# Analyse d'une méthode d'indexation automatique basée sur une analyse syntaxique de texte

(Version finale parue dansCanadian Journal of Information and Library Science / Revue l'information et de bibliothéconomie, 1996, 21(1), 1-21.)

**Najib Faraj†, Robert Godin†, Rokia Missaoui†, Sophie David††, Pierre Plante††**

†*Département d'Informatique*
*Université du Québec à Montréal*
*C.P.8888, Succursale Centre Ville*
*Montréal, Québec*
*Canada, H3C 3P8*
*godin.robert@uqam.ca*

††*Centre d'Analyse de Texte par Ordinateur*
*Université du Québec à Montréal*
*C.P.8888, Succursale Centre Ville*
*Montréal, Québec*
*Canada, H3C 3P8*

*RÉSUMÉ. Une méthode d'indexation automatique basée sur une analyse syntaxique de texte combinée à une analyse statistique de pondération des termes est proposée et évaluée. Plusieurs combinaisons dans le choix des catégories de termes d'index et de méthodes de pondération ont été testées. L'expérimentation effectuée sur un corpus du domaine de l'ingénierie du logiciel révèle une amélioration systématique des performances par l'utilisation de termes composés générés par analyse syntaxique par rapport à l'utilisation de termes simples.*

*ABSTRACT. An automatic indexing method based on syntactical text analysis combined with statistical analysis is proposed and evaluated. Many combinations for the choice of term categories and weighting methods are tested. The experiment conducted on a software engineering corpus shows systematic improvement in the use of syntactic term phrases compared to using only individual words as index terms.*

*MOTS-CLÉS. Indexation automatique, analyse syntaxique de texte.*

## 1. Introduction

Les méthodes statistiques d'indexation par termes simples (termes formés d'un seul mot) basées sur l'analyse des caractéristiques fréquentielles des mots dans une collection de documents montrent des performances en rappel et en précision comparables aux méthodes manuelles (Salton, 1986). Plusieurs approches ont été proposées pour pondérer les termes et les avantages respectifs des diverses combinaisons possibles ont été mis en lumière en particulier dans (Salton & Buckley, 1988). Deux voies sont à considérer pour améliorer ces méthodes:

1. La génération de termes composés. Les termes composés[1] permettent généralement de limiter l'ambiguïté des termes et d'augmenter la précision.

2. L'analyse de la langue naturelle. Une autre voie intéressante pour l'amélioration des performances de l'indexation automatique est de tenter de tirer profit des méthodes d'analyse de la langue naturelle pour raffiner le processus d'indexation en levant par exemple des ambiguïtés de lemmatisation et en suggérant des termes composés.

L'approche d'indexation que nous proposons explore ces deux voies et repose sur la combinaison de méthodes statistiques traditionnelles et d'analyse syntaxique pour des textes français. L'analyse syntaxique est effectuée par le logiciel Termino (David & Plante, 1990). Les termes lemmatisés produits par Termino sont ensuite traités en utilisant des méthodes statistiques. Une évaluation de cette approche dans une application d'ingénierie du logiciel sera présentée.

Les travaux actuels sur l'utilisation de termes composés en indexation automatique ont donné des résultats peu encourageants pour la langue anglaise (Croft, Turtle & Lewis, 1991). Les travaux de Fagan (1987) ont démontré que pour certaines collections, les *termes composés statistiques*,

[1]Nous utilisons l'appellation terme composé dans le sens d'un terme d'index composé de plusieurs mots peu importe la méthode utilisée pour générer le terme.

obtenus par analyse des cooccurrences des termes simples ont produit des augmentations de performance significatives mais que les *termes composés syntaxiques* obtenus par analyse syntaxique du texte n'ont pas produit d'amélioration par rapport aux termes simples. Dans (Lewis, 1992), l'utilisation de termes composés syntaxiques pour le problème de classification n'a pas non plus démontré d'amélioration par rapport aux termes simples.

Les travaux analogues sur des collections en langue française incluant des évaluations formelles sont peu nombreux. Notons en particulier les travaux de (Blosseville, Hébrail, Monteil & Pénot, 1992) qui ont utilisé une méthode de génération de termes composés pour la langue française. Leur méthode se base sur un dictionnaire et est incluse dans un système plus complet comprenant un outil d'analyse statistique et un système expert. La tâche analysée est la classification de projets de recherche. L'évaluation présentée porte sur le système en entier et ne met pas en lumière l'apport spécifique des termes composés.

L'approche que nous utilisons pour l'extraction de termes composés a l'avantage de reposer sur une analyse syntaxique qui n'oblige pas la construction de dictionnaire spécifique du domaine au prix d'un effort important. Diverses combinaisons possibles dans le choix des catégories de termes et des calculs statistiques ont été évaluées formellement sur une collection de textes français.

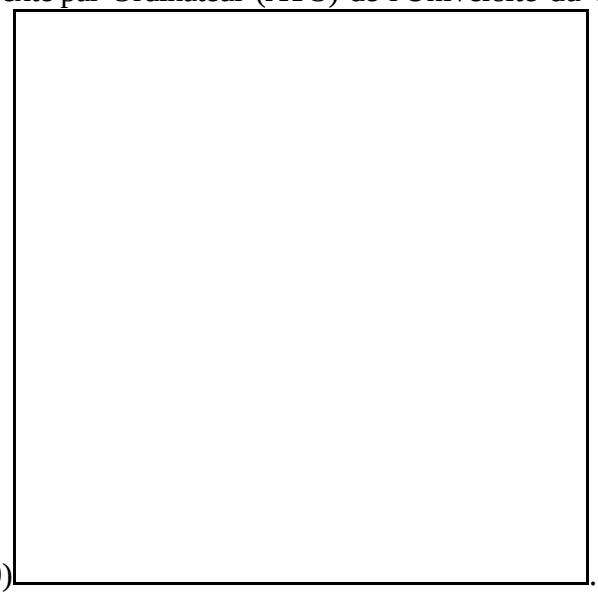
La section deux décrit la méthode d'indexation des points de vues linguistiques et statistiques et explique le modèle de repérage utilisé pour l'expérimentation. La section trois présente une évaluation de la méthode sur un corpus provenant d'une application d'ingénierie du logiciel.

## 2. Méthode d'indexation automatique et modèle de repérage

Dans ce projet, l'analyse faite par le logiciel Termino est le point de départ du processus d'indexation. L'analyse syntaxique sophistiquée produite par Termino permet, entre autres, de lever diverses ambiguïtés de lemmatisation et de générer des termes composés à partir d'une

analyse syntaxique poussée. Nous avons développé un logiciel qui, à partir de la sortie de Termino, produit une indexation automatique en appliquant des méthodes statistiques. Le logiciel est paramétrisé afin de pouvoir étudier diverses combinaisons possibles au niveau du choix des termes et des méthodes de pondération.

### 2.1. Termino: aspects linguistiques et informatiques

Termino est un logiciel de dépouillement terminologique assisté par ordinateur qui a été conçu et réalisé par le groupe *Recherche et Développement en Linguistique Computationnelle* (RDLC) du Centre en Analyse de Texte par Ordinateur (ATO) de l'Université du Québec à Montréal (UQAM) (David & Plante, 1990)  . La théorie linguistique sur lequel il est fondé propose que la grammaire soit organisée en composants autonomes et non hiérarchisés (en d'autres termes, les unités d'un composant et les règles qui lui sont associées sont spécifiques (Marandin, 1992)). Parmi ces composants, celui de la morphologie et celui de la syntaxe sont susceptibles de construire des unités lexicales. Le composant morphologique (Corbin, 1991) les construit selon divers procédés tels que la suffixation, la préfixation, la composition[2],

[2] Aussi, nous réservons le terme de "mot composé" aux seules unités construites par la morphologie via l'opération de composition (Corbin, 1992).

etc. En ce qui concerne les unités nominales, le composant syntaxique les construit dans la position noyau du groupe nominal (GN) (David, 1993; Marandin, 1992). Plus généralement, la théorie syntaxique mise en oeuvre dans Termino est une théorie positionnelle (Cori & Marandin, 1993; Marandin, 1992; Milner, 1989). L'hypothèse fondamentale est la suivante: l'analyse syntaxique d'une phrase n'est pas entièrement déductible des unités lexicales qui la constituent. On distingue dans ce modèle:

- un ensemble de positions, c'est-à-dire un ensemble de "points" hiérarchiquement organisés entre eux, formant ainsi une configuration;

- une relation entre les positions et les constituants lexicaux. Cette relation est appelée relation d'occupation.

L'objet de la syntaxe est alors de rendre compte des positions (identification des positions et de leurs propriétés spécifiques) et de l'occupation de ces positions. Les relations entre les éléments lexicaux ne relèvent pas du composant syntaxique mais du composant lexical.

Dans ce cadre, une tâche essentielle est de retrouver les unités nominales polylexicales (UNP), c'est-à-dire des unités nominales constituées de plusieurs mots, pour lesquelles on peut faire l'hypothèse d'une construction syntaxique. Ces unités sont appelées des synapsies (Benveniste, 1974; David, 1993; Marandin, 1992). Il s'agit d'unités telles que "pomme de terre", "mur du son", "logiciel intégré", "haute cour de justice", etc.

Termino est constitué de plusieurs modules:

- EDITO (*Traitement des marques d'édition*) ;
- LCMF (*Lemmatisation et Caractérisation Morphologique du Français*) ;
- ALSF (*Analyseur Lexico-Syntaxique du Français*) ;
- MRSF (*Module de Reconnaissance des Synapsies du Français*) ;
- MRAF (*Module de Rédaction Assistée des Fiches*) ;

- FX (*Langage de programmation en faisceaux*) [3].

Lorsque l'opération *Description linguistique* est lancée, les textes sont analysés et Termino produit, pour chacun, une série de fichiers dont les principaux sont:

- le fichier des adjectifs ;
- le fichier des noms ;
- le fichier des verbes ;
- le fichier des synapsies.

Chacun des fichiers précédents contient un terme par ligne. Chaque terme est suivi par les numéros de phrases où il apparaît. Par exemple, Termino produit pour le texte de la Figure 1 les fichiers du Tableau 1.

Les besoins en information des nouveaux utilisateurs sont satisfaits par les bonnes performances du système de repérage.
Les avions se rapprochaient trop. Nous avions peur.
La belle porte un parapluie.

**Figure 1.** Exemple de texte.

[3] EDITO a été développé par I. Winter, coll. A. Plante (Centre ATO, UQAM) ; LCMF a été développé par L. Dumas (Centre ATO, UQAM), coll. D. Perras, P. Plante (Centre ATO, UQAM) et A. Plante (Centre ATO, UQAM). ALSF a été développé par J.-M. Marandin (URA 1028, CNRS), S. David (Paris VII & Centre ATO, UQAM) et P. Plante ; MRSF a été développé par S. David. MRAF a été développé pat A. Plante. FX, langage de programmation en faisceaux, a été développé par P. Plante (Plante, 1993).

**Tableau 1.** Fichiers produits par Termino.

| Adjectifs | | Noms | | Verbes | | Synapsies | |
|---|---|---|---|---|---|---|---|
| bon | 1 | avion | 2 | avoir | 3 | besoin en information | 1 |
| nouveau | 1 | belle | 4 | être | 1 | performance du système de repérage | 1 |
| | | besoin | 1 | porter | 4 | système de repérage | 1 |
| | | information | 1 | rapprocher | 2 | | |
| | | parapluie | 4 | satisfaire | 1 | | |
| | | performance | 1 | | | | |
| | | peur | 3 | | | | |
| | | repérage | 1 | | | | |
| | | système | 1 | | | | |
| | | utilisateur | 1 | | | | |

EDITO effectue le découpage du texte en phrases et en mots et reconnaît les noms propres.

Pour chaque mot, LCMF (Dumas, 1990) fournit une catégorie grammaticale (nom commun, adjectif, verbe infinitif, verbe fléchi, participe présent, participe passé) et une caractérisation morphologique (traits de genre, nombre, personne, mode et temps, selon la catégorie). Si un mot est ambigu du point de vue catégoriel, LCMF fournira toutes les hypothèses. La désambiguïsation catégorielle sera opérée à l'étape suivante par le module ALSF. Enfin, LCMF fournit pour chaque mot son lemme (singulier pour les noms communs, masculin singulier pour les adjectifs, infinitif pour les verbes fléchis, les participes passés et les participes présents). LCMF est essentiellement fondé sur une analyse morphologique des formes lexicales. Il a ainsi la possibilité de catégoriser et d'assigner un lemme à des néologismes.

ALSF est appliqué sur les sorties produites par LCMF. Il fournit, à l'aide d'une analyse descendante, une représentation hiérarchisée des différents groupes syntagmatiques de la phrase (groupe nominal (GN), groupe verbal (GV), groupe prépositionnel (GP), groupe adjectival (GA), etc.) ainsi qu'une représentation des relations entre ces groupes (relation sujet par exemple). ALSF est un parseur fondé sur une théorie syntaxique positionnelle, il s'appuie à la fois sur un savoir syntaxique (la géométrie des positions) et sur un savoir lexical (le sous-ensemble des propriétés lexicales qui est mis en jeu dans l'occupation de telle ou telle position). En ce sens, ALSF est un parseur syntaxique autonome. De ce fait, la représentation syntaxique qu'il construit peut être

largement sous-déterminée. Par exemple, en ce qui concerne le problème classique dit de l'ambiguïté de rattachement, ALSF ne construit pas toutes les possibilités de structuration mais laisse une marque dans la représentation pour indiquer qu'il n'a pu déterminer le rattachement du complément envisagé. Quant à la désambiguïsation catégorielle, elle est effectuée en fonction de la position impliquée et de la configuration dans laquelle cette position se situe. Par exemple, si la position considérée est la position noyau de GV et que le mot est catégoriellement ambigu (par exemple, la forme "porte" peut être un nom ou un verbe fléchi), ALSF choisira les informations lexicales associées à "porte" en tant que verbe fléchi. De ce fait, le lemme considéré, dans cette position, sera "porter" et non "porte". ALSF est programmé en FX; ce langage est bien adapté à la manipulation des positions dans un arbre.

Le module MRSF est chargé de construire des synapsies. Une synapsie est une unité nominale polylexicale (UNP), i.e. formée de plusieurs termes, construite syntaxiquement. Elle est composée de groupes prépositionnels, nominaux ou adjectivaux. Elle a toujours une tête (ou base) qui est un nom commun et qui représente le noyau de la synapsie. Le nombre d'éléments composant une synapsie n'est pas limité. Le Tableau 2 présente quelques synapsies ainsi que leur forme syntaxique.

**Tableau 2.** Exemples de synapsies.

| Forme | Exemple | Forme | Exemple |
|---|---|---|---|
| T GA | plante verte | GA T GP | grand livre de banque |
| GA T | haute tension | GA T GA | grand livre auxiliaire |
| T GP | traitement de textes | T GA GP GP | nombre moyen de défauts |
| T GA GA | horaire décalé fixe | | par unité |
| T GA GP | logiciel intégré de gestion | | |
| T GP GA | système de gestion universel | | |
| T GP GP | valeur de l'actif par action | | |

T: tête, GN: groupe nominal, GP: groupe prépositionnel, GA: groupe adjectival.

La détermination des synapsies consiste en une exploration de la représentation syntaxique fournie par ALSF. Une fois que l'on a analysé une phrase, on retrouve tous les groupes nominaux qui la constituent et l'on isole ceux qui sont construits à l'aide de compléments présentant certaines

propriétés (absence de déterminant dans le GN d'un GP (à l'exception des déterminants «la» et «l'»), liste restreinte d'adjectifs situés à gauche de la tête, liste restreinte de prépositions dans un GP, etc.). À chaque synapsie est associée une représentation qui met en jeu à la fois la structuration syntaxique de la synapsie et une trace de son ancrage syntaxique dans la phrase[4]. À la toute fin du texte, toutes les synapsie sont réexaminées. Des règles de suppression de compléments peuvent alors être appliquées. Par exemple, si une synapsie contenant un groupe adjectival à gauche n'apparaît qu'une seule fois, ce groupe adjectival est supprimé. Qu'il s'agisse des règles d'exploration des groupes nominaux ou bien des règles de suppression, ces règles ont un statut d'heuristiques: elles s'appuient sur un savoir lié à la lexicalisation des UNP.

Parmi les quatre catégories de termes produites par Termino, on peut choisir de prendre toutes les catégories ou une combinaison parmi celles-ci pour l'indexation. Les termes composés syntaxiques considérés dans cette expérience correspondent aux synapsies produites par Termino et toutes les combinaisons de catégories de termes ont été testées. Chaque combinaison est identifiée par un code de quatre lettres indiquant quelles catégories de termes sont utilisées. Les codes suivants servent à identifier les catégories:

*S*: synapsies
*A*: adjectifs
*V*: verbes
*N*: noms

Le code *SxVN* représente la combinaison des synapsies, verbes et adjectifs. Le *x* indique que cette catégorie a été omise. Il y a donc 15 combinaisons possibles en tout.

### 2.2. Méthode de pondération statistique

Il existe plusieurs méthodes pour calculer l'importance (ou *poids*) d'un terme simple dans un document. L'approche qui a été utilisée est celle décrite dans Salton et Buckley (1988). Dans cette

[4] Par un jeu de catégories, on qualifie le type de contexte dans lequel la synapsie a été repérée (étant donné le type de représentation que fournit ALSF, tous les contextes syntaxiques ne sont pas équivalents, ils sont entre autres plus ou moins déterminés). Ces catégories sont susceptibles d'être modifiées au cours du texte, chaque fois que l'on retrouve la même synapsie. Elles sont impliquées dans les règles de suppression à la toute fin de l'analyse.

approche, le poids d'un terme simple est déterminé à partir du produit de trois composantes: sa fréquence dans le document (*C*1), sa fréquence dans la collection de documents (*C*2) et un facteur de normalisation (*C*3). Le Tableau 3 présente les méthodes de calcul de ces trois composantes. Il existe donc 18 combinaisons possibles ($3 \times 3 \times 2$). Chaque combinaison sera représentée à l'aide de trois lettres: chaque lettre désignant le paramètre utilisé pour chacune des composantes. Ainsi, la méthode de calcul de poids *tfx* signifie que pour la composant *C*1 le paramètre *t* a été utilisé, que pour la composant *C*2 le paramètre *f* a été utilisé, et que pour la composante *C*3 le paramètre *x* a été utilisé.

**Tableau 3.** Composantes du poids d'un terme simple
(adapté de Salton et Buckley, 1988).

| | | |
|---|---|---|
| ***C1: fréquence du terme*** | | |
| *b* | 1 ou 0 | poids binaire égal à 1 si terme présent |
| *t* | $ft$ | fréquence du terme |
| *n* | $\frac{1}{2}+\frac{1}{2}\cdot\frac{ft}{\max ft}$ | fréquence normalisée augmentée |
| ***C2: fréquence dans la collection*** | | |
| *x* | 1 | aucun changement |
| *f* | $\log\frac{N}{fd}$ | fréquence dans la collection inverse |
| *p* | $\log\frac{N-fd}{fd}$ | fréquence dans la collection inverse probabiliste |
| ***C3: facteur de normalisation*** | | |
| *x* | 1 | aucune normalisation |
| *c* | $\frac{1}{\sqrt{\sum_{termes} p_i^2}}$ | normalisation |

*N*: nombre de documents, *fd*: nombre de documents auxquels le terme est associé
$p_i$: poids du terme non normalisé = $C1 \times C2$.

En ce qui concerne les termes composés, la détermination du poids est une question qui n'a pas encore de réponse précise. Fagan (1989) propose d'utiliser un poids qui est déterminé en fonction des poids de ses termes composants. Croft, Turtle et Lewis (1991) ont utilisé plusieurs approches: la moyenne, le produit et le maximum des poids des mots composant le terme composé. Ils ont

aussi considéré le terme composé comme un terme simple et appliqué les mêmes méthodes de pondération. C'est cette dernière approche que nous avons choisie.

## 2.2. Modèle de repérage

Le modèle vectoriel est utilisé pour notre expérience. Dans le modèle vectoriel chaque document *i* indexé est représenté à l'aide d'un vecteur de la forme $D_i = (p_{i1}, p_{i2}, \ldots, p_{it})$, où $p_{ik}$ représente le poids du terme *k* dans le document *i*, et *t* représente le nombre total de termes. Pour faire une recherche, l'utilisateur soumet une requête *j* en langue naturelle au système. Ce dernier analyse la requête et calcule le poids des termes de la requête. Il représente alors la requête à l'aide du vecteur $R_j = (r_{j1}, r_{j2}, \ldots, r_{jt})$. Les différentes méthodes de pondération décrites à la section précédente peuvent être appliquées non seulement aux documents mais aussi aux requêtes. Le code *D/bxx* représente le fait que la méthode de pondération *bxx* a été utilisée pour les documents et *R/bxx* pour les requêtes. Les poids des termes des requêtes sont calculés selon 6 méthodes au lieu des 18 possibles. En effet, dans le cas des requêtes, nous n'avons pas jugé nécessaire d'utiliser la composante de fréquence dans la collection (*f* et *p*) puisque dans la réalité les requêtes sont fournies une à une au système et non toutes ensembles comme c'est le cas dans les expériences. Il reste donc 6 combinaisons qui ont un *x* comme composante de fréquence dans la collection. On obtient en tout 15 combinaisons possibles pour le choix des catégories de termes, 18 combinaisons pour la pondération des termes dans les documents pour un total de 270 stratégies possibles pour l'indexation des documents et 6 combinaisons pour l'indexation des requêtes pour un grand total de 1 620 possibilités en tout. Toutes ces combinaisons ont été testées dans notre expérience.

Le système calcule un *coefficient de similarité* entre la requête et les documents. Il existe plusieurs formules pour calculer ce coefficient et celle qui a été choisie est la suivante (cosinus de l'angle entre les deux vecteurs):

$$sim(D_i, R_j) = \frac{\sum_{k=1}^{t} p_{ik} \cdot r_{jk}}{\sqrt{\sum_{k=1}^{t} p_{ik}^2 \cdot \sum_{k=1}^{t} r_{jk}^2}}$$

Après avoir calculé ce coefficient pour tous les documents, le système trie les documents par ordre décroissant par rapport à ce coefficient et présente à l'utilisateur cette liste triée. De cette façon, les premiers documents sont les plus similaires à la requête.

## 3. Expérimentation

L'expérience réalisée en est une du type "laboratoire" où un ensemble de requêtes portant sur un corpus test sont soumises au système qui produit un ordonnancement des documents selon le modèle de repérage choisi. Des mesures standards de rappel et de précision sont alors calculées pour évaluer les performances du système.

### 3.1. Évaluation

Les différentes stratégies d'indexation sont évaluées par la performance obtenue en rappel et en précision selon la tradition des expériences de cette nature. Le rappel mesure la proportion des documents pertinents extraits par rapport au total des documents pertinents dans le système et la précision mesure la proportion des documents pertinents par rapport au total des documents extraits. Pour utiliser ces deux mesures, il faut pouvoir déterminer les documents pertinents pour chaque requête. Il faut alors disposer de *jugements de pertinence* pour les requêtes qui donnent pour chaque requête la liste de documents pertinents. Ces jugements de pertinence sont généralement donnés par des personnes qui connaissent la collection de documents.

Puisqu'il y a deux mesures de performance (rappel et précision), et dans le but de comparer plusieurs stratégies d'indexation, le rappel est fixé à des valeurs déterminées (10%, 20%, …) et la précision est calculée pour chacun de ces niveaux de rappel (Raghavan, Jung & Bollmann, 1989). Après ce calcul, une stratégie *A* est dite meilleure qu'une stratégie *B* si, pour chaque niveau de rappel, la précision de la stratégie *A* est supérieure à celle de la stratégie *B*. Si ceci n'est pas vrai pour tous les niveaux de rappel, la moyenne de quelques précisions est calculée, et la comparaison se fait à l'aide de cette moyenne (Raghavan et al., 1989).

### 3.2. Collection de documents

Pour réaliser l'expérience et faire l'évaluation, il faut disposer d'une collection de documents en langue française ainsi que d'un ensemble de requêtes et de leurs jugements de pertinence. La collection utilisée dans notre expérimentation vient du *Guide de développement d'un système* du Groupe DMR Inc. Ces expériences ont été réalisées dans le cadre d'un projet de recherche majeur sur l'ingénierie des logiciels, le projet mobilisateur Le Macroscope Informatique géré par DMR et impliquant le Centre de Recherche Informatique de Montréal (CRIM) et d'autres partenaires industriels et académiques. L'application visée concerne les utilisateurs du guide qui sont appelés à faire des recherches dans le guide pour supporter leurs activités de développement de logiciels. Les requêtes choisies représentent l'expression de besoins typiques pour ce genre d'utilisateurs. Chaque document correspond à une sous-section du guide. Les caractéristiques des requêtes et de la collection utilisées pour notre expérience sont données dans le Tableau 4.

**Tableau 4.** Caractéristiques de la collection, des requêtes et des jugements de pertinence.

| | |
|---|---|
| ***Documents*** | |
| Nombre | 177 |
| Taille minimum (en mots) | 46 |
| Taille maximum (en mots) | 1472 |
| Taille moyenne (en mots) | 271,8 |
| ***Requêtes*** | |
| Nombre | 40 |
| Taille minimum (en mots) | 5 |
| Taille maximum (en mots) | 25 |
| Taille moyenne (en mots) | 12,1 |
| ***Jugements de pertinence*** | |
| Nombre minimum de documents par requête | 1 |
| Nombre maximum de documents par requête | 35 |
| Nombre moyen de documents par requête | 6,7 |

### 3.3. Résultats

Le programme de calcul des valeurs de rappel et précision a fourni 270 tableaux contenant les valeurs de rappel et précision pour les différentes stratégies possibles d'indexation des documents (15 possibilités pour le choix des catégories de termes et 18 possibilités pour la pondération). Le Tableau 5 est un exemple contenant les résultats pour la stratégie d'indexation *Sxxx-D/bxx* pour les documents et les 6 possibilités de calcul du poids des requêtes. La stratégie *Sxxx-D/bxx* correspond à la combinaison des synapsies seules comme catégorie de termes et de *bxx* pour la pondération des termes.

**Tableau 5.** Valeurs de rappel et précision pour *Sxxx–D/bxx*.

| Rappel | Précision | | | | | |
|---|---|---|---|---|---|---|
| | ***R/bxx*** | ***R/bxc*** | ***R/txx*** | ***R/txc*** | ***R/nxx*** | ***R/nxc*** |
| 10% | 60,4% | 60,4% | 60,4% | 60,4% | 60,4% | 60,4% |
| 20% | 53,0% | 53,0% | 53,0% | 53,0% | 53,0% | 53,0% |
| 30% | 44,4% | 44,4% | 44,4% | 44,4% | 44,4% | 44,4% |
| 40% | 40,0% | 40,0% | 40,0% | 40,0% | 40,0% | 40,0% |
| 50% | 36,6% | 36,6% | 36,6% | 36,6% | 36,6% | 36,6% |
| 60% | 27,0% | 27,0% | 27,0% | 27,0% | 27,0% | 27,0% |
| 70% | 22,7% | 22,7% | 22,7% | 22,7% | 22,7% | 22,7% |
| 80% | 18,8% | 18,8% | 18,8% | 18,8% | 18,8% | 18,8% |
| 90% | 13,8% | 13,8% | 13,8% | 13,8% | 13,8% | 13,8% |
| 100% | 13,8% | 13,8% | 13,8% | 13,8% | 13,8% | 13,8% |

Chaque cellule du tableau contient la précision pour un niveau de rappel et pour une méthode de calcul du poids des requêtes. Ainsi, pour un rappel de 70% et pour la méthode *R/bxx*, la précision est de 22,7%. L'étude des 270 tableaux montre que la précision varie peu avec la méthode de calcul du poids des requêtes. On peut donc négliger ce facteur dans l'analyse. Par exemple, dans le Tableau 5, toutes les précisions correspondant à un rappel donné sont égales. Ceci est dû au fait que, puisque les requêtes sont courtes, les mots n'apparaissent habituellement qu'une seule fois dans le texte de la requête. Donc, le fait d'utiliser le paramètre *t* pour la composante *C*1 au lieu du paramètre *b* ne change pas la valeur du poids (pour chaque terme on a $t = 1$). De même, l'utilisation du paramètre *n* équivaut à utiliser le paramètre *b* puisque $\max ft = 1$, et donc $n = 1 \ (= b)$. Finalement, la composante *C*3 est elle aussi toujours égale à 1 puisque $p_i = C1 \times C2 = 1 \times 1 = 1$. En conclusion, le poids d'un terme d'une requête est presque toujours égal à 1. De ce fait, une seule colonne pour la précision ne sera utilisée dans les 270 tableaux.

Dans les résultats de l'expérience, aucune stratégie n'est meilleure par rapport aux autres pour chaque niveau de rappel. Par exemple, la stratégie *SxVN–D/tfx* est celle qui a la plus grande précision pour le rappel de 10% (voir Tableau 6). Mais pour le rappel de 100% cette stratégie n'est plus la meilleure (par exemple la précision de la stratégie *SxxN–D/npx* à ce rappel est supérieure, voir Tableau 6).

**Tableau 6.** Valeurs de rappel et précision pour *SxVN–D/tfx* et *SxxN–D/npx*.

| Rappel | Précision | |
|---|---|---|
| | ***SxVN–D/tfx*** | ***SxxN–D/npx*** |
| 10% | 80,4% | 73,9% |
| 20% | 64,4% | 65,1% |
| 30% | 54,4% | 60,2% |
| 40% | 49,3% | 52,2% |
| 50% | 47,8% | 49,8% |
| 60% | 40,4% | 39,4% |
| 70% | 35,5% | 35,5% |
| 80% | 30,3% | 31,6% |
| 90% | 25,6% | 28,6% |
| 100% | 24,9% | 27,5% |

## 3.4 Analyse des moyennes des précisions

L'étude des moyennes des précisions montre que les meilleures stratégies sont *SAVN–D/npx* *SxxN–D/nfx* et *SxxN–D/nfc* avec 47,1% comme précision moyenne. Pour connaître l'impact de l'utilisation des synapsies, une première étude des résultats concernera le choix des catégories de termes sans tenir compte des méthodes de calcul du poids. Le Tableau 7 présente les moyennes des précisions des 15 combinaisons de catégories de termes (pour chaque combinaison la précision moyenne est celle de la meilleure stratégie utilisant cette méthode, et le rang est le classement de cette stratégie parmi les 270 existantes).

**Tableau 7.** Moyennes des précisions des méthodes de représentation de documents.

| Rang | Méthode de représentation | Précision moyenne |
|---|---|---|
| 1 | *SAVN* | 47,1% |
| 1 | *SxxN* | 47,1% |
| 4 | *SAxN* | 47,0% |
| 12 | *SxVN* | 45,7% |
| 45 | *xAVN* | 41,4% |
| 51 | *xAxN* | 40,2% |
| 55 | *SAxx* | 39,1% |
| 57 | *xxVN* | 39,0% |
| 69 | *xxxN* | 37,8% |
| 83 | *SAVx* | 36,6% |
| 94 | *Sxxx* | 35,8% |
| 135 | *SxVx* | 32,8% |
| 215 | *xAVx* | 17,5% |
| 221 | *xAxx* | 14,7% |
| 245 | *xxVx* | 12,4% |

Une conclusion est apparente à partir de ce tableau:

*Conclusion 1. Les combinaisons de catégories de termes incluant des synapsies (termes composés) sont meilleures que celles qui n'en utilisent pas toute méthode de pondération confondue.*

Ainsi, les deux meilleures stratégies utilisant des synapsies (*SAVN* et *SxxN*) ont comme précision moyenne 47,1%, alors que cette valeur est de 41,4% pour la meilleure stratégie utilisant seulement des termes simples (*xAVN*). Il y a donc une différence dans la précision moyenne de 5,7% lors de l'utilisation des synapsies. D'après Spark Jones (1974), cette différence est visible. En plus, l'amélioration de la précision moyenne est de 13,8%. Cette amélioration est calculée en effectuant la différence entre les deux précisions moyennes ($47{,}1 - 41{,}4 = 5{,}7$) et en divisant le résultat par la précision moyenne de la stratégie par rapport à laquelle est calculée cette amélioration (Keen, 1992).

Le Tableau 8 contient les différences et les améliorations dans la précision moyenne entre chaque stratégie utilisant des termes simples, et la strétégie correspondante utilisant en plus des synapsies. La conclusion suivante est apparente à partir de ce tableau:

*Conclusion 2. L'ajout de synapsies à une combinaison de catégories de termes utilisant seulement des termes simples augmente la précision moyenne.*

**Tableau 8.** Différences des précisions moyennes.

| **Méthode de représentation** | **Précision moyenne** | | **Différence** | **Amélioration** |
|---|---|---|---|---|
| | **sans synapsie** a | **avec synapsie** b | b - a | (b - a) / a |
| ·*AVN* | 41,4% | 47,1% | +5,7% | +13,8% |
| ·*xVN* | 39,0% | 45,7% | +6,7% | +17,2% |
| ·*AxN* | 40,2% | 47,0% | +6,8% | +16,9% |
| ·*xxN* | 37,8% | 47,1% | +9,3% | +24,6% |
| ·*AVx* | 17,5% | 36,6% | +19,1% | +109,1% |
| ·*xVx* | 12,4% | 32,8% | +20,4% | +164,5% |
| ·*Axx* | 14,7% | 39,1% | +24,4% | +166,0% |

Ces deux conclusions sont valables même lorsque la méthode de calcul du poids des documents est fixée. Par exemple, le Tableau 9 reprend les différences dans les précisions moyennes dans le cas de l'utilisation de la méthode de calcul du poids *D/tfx*.

**Tableau 9.** Différences des précisions moyennes pour *D/tfx*.

| **Méthode de représentation** | **Précision moyenne** | | **Différence** | **Amélioration** |
|---|---|---|---|---|
| | **sans synapsie** a | **avec synapsie** b | b - a | (b - a) / a |
| ·*AVN* | 41,4% | 47,0% | +5,6% | +13,5% |
| ·*xVN* | 39,0% | 45,3% | +6,3% | +16,2% |
| ·*AxN* | 40,2% | 47,0% | +6,8% | +16,9% |
| ·*xxN* | 37,8% | 46,4% | +8,6% | +22,8% |
| ·*AVx* | 17,2% | 36,6% | +19,4% | +112,8% |
| ·*xVx* | 10,0% | 32,7% | +22,7% | +227,0% |
| ·*Axx* | 14,7% | 38,5% | +23,8% | +161,9% |

### 3.5. Courbes de rappel/précision

Une comparaison plus fine des méthodes d'indexation avec et sans termes composés peut se faire en regardant les variations de précision à différents niveaux de rappel (Keen, 1992). Les Figures 2 à 7 représentent les courbes de rappel/précision pour diverses combinaisons. Chaque figure contient 2 courbes: la courbe de rappel/précision pour une des stratégies de représentation à base de termes simples, et la courbe de la même stratégie mais utilisant en plus des synapsies. Il apparaît ainsi de ces figures, qu'à tous les niveaux de rappels, les méthodes utilisant des synapsies sont plus performantes que celles qui n'en utilisent pas.

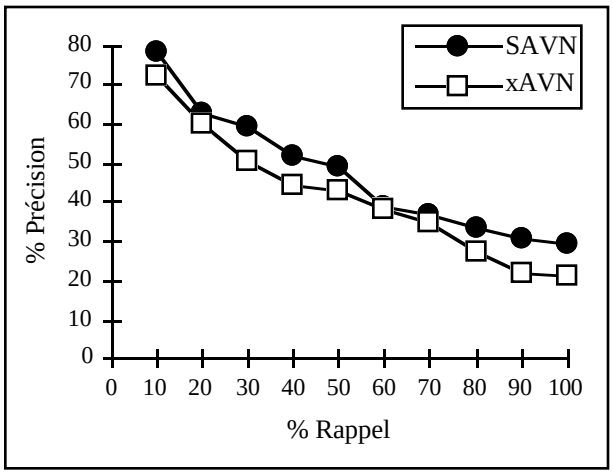



**Figure 2.** Rappel/précision pour *·AVN*.

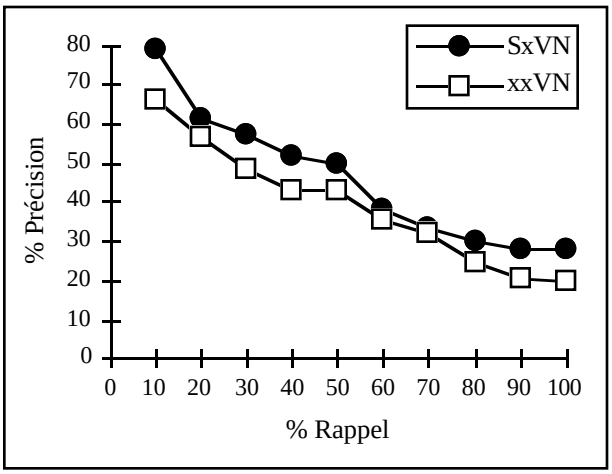



**Figure 3.** Rappel/précision pour *·xVN*.

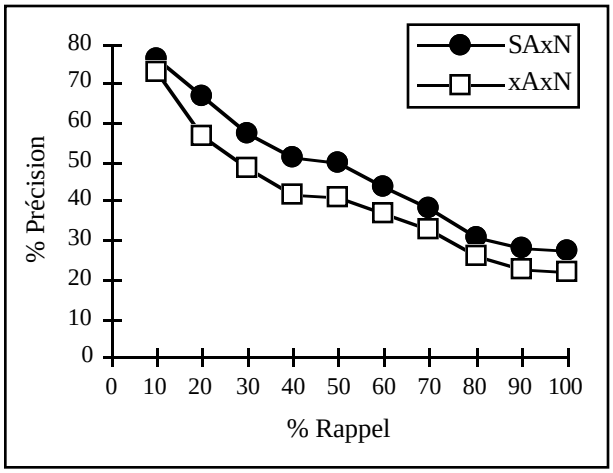



**Figure 4.** Rappel/précision pour *·AxN*.

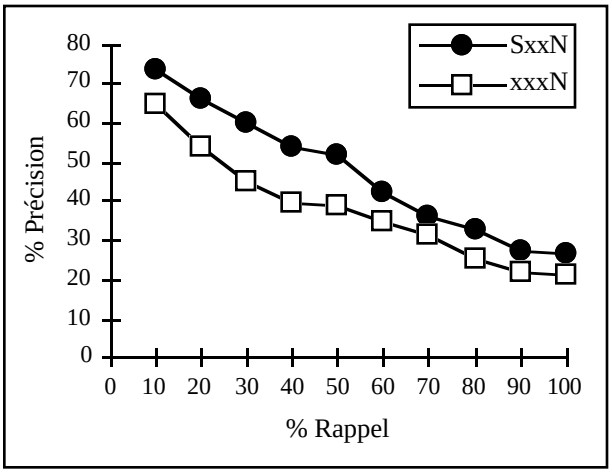



**Figure 5.** Rappel/précision pour *·xxN*.

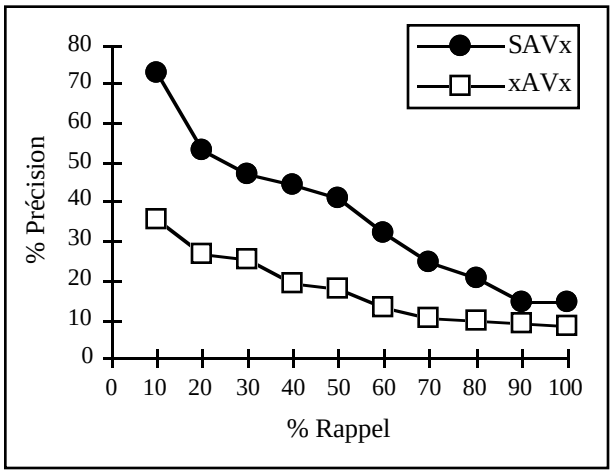



**Figure 6.** Rappel/précision pour *·AVx*.

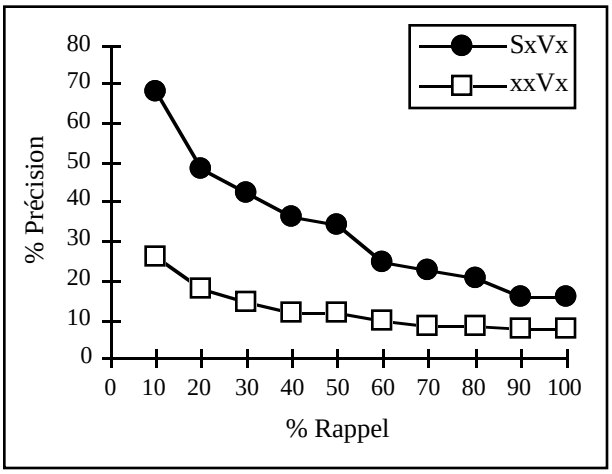



**Figure 7.** Rappel/précision pour *·xVx*.

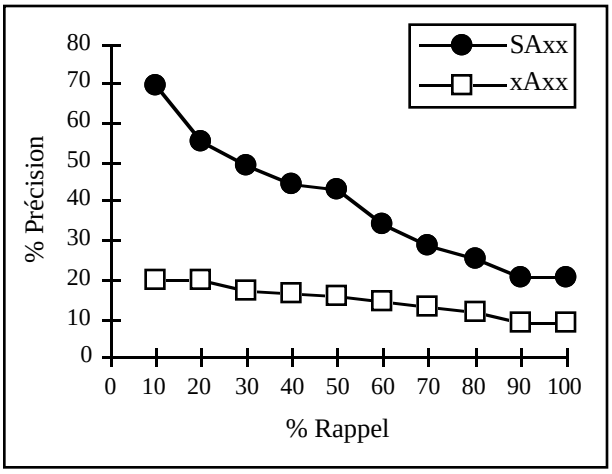



**Figure 8.** Rappel/précision pour ·*Axx*.

### 3.6. Discussion des résultats

Le résultat le plus apparent de cette expérience est la contribution importante des termes composés syntaxiques, ici représentés par les synapsies, à l'indexation automatique. La bonne performance des synapsies peut être due en partie au fait que le texte contient un vocabulaire technique dans un domaine en rapide évolution. Les nouveaux concepts sont souvent décrits par des termes composés étant donné qu'il n'y a pas encore de termes simples pour identifier ces nouveaux concepts. Le Tableau 10 contient quelques exemples de synapsies tirées de la collection de documents, qui représentent des concepts importants du domaine du développement de logiciels.

**Tableau 10.** Exemples de synapsies de la collection

| | |
|---|---|
| analyse des système[a] d'information | modèle de traitement |
| analyse préliminaire | modèle dynamique |
| architecture de système | phase du développement |
| complexité du système d'information | principe de la démarche |
| construction de modèle | proposition de projet |
| construction de modèle de traitement | représentation graphique |
| développement du système | scénario d'implantation |
| évaluation d'opportunité | stratégie d'affaire de l'organisation |
| hiérarchie de modèle | système d'information |
| modèle de donnée | technique de construction de modèle |

[a] les mots des synapsies sont lemmatisés.

D'autre part, les caractéristiques des requêtes peuvent aussi influencer ces résultats. Ainsi, la plupart des requêtes recherchent des concepts de la méthodologie. En conséquence, 35 requêtes parmi les 40 contiennent des synapsies.

## Conclusion

Comme conclusion générale, on peut retenir que, dans le cas de la présente collection de documents, l'utilisation de termes composés générés par analyse syntaxique produit une amélioration systématique de la performance par rapport aux termes simples. Ces résultats contrastent avec d'autres études du même genre sur la langue anglaise (Fagan, 1987). Plusieurs explications sont possibles. Un facteur à considérer est la plus grande richesse syntaxique de la langue française par rapport à la langue anglaise. Il y a donc un plus grand potentiel à exploiter lors de l'analyse syntaxique, ce qui peut produire une analyse de meilleure qualité qui se rèflète dans une meilleure qualité d'indexation. Il y a évidemment d'autres facteurs à considérer dont le domaine d'application, la collection particulière utilisée, etc. Le danger de généraliser les conclusions de telles expériences a été amplement mis en lumière dans la littérature (Robertson & Hancock-Beaulieu, 1992) et il faudra réaliser d'autres expériences utilisant diverses collections de documents afin d'augmenter notre confiance en ces résultats. D'autres variations dans la méthode

de pondération des termes composés sont aussi à considérer. Finalement, il serait intéressant de comparer les termes composés syntaxiques avec les approches statistiques dans le même esprit que les travaux de Salton, Buckley et Smith (1990).

## Remerciements

Ces travaux ont été supportés en partie par des subventions du Conseil de Recherche en Sciences Naturelles et en Génie du Canada (CRSNG), du Centre en Analyse de Texte (ATO) de l'Université du Québec à Montréal (UQAM) dans le cadre du projet ALEX et du Centre de Recherche en Informatique de Montréal (CRIM) dans le cadre du projet mobilisateur Le Macroscope Informatique géré par le Groupe DMR Inc.

## Références


Benveniste, É. (1974). Formes nouvelles de la composition nominale. In *Problèmes de linguistique générale*, (pp. 163-176). Paris: Gallimard.

Blosseville, M. J., Hébrail, G., Monteil, M. G. & Pénot, N. (1992). Automatic Document Classification: Natural Language Processing, Statistical Analysis, and Expert System Techniques used together. In *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, N. Belkin, P. Ingwerson, & A. M. Pejtersen (Ed.), Copenhagen, Denmark: ACM Press, pp. 51-59.

Corbin, D. (1991). Introduction. La formation des mots : structures et interprétations. *Lexique, Presses Universitaires de Lille*, **10**, 7-31.

Corbin, D. (1992). Hypothèses sur les frontières de la composition nominale. *Cahiers de grammaire*, **17**, 26-55.

Cori, M. & Marandin, J.-M. (1993). Grammaire d'arbres polychromes. *Traitement Automatique des Langues*, **34**(1), 101-132.

Croft, W. B., Turtle, H. R. & Lewis, D. D. (1991). The Use of Phrases and Structured Queries in Information Retrieval. In *Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, A. Bookstein, Y. Chiaramella, G. Salton, & V. V. Raghavan (Eds.), Chicago, Illinois: pp. 32-45.

David, S. (1993). *Les unités nominales polylexicales. Éléments de description et reconnaissance automatique*. Thèse de doctorat Thesis, Université Paris 7.

David, S. & Plante, P. (1990). De la Nécessité d'une Approche Morpho-Syntaxique en Analyse de Texte. *ICO*, **2**(3), 140-155.

Dumas, L. (1990). *LCMF : Lemmatisation et caractérisation morphologique du français*. RDLC, Centre ATO, UQAM.

Fagan, J. (1987). *Experiments in Automatic Phrase Indexing Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods*. Doctoral Dissertation Thesis, Cornell Univeristy.

Fagan, J. L. (1989). The Effectiveness of a Nonsyntactic Approach to Automatic Phrase Indexing for Document Retrieval. *JASIS*, **40**(2), 115-132.

Keen, E. M. (1992). Presenting Results of Experimental Retrieval Comparisons. *Information Processing & Management*, **28**(4), 491–502.

Lewis, D. D. (1992). An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. In *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, N. Belkin, P. Ingwerson, & A. M. Pejtersen (Eds.), Copenhagen, Denmark: ACM Press, pp. 37-50.

Marandin, J.-M. (1992). La perception syntaxique. *Le gré des langues*, **4**, 64-91.

Milner, J.-C. (1989). *Introduction à une science du langage*. Paris: Seuil.

Plante, P. (1993). *FX, La programmation en faisceaux ; Version 5.1*. Centre ATO, UQAM.

Raghavan, V. V., Jung, G. S. & Bollmann, P. (1989). A Critical Investigation of Recall and Precision as Measures of Retrieval System Performance. *ACM TOIS*, **7**(3), 205-229.

Robertson, S. E. & Hancock-Beaulieu (1992). On the Evaluation of IR Systems. *Information Processing and Management*, **28**(4), 457-466.

Salton, G. (1986). Another Look at Automatic Text-Retrieval Systems. *Communications ACM*, **29**(7), 648-656.

Salton, G. & Buckley, C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing and Management*, **24**(5), 513-523.

Salton, G., Buckley, C. & Smith, M. (1990). On the Application of Syntactic Methodologies in Automatic Text Analysis. *Information Processing & Management*, **26**(1), 73-92.

Spark Jones, K. (1974). Automatic Indexing. *Journal of Documentation*, **30**(4), 393–432.