract: This paper describes a research project aiming to verify the existence of relations between two indexing approaches, traditional image indexing recommending the use of controlled vocabularies or free image indexing using uncontrolled vocabulary, and their respective performance for image retrieval, in a multilingual context. This paper presents the context of the research project, as well as the experiment carried out.

1. Énoncé du problème

Depuis quelques années, Internet est devenu un média incontournable pour la diffusion de ressources multilingues. De nos jours, on considère qu'il existe environ 7 000 langues vivantes à travers le monde. Même s'il est difficile d'estimer le nombre exact de langues écrites parmi celles-ci, en raison principalement du manque de sources d'information fiables et disponibles, nous pouvons supposer que la plupart de ces langues généreront éventuellement des documents sous forme textuelle ou autre. Sur le web, les différences linguistiques constituent souvent un obstacle majeur aux échanges de documents scientifiques, culturels, pédagogiques et commerciaux. En outre, la recherche d'information est confrontée au problème de la surabondance des résultats. Ce problème, loin de s'amenuiser, prend de l'ampleur avec l'accroissement du web et l'émergence d'une variété de langues dans ce dernier. À l'heure actuelle, le web constitue un vaste univers de connaissances et de cultures humaines diverses, permettant le partage d'idées et d'information et cela, dans une multitude de langues. Par conséquent, l'accès à ce foisonnement d'information multilingue est devenu un défi de taille.

En plus de cette diversité linguistique, on constate le développement croissant de bases de données et de collections composées de différents types de documents textuels ou multimédias, ce qui complexifie également le processus du repérage. Par exemple, le repérage d'images présente un niveau de difficulté important. Parmi les nombreux types d'images, l'image ordinaire (i.e., non artistique) occupe une place importante dans les nombreuses recherches des utilisateurs du web. Par images ordinaires, on entend les images représentant des objets de la vie quotidienne, par exemple : un téléviseur, une tondeuse à gazon, un biberon, etc. Le qualificatif « ordinaire » distingue ce type d'images avec les images dites d'art dont la valeur première à titre de document visuel est d'être la manifestation d'une expression artistique (Turner, 1998). La croissance du web a mis en relief le besoin pressant de se doter d'outils propres à la description des images dans le but de faciliter leur repérage, puisque l'on retrouve celles-ci dans la plupart des

ressources web : pages et sites personnels, bibliothèques et collections muséales virtuelles, catalogues de services et produits commerciaux, information gouvernementale, etc.

En général, deux catégories de requêtes sont utilisées pour le repérage de l'image. L'individu peut soumettre une requête « image » (utilisation d'une image ou d'un dessin). Le système tente alors de repérer des images similaires par l'utilisation de certaines caractéristiques physiques de l'image (forme, couleur, texture, etc.). L'individu peut également formuler une requête textuelle, comme dans la majorité des recherches d'images sur le web. Le succès du repérage dépend alors de la correspondance devant être établie entre la requête du chercheur d'images et le texte associé aux images. Étant donné la grande diversité linguistique qui existe sur le web, on doit s'attendre à ce que les textes associés aux images existent en de nombreuses langues différentes. Par conséquent, lorsqu'il utilise une requête textuelle, le chercheur d'images fait face à un double défi lorsque vient le moment de repérer celles-ci. D'une part, les termes de la requête doivent correspondre au texte associé aux images et d'autre part, la langue de sa requête doit concorder à la langue du texte associé aux images. Cette recherche propose d'étudier cette double problématique à laquelle fait face l'individu, lors du repérage d'images en contexte multilingue.

Bien que démontrant certaines similitudes avec le repérage de documents textuels à proprement parler, le repérage de l'image se distingue quant à certains éléments tels que la formulation des requêtes, la méthode utilisée pour le repérage, la manière dont la pertinence des résultats est évaluée, la participation de l'individu au processus de recherche et les différences cognitives fondamentales d'interprétation du matériel visuel plutôt que textuel (Clough et Sanderson, 2003). En général, on considère l'image comme libre au point de vue linguistique. En d'autres termes, on ne peut attribuer d'emblée une langue à ce type de documents. Par exemple, on ne dira pas d'une image qu'elle est de langue chinoise ou allemande, comme on le dirait d'un document textuel rédigé dans une de ces langues. Toutefois, l'indexation de l'image lui confère un statut linguistique au même titre que tout document textuel, ce qui peut avoir une incidence sur le repérage. Par exemple, si le chercheur d'images fait une requête en français et que les images à repérer sont indexées en anglais (ou tout autre langue différente de la langue de la requête), le repérage se fait à condition que le système de repérage prévoit un mécanisme de recherche d'information multilingue (RIML) permettant la mise en correspondance des termes de la requête avec les termes d'indexation, même si la langue de la requête est différente de la langue des termes d'indexation.

Pour que cette correspondance se fasse, deux approches sont habituellement proposées en recherche d'information multilingue. La première consiste à traduire les documents au complet dans toutes les langues, alors que la deuxième suppose la traduction des requêtes dans la langue des documents à repérer. Ces deux approches font généralement appel à trois types de ressources linguistiques pour la traduction des requêtes ou des documents : les dictionnaires bilingues ou multilingues, les systèmes de traduction automatique (TA) et les corpus parallèles ou comparables. Pour cette recherche, l'approche privilégiant la traduction des requêtes est utilisée, de même qu'un système de traduction automatique comme processus de traduction. D'une part, la traduction des requêtes, plutôt que des documents, semble mieux adaptée aux fins de la recherche puisque généralement, les requêtes contiennent peu de mots. La traduction de celles-ci peut donc se faire rapidement et à un moindre coût. D'autre part, l'utilisation d'un système de traduction automatique comme ressource linguistique a été retenue, car ce type de mécanismes peut

être facilement intégré à l'intérieur d'un système de repérage. En outre, on considère que la performance de la recherche d'information utilisant un mécanisme de traduction automatique, en contexte multilingue, se situe généralement entre 80 et 100 pour cent de la performance du repérage en contexte monolingue (Nie, 2006). Toutefois, peu importe l'approche utilisée (traduction de la requête ou des documents) et la ressource linguistique employée pour la traduction (dictionnaire, traduction automatique ou corpus), les risques d'ambiguïté au moment de la traduction, demeurent importants.

Depuis le XIX^e siècle, le catalogage, la classification et l'indexation ont pour principal objectif de traiter essentiellement les documents textuels. Pourtant, l'accès au matériel audiovisuel soulève autant d'intérêt pour les chercheurs, sinon plus. Selon Guinchat et Menou (1990), « l'indexation des documents non écrits pose des problèmes particuliers en raison de leur nature propre et de leur mode de consultation ». Comme on peut le supposer, le choix des termes d'indexation appropriés pour la description des images revêt une importance capitale, que ce soit en contexte de repérage monolingue ou multilingue. Cette recherche propose d'examiner ce qui se produit au moment du repérage de l'image lorsque celle-ci est indexée selon l'une ou l'autre des deux approches suivantes : (1) indexation préconisant l'utilisation du vocabulaire contrôlé, c'est-à-dire que les termes d'indexation sont tirés d'un langage artificiel dont la principale fonction est de générer des représentations formelles de documents (Hudon, 2007), et (2) indexation libre en vocabulaire non contrôlé, c'est-à-dire que les termes d'indexation servant à décrire l'image n'obéissent à aucune règle préétablie. Ces deux approches d'indexation présentent des caractéristiques fondamentalement différentes pouvant influencer le repérage de l'image en contexte multilingue.

La première approche d'indexation à l'étude, préconisant l'utilisation du vocabulaire contrôlé, suppose généralement qu'un schéma de métadonnées prévoit la structure, les éléments et le vocabulaire à utiliser. Un schéma de métadonnées se définit comme un ensemble d'éléments conçus pour répondre aux besoins de communautés spécifiques (Smiraglia, 2005). Certains schémas de métadonnées sont de nature générale, alors que d'autres ont été conçus à des fins spécifiques comme la description de l'information gouvernementale, des données géospatiales, des ressources multimédias, etc. En plus de la structure et des éléments devant être utilisés, certains schémas de métadonnées recommandent également quels vocabulaires doivent être employés pour l'indexation. Généralement, les schémas de métadonnées suggèrent l'utilisation de deux types de vocabulaires : le vocabulaire contrôlé ou le vocabulaire non contrôlé. Toutefois, la plupart des schémas de métadonnées préconisent l'utilisation du vocabulaire contrôlé, afin d'accroître la cohérence descriptive (Jørgensen, 2003).

De prime abord, les vocabulaires contrôlés ont pour objectif de faciliter le processus d'indexation. En outre, l'utilisation des vocabulaires contrôlés offre de nombreux avantages pour le repérage, le furetage et l'interopérabilité. Toutefois, les vocabulaires contrôlés présentent également certaines faiblesses dont la principale est de représenter les concepts de manière artificielle. En effet, la plupart des vocabulaires contrôlés incluent des termes corrects au niveau linguistique, mais ces termes sont souvent peu utilisés par les individus dans leur vie quotidienne. Par exemple, un vocabulaire contrôlé contiendra des expressions comme « brodequin de travail », alors que « botte de travail » est beaucoup plus usuelle. Par conséquent, on remarque souvent une grande divergence entre les termes d'indexation offerts par le vocabulaire contrôlé et les termes généralement utilisés dans la formulation des requêtes des individus, au moment du repérage (Furnas et al., 1987). À souligner également que les vocabulaires contrôlés

deviennent rapidement désuets, ce qui constitue un autre désavantage puisque les néologismes mettront souvent beaucoup de temps à apparaître dans les différents vocabulaires contrôlés. En outre, l'utilisation de ces vocabulaires demeure une tâche complexe pour la plupart des indexeurs. Finalement, la majorité des vocabulaires contrôlés prévus par les schémas de métadonnées et habituellement utilisés pour l'indexation de l'image n'existe qu'en anglais. Par conséquent, un indexeur possédant peu de connaissances de la langue anglaise et désirant utiliser ces vocabulaires contrôlés, risque de se heurter à un problème linguistique important, à moins qu'un mécanisme de traduction facilitant leur utilisation ne soit prévu.

La deuxième approche d'indexation à l'étude constitue une alternative à l'indexation traditionnelle effectuée avec le vocabulaire contrôlé. En effet, l'indexation libre qui, en plus de n'imposer aucune structure ou élément obligatoire, privilégie généralement le vocabulaire non contrôlé. Cette forme d'indexation, longtemps négligée, a été remise récemment au goût du jour. À titre d'exemple de l'indexation libre en vocabulaire non contrôlé, mentionnons l'indexation que l'on retrouve notamment dans les systèmes de flickrs qui permettent le stockage massif et le partage d'images sur le web. Ces systèmes de partage offrent la possibilité aux utilisateurs d'indexer leurs propres images et de rendre celles-ci publiques, c'est-à-dire que les images peuvent être vues par tous ou par un ensemble de personnes choisies par l'utilisateur du système, formant ainsi une très grande base communautaire d'images. Dans un système de partage d'images, un utilisateur qui téléverse une image peut ainsi déterminer qui aura accès à cette image en établissant certaines règles de contrôle d'accès. Parallèlement, les autres utilisateurs du système ont la possibilité de mettre quotidiennement à jour l'indexation des images en ajoutant des mots-clés (*tags*) ou en inscrivant des commentaires sur toute image à laquelle ils ont accès. Les annotations attribuées par le téléverseur ou par tout autre utilisateur du système de partage constituent une forme d'indexation en vocabulaire non contrôlé appelée étiquetage collaboratif (*collaborative tagging*). En outre, dans ce type de systèmes, l'indexation peut se faire en une seule langue ou combiner plusieurs langues. Cependant, malgré sa popularité grandissante et à l'instar de l'indexation en vocabulaire contrôlé, l'indexation en vocabulaire non contrôlé présente aussi plusieurs lacunes. Par exemple, certaines ambiguïtés émergent puisque souvent, un même mot-clé est employé par plusieurs individus, mais dans des contextes divers. De la même manière, le manque de contrôle synonymique entraîne l'utilisation de nombreux mots-clés différents pour décrire un même concept. Par conséquent, et pour beaucoup d'autres raisons, l'indexation en vocabulaire non contrôlé est souvent perçue comme de piètre qualité.

L'utilisation du vocabulaire contrôlé ou non contrôlé mène à un certain nombre de difficultés au moment de l'indexation. Ces difficultés auront nécessairement des conséquences au moment du repérage de l'image. L'indexation avec l'un ou l'autre de ces vocabulaires est une question abondamment débattue dans la littérature. Au fil des années, plusieurs chercheurs se sont penchés sur la question du vocabulaire devant être utilisé pour le processus d'indexation. De nombreuses études ont examiné les avantages de chacune des deux approches d'indexation décrites précédemment, dans divers contextes, sans toutefois parvenir à un consensus clair : « l'éternel débat sur l'avantage du vocabulaire contrôlé par opposition au vocabulaire libre dans les systèmes de recherche d'information persiste » (Arsenault, 2006, 142). Or, il semble que les nombreuses difficultés associées à l'indexation en vocabulaire non contrôlé peuvent seulement être comprises par l'entremise d'une analyse comparative avec le vocabulaire contrôlé (Macgregor et McCulloch, 2006).

Deux éléments d'importance ressortent de cet état de la question sur l'indexation et le repérage d'images. Dans un premier temps, on constate qu'on connaît très peu de choses sur l'influence du vocabulaire utilisée pour l'indexation de l'image au moment de son repérage en contexte monolingue, et encore moins en contexte multilingue. Dans un deuxième temps, cet état de la question met en relief la grande variabilité du processus du repérage d'images. En effet, depuis quelques années, on assiste à une évolution dans la manière de formuler les requêtes pour le repérage d'images. Cette mutation amène, entre autres, à repenser la manière dont l'image doit être indexée et à se demander si le vocabulaire contrôlé, traditionnellement employé pour l'indexation, est bien adapté à ce type de document. Par conséquent, l'étude des besoins des chercheurs d'images sur le web semble nécessaire afin de concevoir de meilleurs systèmes de repérage, tout comme il est essentiel de poursuivre l'étude des stratégies de repérage des chercheurs d'images, afin de prévoir une meilleure indexation, et plus spécifiquement un vocabulaire mieux adapté aux besoins et comportements réels et actuels des chercheurs d'images.

Compte tenu de ce qui précède, plusieurs questions se posent au sujet du repérage de l'image en contexte multilingue. Ainsi, on peut se demander si le choix du vocabulaire utilisé pour l'indexation de l'image a une influence sur les résultats de son repérage en contexte multilingue et si tel est le cas, quelle est cette influence. En outre, on peut se demander quelle approche d'indexation il est préférable d'adopter pour l'image, afin d'en faciliter son repérage en contexte multilingue. Ces questions, à notre connaissance, ont été peu étudiées jusqu'à maintenant. Pourtant, le processus d'indexation de l'image, dans le but de permettre son repérage efficace, efficient et satisfaisant, en contexte de repérage multilingue, demeure une question délicate qui mérite une attention particulière.

2. Objectifs de l'étude

Cette étude examine le repérage de l'image numérique, en contexte de repérage multilingue, lorsque que celle-ci est indexée selon l'une ou l'autre des deux approches suivantes : indexation préconisant l'utilisation du vocabulaire contrôlé ou indexation libre en vocabulaire non contrôlé. Le but de cette recherche est de vérifier l'existence de relations entre chacune de ces deux approches d'indexation et leur performance respective, lors du repérage de l'image, en contexte multilingue. Trois objectifs spécifiques ont ainsi été définis :

1. Identifier les caractéristiques de chacune des deux approches d'indexation de l'image pouvant influencer le repérage, en contexte multilingue.
2. Explorer l'existence de relations entre les caractéristiques de chacune des deux approches d'indexation de l'image et les résultats du repérage, en contexte multilingue.
3. Déterminer la direction des relations entre les caractéristiques de chacune des deux approches d'indexation de l'image et les résultats du repérage, en contexte multilingue.

La réalisation de ces trois objectifs permettra de mesurer empiriquement l'influence de l'approche d'indexation de l'image, en contexte de repérage multilingue. Cette étude suppose que les deux approches d'indexation présentent des caractéristiques communes, mais également des différences pouvant influencer le repérage de l'image. Cette recherche permet ainsi d'estimer si l'une ou l'autre de ces approches d'indexation surclasse l'autre, en termes d'efficacité, d'efficience et de satisfaction du chercheur d'images. Cette recherche permet ainsi de dégager les éléments essentiels d'un traitement

adapté aux collections d'images numériques, puisque celles-ci renferment une richesse informationnelle inégalée. Il sera éventuellement possible d'étendre la recherche à d'autres environnements linguistiques et d'autres types d'images.

3. Méthodologie

3.1 Approche générale

Cette étude examine le repérage de l'image, en contexte de repérage multilingue, lorsque celle-ci est indexée selon l'une ou l'autre des deux approches suivantes : indexation en vocabulaire contrôlé ou indexation libre en vocabulaire non contrôlé. Deux contextes linguistiques sont comparés : (1) le contexte de repérage monolingue, c'est-à-dire que la langue de la requête (français) est la même que la langue d'indexation (français), et (2) le contexte de repérage multilingue, c'est-à-dire que la langue de la requête (français) est différente de la langue d'indexation (anglais). Soulignons toutefois que le contexte monolingue est considéré ici uniquement dans le seul but d'établir la comparaison avec le repérage en contexte multilingue. En outre, cette recherche utilise une approche quantitative. Avec cette approche, régulièrement utilisée en sciences de l'information pour mesurer et comparer la performance du repérage des documents textuels et visuels, les concepts sont étudiés et testés à l'aide d'hypothèses de recherche basées sur des variables observables et clairement définies.

3.2 Collecte des données

Trois méthodes de collecte de données sont employées pour cette recherche : (1) l'analyse des termes utilisés pour l'indexation des images, (2) la simulation du repérage d'un sous-ensemble d'images indexées selon chacune des approches d'indexation à l'étude, et (3) le questionnaire administré aux chercheurs d'images au cours de la simulation du repérage.

Dans un premier temps, une base d'images ordinaires a été constituée à partir d'images tirées d'un catalogue commercial en ligne. La collecte des images, effectuée entre le 26 décembre 2006 et le 10 janvier 2007, a permis de recueillir 3 950 images puisées parmi les huit grandes sections du catalogue : sports et loisirs, maison, auto, outils, cuisine et bain, électronique, côté jardin et habillement. La sélection des images était basée principalement sur trois critères : (1) la qualité visuelle de l'image illustrant un objet spécifique, (2) la facilité d'identification de l'image par un non-spécialiste, et (3) la diversité maximale des images incluses dans la base d'images. Chacune des images a ensuite été indexée de quatre manières différentes : en vocabulaire contrôlé français et anglais et en vocabulaire libre français et anglais. Pour cette recherche, le français et l'anglais ont été retenus pour deux raisons. D'une part, la langue française est la langue utilisée par l'échantillon de participants visé par la recherche et d'autre part, la plupart des vocabulaires contrôlés destinés à l'indexation des images existent en anglais ce qui rend presque obligatoire la sélection de l'anglais comme deuxième langue d'indexation.

L'indexation en vocabulaire libre français a été effectuée par un indexeur francophone, alors que l'indexation en vocabulaire libre anglais a été effectuée par un indexeur anglophone. L'indexation en vocabulaire contrôlé français et anglais a été réalisée par un indexeur maîtrisant ces deux langues. Pour cette partie de l'indexation, le *Nouveau dictionnaire visuel multilingue* a été retenu comme vocabulaire contrôlé. Trois raisons justifient ce choix. Premièrement, ce dictionnaire contient les termes pouvant convenir au type d'images contenues dans la base, c'est-à-dire des images ordinaires d'objets de la vie quotidienne. Deuxièmement, le *Visuel* offre une forme de normalisation des termes

permettant une identification claire et précise des images, tout en exerçant un contrôle maximal sur les régionalismes, les archaïsmes, les anglicismes, etc. Finalement, le *Visuel* comporte à la fois une terminologie française et anglaise ce qui facilite le processus d'indexation, tout en assurant une certaine forme de cohérence interlinguistique. L'indexation en vocabulaire contrôlé a été faite à l'aide du dictionnaire format papier et de sa version CD-ROM (*Le Visuel 3*). La figure 1 présente un extrait de la version CD-ROM :

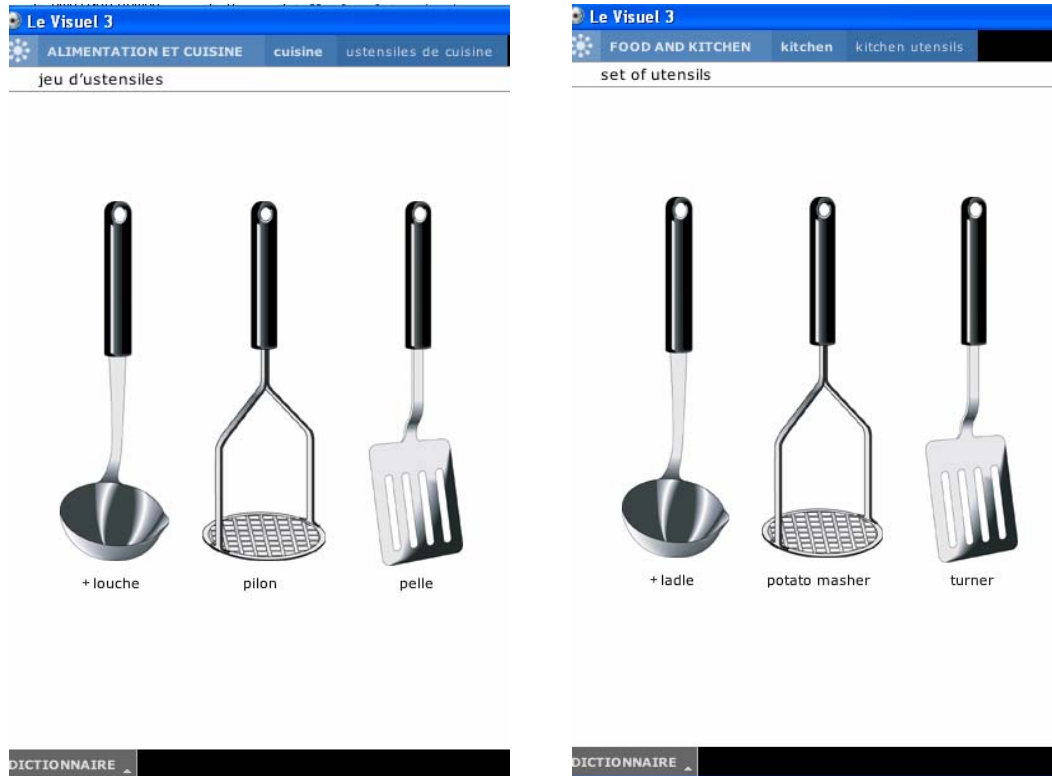


Figure 1 - Extrait du Visuel 3

Le processus d'indexation de la base d'images s'est déroulé du 10 février et le 31 mars 2007. Une fois toutes les images indexées avec chacune des quatre formes d'indexation, l'ensemble des termes d'indexation du vocabulaire libre et contrôlé français a été fusionné, de même que l'ensemble des termes d'indexation en vocabulaire libre et contrôlé anglais, produisant ainsi deux autres formes d'indexation. La base d'images associe donc six formes d'indexation à chacune des 3 950 images. Suite au processus d'indexation, une grille d'analyse des termes d'indexation est en cours d'élaboration. L'objectif de cette grille d'analyse est de recenser les caractéristiques des termes d'indexation tirés de chacune des approches d'indexation à l'étude.

Parallèlement au développement de la base d'images, un système de repérage d'images est en voie de conception. Ce système tente de reproduire les conditions du repérage d'images que l'on trouve généralement sur le web. Ce système de repérage inclut : (1) une interface de recherche permettant au chercheur d'images de formuler une requête textuelle (voir annexe 1), (2) un mécanisme de traduction de la requête du français vers l'anglais, (3) un module de mise en correspondance des termes des requêtes (ou de la traduction des requêtes) avec les termes d'indexation associés aux images, et (4) une interface d'affichage des résultats du repérage permettant de sélectionner les images satisfaisant la requête (voir annexe 2). Pour la simulation du repérage, un sous-ensemble de 30 images a été constitué aléatoirement à partir de la base et sont recherchées par tous

les participants. La sélection des 30 images à repérer lors de la simulation de repérage vise à : (1) maximiser la variété des images à repérer pour contrer l'effet d'apprentissage, et (2) privilégier le choix d'images présentant des niveaux de difficulté de repérage similaires. Chacune des 30 images est associée aléatoirement à une forme d'indexation. L'appariement image/forme d'indexation a été distribué de manière balancée (5 images par forme d'indexation). Ces 30 paires image/forme d'indexation constituent les 30 tâches de repérage que chaque participant doit effectuer au moment de la simulation du repérage. Ainsi, lors de l'expérience, chacune des 30 images est montrée à tour de rôle au participant. Ce dernier formule une requête textuelle en français pour repérer chacune des images. Les requêtes pour la moitié des images à repérer sont mises en correspondance avec une forme d'indexation en français, alors que les requêtes pour l'autre moitié des images à repérer sont d'abord traduites en anglais à l'aide du mécanisme de traduction automatique intégré au système de repérage, puis mises en correspondance avec une forme d'indexation en anglais, tel qu'illustré à la figure 2 :

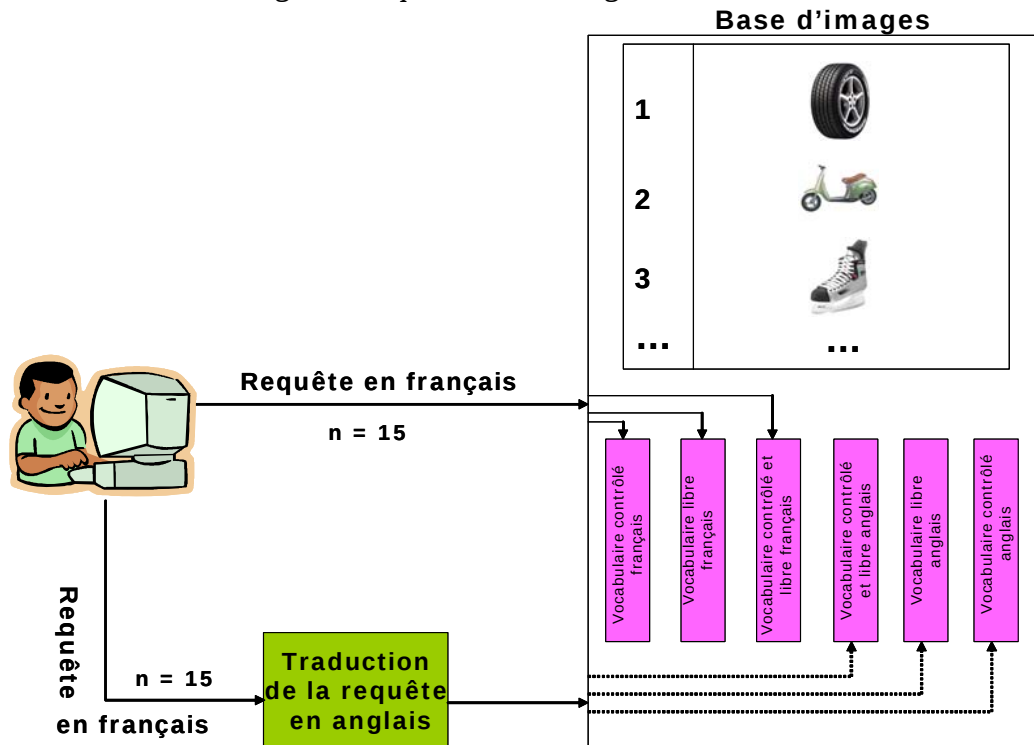


Figure 2 - Schéma de la simulation du repérage

À partir de l'interface d'affichage des résultats, le participant doit ensuite sélectionner l'image qui satisfait la tâche de repérage. Chaque image considérée satisfaisante servira au calcul du taux de succès du repérage. Le participant peut également décider de reformuler une ou plusieurs autres requêtes. Afin d'éviter l'effet d'apprentissage, l'ordre des tâches de repérage est déterminé de manière aléatoire et modifié pour chaque participant. Chacune des paires image/forme d'indexation est recherchée par des groupes de sept participants, ce qui permet d'évaluer individuellement chacune des formes d'indexation, en termes d'efficacité, d'efficacité et de satisfaction du participant. La simulation du repérage sera effectuée à l'automne 2007 auprès de quarante-deux participants. La sélection des participants à la simulation du repérage se fera en fonction des quatre critères suivants : (1) participants âgés de 18 ans et plus, (2) participants dont la langue maternelle est le français, (3) étudiants de l'Université de Montréal de niveau 1^{er} cycle, et (4) étant donné la nature de la tâche à effectuer lors de l'expérience, c'est-à-

dire le repérage d'images, les participants ne doivent avoir aucune expérience professionnelle dans un domaine utilisant le repérage d'images.

Finalement, afin de compléter les données recueillies lors de la simulation du repérage, un questionnaire visant à recueillir les impressions des participants face au processus de repérage et aux résultats obtenus a été développé. Ce questionnaire est administré aux participants en deux étapes. En premier lieu, puisque nous désirons mesurer la satisfaction du chercheur d'images au sujet de chacune des approches d'indexation à l'étude, nous demandons au participant d'estimer son degré de satisfaction face aux résultats obtenus après chacune des tâches de repérage. En second lieu, un questionnaire global contenant des questions fermées et ouvertes est soumis au participant, une fois que toutes les tâches de repérage sont complétées. Ce questionnaire compte des questions sur : (1) la perception de la tâche de repérage en général, (2) les habitudes de recherche et la familiarité avec le repérage d'images et, (3) le profil démographique des participants.

4. Observations et quantifications

Lors du processus d'indexation en vocabulaire contrôlé, l'indexeur a reçu comme directive d'utiliser la politique d'indexation spécifiquement conçue pour cette recherche et qui préconise l'utilisation du *Nouveau dictionnaire visuel multilingue* pour l'indexation des 3 950 images contenues dans la base. L'indexation en vocabulaire contrôlé en français et en anglais a été réalisée par un seul et même indexeur. Afin d'assurer la cohérence interindexeurs, un certain nombre d'images (7 pour cent du nombre total d'images contenues dans la base) a été indexé par un deuxième indexeur. On considère la cohérence interindexeurs comme la capacité chez plusieurs indexeurs d'analyser un même document de la même façon et de représenter les mêmes concepts par les mêmes termes (Bertrand-Gastaldy, 1986). Le calcul du taux de cohérence interindexeurs est le rapport, exprimé en pourcentage, entre le nombre de termes communs (C) de deux indexeurs (A et B), multiplié par deux, divisé par le nombre total de termes choisis par ces mêmes indexeurs, selon la formule suivante : $2C \div (A + B)$ (David, 2003). En ce qui concerne la cohérence interindexeurs en vocabulaire contrôlé français et anglais, un taux de 73,06 pour cent a été obtenu. Ce taux a été calculé à l'identique, c'est-à-dire que les deux indexeurs doivent avoir choisi exactement les mêmes termes d'indexation pour que ceux-ci soient considérés comme des termes communs. Par exemple, les termes « raquette » et « raquette de tennis » ne sont pas considérés comme des termes communs.

Une politique spécifique pour l'indexation libre a également été élaborée pour cette recherche. Cette politique prévoit que l'indexeur utilise ses propres termes pour décrire chacune des images contenues dans la base, sans aucune forme de contrôle du sens ou de la forme. L'indexation en vocabulaire libre en français et en anglais se fait respectivement par un seul indexeur possédant les compétences linguistiques nécessaires pour accomplir cette tâche. À l'instar de l'indexation en vocabulaire contrôlé et afin d'assurer la cohérence interindexeurs, un certain nombre d'images (10 pour cent du nombre total d'images contenues dans la base) a été indexé en vocabulaire libre par un deuxième indexeur. La cohérence interindexeurs en vocabulaire libre a également été calculée à l'identique. Pour cette forme d'indexation, un taux de 9,75 pour cent a été obtenu pour l'indexation en vocabulaire libre français et de 11,00 pour cent pour le vocabulaire libre anglais. À noter que la divergence entre les taux de cohérence interindexeurs en vocabulaire contrôlé et en vocabulaire libre concorde avec les conclusions des études portant sur l'utilisation d'un vocabulaire contrôlé qui soutiennent

que le taux de cohérence interindexeurs augmente avec l'utilisation d'un tel vocabulaire comparativement à l'emploi d'un vocabulaire libre (David, 2003).

Le processus d'analyse des termes d'indexation, actuellement en cours, permettra de caractériser chacune des approches d'indexation à l'étude et de mettre en relief les similitudes et les différences qui existent entre les deux approches d'indexation et qui pourraient influencer le repérage de l'image en contexte multilingue. Les résultats de cette analyse seront disponibles en septembre 2007.

La quantification de la performance du repérage de chacune des approches d'indexation à l'étude est basée sur les mesures habituellement employées pour évaluer la performance des systèmes de repérage (précision et rappel), ainsi que sur les mesures d'utilisabilité préconisées par la norme ISO 9241-11 (efficacité et efficience). La troisième mesure, la satisfaction, est quantifiée à l'aide du troisième mode de collecte des données prévu par la recherche, le questionnaire. En général, l'efficacité réfère à la capacité à atteindre un objectif donné, alors que l'efficience fait plutôt référence à la capacité d'effectuer une tâche donnée avec le minimum d'efforts. Traditionnellement, plusieurs indicateurs peuvent être pris en compte pour ces deux mesures. Pour cette recherche, le taux de succès, à savoir la capacité à atteindre partiellement ou totalement l'objectif fixé par la tâche, est considéré comme le principal indicateur de l'efficacité. En ce qui concerne l'efficience, une distinction est faite entre deux formes spécifiques : (1) l'efficience humaine du repérage d'images qui est mesurée par l'efficacité divisée par l'effort humain, soit le nombre total de requêtes formulées pour une tâche de repérage donnée et (2) l'efficience temporelle du repérage d'image qui est mesurée par l'efficacité divisée par le temps, en secondes, passé pour effectuer la sélection de l'image satisfaisant une tâche de repérage donnée.

Pour chaque tâche de repérage, les variables suivantes sont enregistrées directement par le système de repérage :

- Formulation des requêtes : termes employés dans chaque requête formulée pour repérer chacune des images montrées ;
- Images repérées : numéro d'identification des images retenues par le participant pour chaque tâche de repérage ;
- Nombre d'essais : nombre total de requêtes formulées pour repérer chacune des images montrées ;
- Durée du repérage : temps requis, en secondes, pour effectuer la tâche de repérage, de la première à la dernière requête, pour chacune des images à repérer ;
- Temps consacré par image repérée : ratio calculé à l'aide de la durée du repérage divisé par le nombre d'images repérées.

En outre, les transactions effectuées par les participants pendant la simulation du repérage seront également consignées à l'aide d'un logiciel d'enregistrement de transactions. Ce mode de collecte non obstructif a pour principal avantage d'enregistrer les transactions et les interactions réelles des utilisateurs avec le système (Spink et Jansen, 2004). La simulation du repérage sera effectuée à l'automne 2007. Les résultats de l'analyse des données enregistrées seront disponibles à l'été 2008.

Finalement, le questionnaire a pour principal objectif de recueillir des informations sur la troisième mesure d'utilisabilité préconisée par la norme ISO 9241-11, c'est-à-dire la satisfaction du chercheur d'images. La quantification de la mesure de satisfaction se fait avec les trois indicateurs suivants :

- Le degré de *satisfaction face aux résultats obtenus* : mesurée par une échelle de satisfaction du participant face aux résultats obtenus pour une tâche de repérage donnée ;
- Le degré de *facilité d'apprentissage* : mesurée par une échelle de perception du participant de la facilité avec laquelle il réussit à effectuer une tâche de repérage donnée ;
- Le degré de *tolérance aux erreurs* : mesurée par une échelle de tolérance du participant envers les erreurs de repérage pour une tâche de repérage donnée.

Les résultats de l'analyse des questionnaires seront également disponibles à l'été 2008.

5. Conclusion

La croissance du web a mis en relief le besoin pressant de se doter d'outils propres à la description documentaire des images numériques, puisque l'on retrouve celles-ci dans la plupart des ressources web. Cette recherche tente d'apporter une contribution à trois niveaux spécifiques. Au niveau théorique, elle permet d'identifier les caractéristiques indispensables à l'indexation de l'image pour en faciliter le repérage en contexte multilingue. Cette recherche tente également d'apporter une contribution au niveau méthodologique grâce à un mode d'évaluation du vocabulaire approprié, pour la description de l'image numérique. Finalement, au niveau pratique, cette recherche permet de dégager les éléments essentiels d'un traitement adapté aux collections d'images numériques. Bref, ce projet de recherche met en lumière l'importance du traitement des documents multimédias, afin de faciliter leur repérage et leur diffusion en contexte multilingue, à une époque où l'information numérique prend une part de plus en plus grande dans nos systèmes d'information, et tente d'offrir une solution pour un meilleur accès aux collections d'images numériques qui constituent des composantes essentielles de la mémoire collective et du patrimoine mondial.

Remerciements

Cette recherche a été réalisée grâce à une subvention du Conseil de recherches en sciences humaines du Canada (CRSH). L'auteure tient également à remercier les membres du Comité de recherche, Monsieur Clément Arsenault, Madame Christine Dufour et Monsieur James Turner, la programmeuse du système de repérage, Madame Geneviève Bastien, ainsi que l'équipe d'indexeurs, Clément Arsenault, Manon Gagnon Rita Maalouf et Chantal Ménard.

Références

- Arsenault, Clément. 2006. L'utilisation des langages documentaires pour la recherche d'information. *Documentation et Bibliothèques* 52 (2): 139-48.
- Association française de normalisation. 1998. *Exigences ergonomiques pour travail de bureau avec terminaux à écrans de visualisation (TEV) - partie 11 : lignes directrices relatives à l'utilisabilité*. Genève: Organisation internationale de normalisation.
- Bertrand-Gastaldy, Suzanne. 1986. De quelques éléments à considérer avant de choisir un niveau d'analyse ou un langage documentaire. *Documentation et bibliothèques* (janvier-juin): 3-23.

- Clough, Paul, et Mark Sanderson. 2003. *The CLEF 2003 cross language image retrieval task*. http://clef.isti.cnr.it/2003/WN_web/45.pdf .
- David, Claire. 2003. *Analyse cognitive de la cohérence interindexeurs lors de l'indexation de documents*. Thèse de doctorat, Université de Montréal.
- Furnas, George W., et al. 1987. The vocabulary problem in human-system communication. *Communications of the ACM* 30 (11): 964-71.
- Guinchat, Claire, et Michel Menou. 1990. *Sciences et techniques de l'information et de la documentation*. Paris: Unesco.
- Hudon, Michèle. 2007. *Notes de cours BLT- 6055 Analyse et représentation documentaires 2*. Montréal: Université de Montréal.
- Jørgensen, Corinne. 2003. *Image retrieval theory and research*. Lanham, MD: Scarecrow Press.
- Macgregor, George, et Emma McCulloch. 2006. Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review* 55 (5): 291-300.
- Nie, Jian-Yun. 2006. *Notes de cours IFT 6255 – Présentation sur SMART*. <http://www.iro.umontreal.ca/~nie/IFT6255/CLIR.html>.
- QA International. 2003. *Le Nouveau dictionnaire visuel multilingue*. Montréal: Québec Amérique.
- QA International. 2006. *Le Visuel 3 multimédia (V3.5)*. Montréal: Québec Amérique.
- Smiraglia, Richard P. 2005. Introducing metadata. *Cataloging & Classification Quarterly* 40 (3/4): 1-15.
- Spink, Amanda, et Bernard J. Jansen. 2004. *Web search: public searching of the Web*. Boston: Kluwer Academic.
- Turner, James. 1998. *Images en mouvement : stockage, repérage, indexation*. Sainte-Foy: Presses de l'Université du Québec.



Objet à rechercher

Tâche 1 de 30



Recherche

Rechercher



Résultats 1 - 12 sur un total de 50 pour caméra numérique

Cochez les images qui vous conviennent



☐ ok



☐ ok



☐ ok



☐ ok



☐ ok



☐ ok



☐ ok



☐ ok



☐ ok



☐ ok



☐ ok



☐ ok

Je veux modifier ma recherche

J'ai terminé ma recherche