Intégrer plus de connaissances linguistiques en recherche d'information peut-il augmenter les performances des systèmes ?

Fabienne Moreau* — **Vincent Claveau*** — **Pascale Sébillot***

** IRISA*
Campus universitaire de Beaulieu
35042 Rennes cedex, France
{Fabienne.Moreau, Vincent.Claveau, Pascale.Sebillot@irisa.fr}

RÉSUMÉ. *Cet article pose la question de l'intérêt en RI de la combinaison au sein d'un même système de plusieurs informations linguistiques de types divers (appartenant aux niveaux morphologique, syntaxique et sémantique de la langue). Son objectif est d'évaluer dans un cadre unifié, par le biais d'une plate-forme conçue pour intégrer de multiples connaissances linguistiques au sein d'un même SRI, l'impact respectif de ces diverses informations et d'analyser les relations susceptibles d'exister entre elles : sont-elles complémentaires ou redondantes pour retrouver les documents pertinents ? Il montre en particulier l'intérêt des connaissances morphologiques et sémantiques et de leur association.*

ABSTRACT. *This paper investigates the potential added-value of a combination of different kinds of linguistic information (i.e. information that belongs to morphological, syntactic and semantic levels of language). In particular, it aims at determining whether those various kinds of knowledge, when integrated within a single information retrieval system, have separately the same impact on its overall performance, or whether some degree of correlation exists between them, therefore evaluating whether they are either complementary or redundant for finding relevant documents. The interest of morphological and semantic information, and their combinations, stands out from the described experiments.*

MOTS-CLÉS : *recherche d'information, traitement automatique des langues, couplage d'informations morphologiques, syntaxiques et sémantiques, analyse de corrélations*

KEYWORDS: *information retrieval, natural language processing, combination of morphological, syntactical and semantic informations, correlation analysis*

1. Introduction[1]

La principale difficulté d'un système de recherche d'information (SRI) textuelle est d'établir une correspondance entre l'information demandée par un utilisateur (par le biais d'une requête) et celle contenue dans les documents. Pour y parvenir, la méthode généralement utilisée repose sur un appariement des mots utilisés dans la requête avec ceux représentant le contenu des documents. Compte tenu de ce mécanisme fondé sur une simple comparaison de chaînes de caractères, les SRI se retrouvent rapidement confrontés à deux problèmes liés à la complexité du langage naturel. Le premier est la polysémie : un même terme peut avoir plusieurs sens et renvoyer à des concepts différents, ce qui entraîne potentiellement la récupération par le SRI de documents non pertinents. Le second problème, dual du premier, concerne la possibilité offerte par le langage naturel de formuler de différentes manières un même concept. Un document pertinent peut ainsi contenir des termes différents de ceux de la requête mais « sémantiquement » proches.

Une solution pour faire face à ces deux difficultés est de se tourner vers le traitement automatique des langues (TAL) et les informations linguistiques qu'il permet d'extraire des documents et requêtes. L'objectif de l'introduction de connaissances linguistiques en recherche d'information (RI) est de disposer de descripteurs plus robustes et plus pertinents que de simples chaînes de caractères. De nombreuses expériences d'intégration de telles connaissances en RI ont donc été menées, dont (Moreau, 2006) présente un état de l'art détaillé. La majorité de ces études cherche à évaluer l'apport d'un seul type d'information linguistique appartenant à l'un des trois niveaux d'analyse linguistique suivants : morphologique (qui s'intéresse à la forme des mots), syntaxique (qui concerne la façon dont les mots s'articulent entre eux pour former des syntagmes ou des phrases) ou sémantique (qui étudie le sens des mots). La prise en compte d'informations morphologiques en RI, généralement par racinisation ou lemmatisation (Gaussier *et al.*, 2000; Savoy, 2002, *inter alia*), permet aux systèmes de reconnaître au sein des documents et requêtes les différentes formes d'un même mot et de pouvoir les apparier (*e.g.* une requête avec le terme *cheval* pourra être mise en correspondance avec un document contenant *chevaux*). Les informations d'ordre syntaxique (*e.g.* les termes complexes (Fagan, 1987; Gaussier *et al.*, 2000, *inter alia*) ou d'autres types de syntagmes (Strzalkowski *et al.*, 1996; Haddad, 2002, *inter alia*)) présentent l'avantage d'exploiter les relations et dépendances qu'entretiennent les termes entre eux et de dépasser ainsi les faiblesses d'une représentation classique en « sacs de mots » des documents et requêtes. Enfin, l'intégration de connaissances sémantiques au sein d'un SRI peut contribuer à améliorer ses performances en cherchant, par exemple, à associer à chaque terme des documents et requêtes un ensemble de sens non ambigus (*cf.* les travaux portant sur la désambiguïsation automatique en RI (Sanderson, 1997; Kilgarriff *et al.*, 2000)), ou à ajouter des termes sémantiquement liés aux mots initialement présents dans les requêtes (Claveau *et al.*, 2004; Voorhees, 1998, *inter alia*). Quelques rares travaux ont intégré des informations appartenant à plusieurs niveaux de la langue (morphologique, syntaxique et sémantique).

1. Les travaux présentés dans cet article ont été réalisés avec le soutien de l'INRIA.

Parmi ceux-ci, on peut citer la combinaison de connaissances du même niveau (Zhai *et al.*, 1997) ou celle de deux niveaux (Friburger *et al.*, 2002; Haddad, 2002, *inter alia*). Enfin Strzalkowski *et al.* (1996) proposent un système de descripteurs en parallèle où chaque descripteur reflète une stratégie particulière de représentation linguistique des documents et requêtes, et l'expérimentent sur trois descripteurs provenant de trois niveaux de langue distincts : les documents et requêtes sont représentés à l'aide d'informations morphologiques (racinisation), syntaxiques (sous la forme de syntagmes normalisés) et sémantiques (par le biais d'entités nommées). Ces trois descripteurs sont tout d'abord utilisés individuellement dans le processus de recherche, et le classement final des documents est obtenu par fusion des résultats.

Malgré le nombre et la diversité des pistes explorées, l'impact réel de l'insertion de connaissances linguistiques dans des SRI est difficile à évaluer, les résultats obtenus, tant par les systèmes très majoritaires intégrant un unique type de connaissance linguistique que par les quelques SRI qui en combinent plusieurs, étant mitigés. Les améliorations obtenues sont en effet souvent modestes (Strzalkowski *et al.*, 1996) et les études aboutissent fréquemment à des conclusions contradictoires (Moreau, 2006). De plus, les expériences réalisées sont très peu aisées à comparer (différence de langue, de collection de test, de technique d'évaluation...) et leurs résultats sont tributaires d'un nombre important de paramètres (richesse morphologique de la langue, longueur des requêtes...).

Par conséquent, notre objectif, au sein de cet article, est double : d'une part, nous voulons bâtir un cadre homogène pour mesurer l'apport de multiples connaissances linguistiques provenant des trois niveaux de langue, et évaluer dans ce cadre unifié et sur des données identiques l'efficacité respective de chacune d'entre elles pour retrouver des documents pertinents. D'autre part – et ceci constitue le point central de cet article –, nous voulons poser la question de la pertinence de la combinaison de plusieurs informations au sein d'un même SRI. S'il peut paraître imaginable, bien que non établi jusqu'à présent, que des informations de types divers puissent amener des connaissances différentes et pallier conjointement les faiblesses des SRI, il convient tout d'abord de répondre, avant même de chercher une méthode de combinaison, à un certain nombre de questions fondamentales : quels résultats peut-on attendre d'une éventuelle combinaison ? Les informations sont-elles complémentaires ou redondantes ? Les éventuels gains de performances sont-ils susceptibles de s'additionner ?... Pour répondre à ce double objectif, nous avons d'une part construit une plate-forme permettant d'intégrer en parallèle dans un SRI de multiples connaissances de différents niveaux de langue, autorisant ainsi la mesure de l'impact de chacune d'elles. Cette architecture nous offre également un cadre pour réaliser une analyse originale des corrélations entre ces diverses informations du point de vue de leur efficacité à accroître les performances, et proposer une méthodologie répétable pour évaluer la pertinence de combiner des connaissances linguistiques au sein d'un même système.

Après avoir décrit en section 2 l'architecture de la plate-forme d'intégration au sein d'un SRI de connaissances linguistiques de trois niveaux de langue, nous présentons en section 3 l'évaluation, dans ce cadre unifié, de l'impact individuel de ces diverses

informations. La section 4 est dédiée à l'analyse des relations susceptibles d'exister entre elles, afin de donner des indications sur la pertinence de leur couplage en RI.

2. Architecture expérimentale

Nous présentons dans cette section l'architecture mise en œuvre pour intégrer diverses informations linguistiques de natures variées au sein d'un même SRI. La plateforme proposée s'inspire de celle réalisée par (Strzalkowski *et al.*, 1996) qui, comme évoqué en introduction, permet de représenter au sein d'un SRI de manière simultanée plusieurs informations linguistiques sous la forme de descripteurs mis en parallèle. Notre système se distingue cependant des travaux de Strzalkowski *et al.* d'une part par la nature et la diversité des informations linguistiques prises en compte et, d'autre part, par sa finalité. Notre objectif n'est pas de concevoir directement un système qui incorpore une multitude de descripteurs linguistiques mais d'étudier comment ces derniers se comportent les uns par rapport aux autres dans la recherche de documents pertinents et d'évaluer l'intérêt d'un tel couplage. Après avoir présenté les informations linguistiques appartenant à la fois aux niveaux morphologique, syntaxique et sémantique de la langue que nous avons choisi d'exploiter ici, nous décrivons l'architecture mise en place pour leur intégration au sein d'un SRI puis la collection de test sur laquelle nous nous appuyons pour nos expérimentations.

2.1. *Informations linguistiques prises en compte*

Pour sélectionner les informations linguistiques qui seront ensuite combinées pour enrichir la représentation textuelle des documents et requêtes, nous nous sommes imposés deux contraintes fortes. La première concerne la volonté de recourir à des informations linguistiques « standards », *i.e.* traditionnellement exploitées en RI. Il ne s'agit donc pas de chercher à exploiter de nouvelles méthodes complexes d'acquisition d'informations linguistiques, mais plutôt d'utiliser des connaissances facilement extractibles à l'aide d'outils disponibles et communément employés en RI. La seconde contrainte est directement liée à la collection de documents et requêtes utilisée pour nos expérimentations : les outils doivent être capables de manipuler des volumes de données assez importants et en anglais.

Nous présentons successivement les informations linguistiques de nature morphologique, syntaxique et sémantique respectant les deux contraintes énoncées ci-dessus. L'ensemble des traitement évoqués ici est appliqué sur les documents et requêtes directement issus de la collection. Tous les mots des textes et questions sont pris en compte et considérés comme des termes d'indexation potentiels. Leur pondération et la suppression des mots vides sont effectuées ultérieurement (lors de l'intégration des représentations linguistiques au sein du SRI).

Pour les connaissances d'ordre morphologique, nous avons choisi tout d'abord des informations d'ordre flexionnel. Un traitement de lemmatisation est appliqué à tous les documents et requêtes avec l'outil TREETAGGER (disponible à http ://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger). Il permet d'identifier, pour chaque mot des textes et requêtes, son lemme (*i.e* sa forme de base débarrassée de ses flexions). Nous prenons

également en compte des connaissances d'ordre flexionnel et dérivationnel. Pour cela, nous appliquons aux textes (et questions) une procédure de racinisation (*stemming*), en nous appuyant sur l'algorithme de Porter (Porter, 1980), qui permet d'extraire pour chaque mot sa pseudo-racine (*stem*). Enfin, des informations d'ordre morpho-syntaxique sont également exploitées. Une analyse morpho-syntaxique des documents et requêtes est réalisée — à l'aide de TREETAGGER — dans le but d'associer à chaque mot sa catégorie grammaticale (nom, verbe, adjectif...).

Pour le niveau syntaxique de la langue, nous avons retenu plusieurs structures permettant de rendre compte des relations et dépendances entre les termes. Ainsi, la reconnaissance des termes complexes et de leurs variantes est effectuée à l'aide de l'outil FASTR (Jacquemin, 1994). Nous utilisons également des bigrammes et trigrammes. L'outil employé (NGRAM STATISTIC PACKAGE (http ://www.d.umn.edu/~tpederse/nsp.html) repère et extrait les suites de n termes (dans notre cas $n = 2$ (bigrammes) ou 3 (trigrammes)) qui apparaissent de manière statistiquement significative dans la collection. Nous proposons enfin d'exploiter les syntagmes nominaux présents dans les textes et les requêtes. Pour leur extraction, nous utilisons l'outil développé par Ramshaw et Marcus (Ramshaw *et al.*, 1995).

Enfin, le niveau sémantique est représenté par le biais de connaissances acquises pour la plupart à l'aide de ressources sémantiques (WORDNET). Nous procédons tout d'abord à un traitement simple de désambiguïsation automatique des termes de la collection à l'aide du module proposé par Pedersen *et al.* (Pedersen *et al.*, 2004). Nous récupérons donc, pour chaque terme désambiguïsé, son étiquette sémantique (numéro de sens dans WORDNET), l'ensemble de ses synonymes, et l'ensemble des mots reliés (relations intercatégorielles) à ce terme par un lien de morphologie dérivationnelle. Par exemple, le nom *adoption* est lié dans WORDNET au cinquième sens du verbe *adopt#v#5* qui a lui-même pour synonyme le verbe *take in*. Nous formons donc à partir de ces informations la famille de mots {*adoption, adopt, take in*}. La dernière information sémantique prise en compte correspond aux noms propres. Après avoir évalué plusieurs outils disponibles pour leur extraction, nous utilisons une méthode plus basique mais évaluée plus performante qui consiste à s'appuyer sur l'étiquetage morpho-syntaxique des textes (à l'aide de TREETAGGER) et à extraire tous les mots associés à l'étiquette *nom propre*.

Chaque type de connaissance linguistique extraite d'un document correspond à un descripteur (ou index). Ainsi, à la suite de ces divers traitements linguistiques, un document (ou une requête) de la collection peut donc être représenté par 11 descripteurs différents. Il peut être vu comme un ensemble de lemmes, de racines, de termes simples associés à leur étiquette grammaticale, de termes complexes, de bigrammes, de trigrammes, de groupes nominaux, de noms propres, de termes simples associés à leurs étiquettes sémantiques, de termes simples associés à un groupe de synonymes ou encore à un groupe de mots reliés morphologiquement. Nous proposons également de représenter les documents (et requêtes) à l'aide d'un index « standard » (représentant le $12^{ème}$ type de descripteurs). Nous extrayons pour cela l'ensemble des termes simples qu'ils contiennent. Nous décrivons à présent la façon dont ces multiples représentations peuvent être combinées et intégrées au sein d'un même SRI.

2.2. *Architecture d'intégration au sein d'un SRI*

Les mécanismes traditionnels de RI sont généralement conçus pour manipuler un type de représentation des documents et requêtes (un descripteur par document) à la fois. Pour pouvoir intégrer les différentes informations linguistiques au sein du même SRI, et ainsi évaluer la contribution respective de chaque information et étudier les relations qu'elles entretiennent entre elles, nous nous appuyons, comme nous l'avons dit précédemment, sur une architecture semblable à celle de Strzalkowski *et al.* (1996). Il s'agit donc de concevoir un système de descripteurs en parallèle, où chaque index reflète une représentation linguistique particulière des documents et requêtes.

De manière plus précise, l'architecture proposée peut être synthétisée de la façon suivante (*cf.* figure 1) : les documents et requêtes passent tout d'abord par un module d'analyse linguistique (combinant les différents outils décrits précédemment) qui permet d'obtenir 12 représentations différentes d'un même document (ou requête).

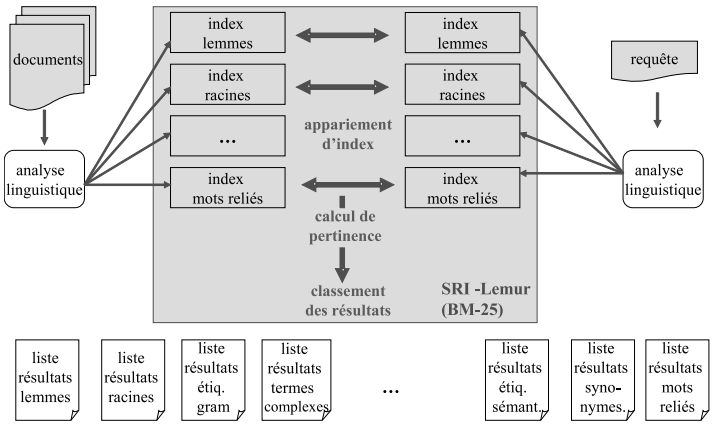

Figure 1. *Intégration au sein du SRI des représentations multi-index*

Ces représentations sont ensuite intégrées de manière parallèle au sein du SRI. Le SRI utilisé est LEMUR (http ://www.lemurproject.org/), configuré de manière à fonctionner comme le système Okapi (BM-25) (Robertson *et al.*, 1998). Il compare (pour les 12 index pris en compte) chacune des représentations des documents à celles correspondantes des requêtes et calcule un score de similarité. À la suite de cette phase d'appariement, nous obtenons finalement 12 listes ordonnées de résultats.

2.3. *Collection de test*

Les documents et les requêtes sur lesquels nous nous appuyons pour nos évaluations présentées en sections suivantes proviennent de la collection TIPSTER utilisée lors des campagnes d'évaluation TREC. Pour nos expérimentations, nous utilisons une sous-partie de ce corpus qui regroupe environ 175000 articles de journaux issus

du *Wall Street Journal* des années 1986 à 1992, de longueurs et thématiques variées. Le jeu de 50 requêtes (champs *titre* et *description*) et les jugements de pertinence associés sur lesquels nous nous appuyons correspondent à ceux utilisés lors de la campagne TREC-3 de 1994.

La plate-forme proposée nous permet donc de représenter les documents et les requêtes à l'aide de divers descripteurs linguistiques – 12 descripteurs dans les expériences présentées ici. Puisque ces descripteurs sont obtenus à partir d'une même collection, nous pouvons à présent répondre à notre premier objectif : évaluer leur impact respectif dans un cadre homogène.

3. Impact individuel de chaque information linguistique sur les performances

Comme évoqué en introduction, les diverses connaissances linguistiques que nous avons choisies de prendre en compte ont déjà été exploitées en RI. Leur apport a cependant toujours été évalué sur des données différentes et les résultats obtenus sont par conséquent difficilement comparables. L'objectif de cette section est donc de chercher à mesurer leur impact respectif dans un cadre unifié et sur des données identiques.

À partir des 12 listes de résultats produites pour chaque requête par le biais de notre architecture de test, nous pouvons estimer les performances du SRI pour chaque type de représentation considéré en utilisant les mesures d'évaluation communément utilisées en RI. Les résultats présentés dans la tableau 1 indiquent les performances classiquement mesurées en termes de précision et rappel sur les n premiers documents trouvés ($P(n)$ et $R(n)$ par la suite) et de précision moyenne non interpolée (MAP) obtenues par le SRI pour chacun des index exploités sur les 50 requêtes de notre jeu de test. Les performances d'un SRI « standard », n'exploitant aucune information linguistique, correspondent dans le tableau aux résultats obtenus par l'index des termes simples. Il nous sert de référence pour évaluer l'apport des index « linguistiques ». L'amélioration (en %) des performances des divers index linguistiques obtenue par rapport à cet index de référence est indiquée entre parenthèses.

Il ressort tout d'abord de ces résultats que, contrairement à ce qui est parfois observé dans les travaux de l'état de l'art, la prise en compte d'informations morphologiques (lemmes ou racines), même exploitées simplement, s'avère intéressante en RI. L'exploitation de certaines connaissances sémantiques et plus particulièrement d'informations de synonymie a également un impact positif sur les performances. Là encore, ces connaissances sont obtenues avec des méthodes qui ne sont que partiellement efficaces, laissant penser qu'en recourant à des informations de synonymie plus fines, les performances pourraient être améliorées. Enfin, les résultats les plus décevants sont ceux obtenus en intégrant au sein du SRI des informations de nature syntaxique (bigrammes, termes complexes, groupes nominaux et trigrammes). Parmi les différentes explications possibles à ces faibles résultats, nous privilégions l'idée qu'en considérant les documents et requêtes uniquement à partir des informations syntaxiques qu'ils contiennent, on entraîne une sous-représentation de leur contenu textuel.

Le fait que certaines informations linguistiques, et plus particulièrement les informations d'ordre (morpho-)syntaxique ne produisent pas de meilleurs résultats n'est pas surprenant. Nous avons volontairement adopté une représentation simpliste des

	MAP	P(20)	P(100)	P(1000)	R(20)	R(100)	R(1000)
termes simples	25.53	42.10	23.70	4.63	20.76	45.48	71.51
racines	30.54 **(+19.62%)**	44.90 **(+6.65%)**	25.36 **(+7.01%)**	5.40 **(+16.63%)**	22.01 **(+6.02%)**	46.22 **(+1.62%)**	78.74 **(+10.11%)**
lemmes	28.12 **(+10.14%)**	41.90 (-0.47%)	25.46 **(+7.42%)**	5.42 **(+17.06%)**	20.24 (-2.50%)	46.98 **(+3.29%)**	77.79 **(+8.78%)**
termes simples + synonymes	26.67 **(+4.46%)**	39.60 (-5.93%)	23.86 **(+0.67%)**	5.29 **(+14.25%)**	19.59 (-5.63%)	43.15 (-5.12%)	74.97 **(+4.83%)**
termes simples + étiq. sémant.	24.42 (-4.34%)	39.40 (-6.41%)	22.88 (-3.45%)	4.64 **(+0.21%)**	19.05 (-8.23%)	42.32 (-6.94%)	69.69 (-2.54%)
termes simples + mots reliés	22.18 (-13.12%)	35.60 (-15.43%)	21.68 (-8.52%)	4.69 **(+1.29%)**	17.10 (-17.63%)	39.23 (-13.74%)	67.64 (-5.41%)
termes simples + étiq. gram.	17.80 (-30.27%)	29.80 (-29.21%)	16.86 (-28.86%)	3.41 (-26.34%)	13.36 (-35.64%)	32.35 (-28.86%)	54.71 (-23.49%)
bigrammes	10.65 (-58.28%)	17.76 (-57.81%)	13.20 (-44.30%)	2.17 -53.13%)	8.41 (-59.48%)	22.67 (-50.15%)	37.68 (-47.30%)
termes complexes	7.63 (-70.11%)	17.23 (-59.07%)	9.83 (-58.52%)	1.42 (-69.33%)	6.37 (-69.31%)	15.66 (-65.56%)	23.97 (-66.48%)
groupes nominaux	6.39 (-74.97%)	16.67 (-60.40%)	7.31 (-69.15%)	0.81 (-82.50%)	7.24 (-65.12%)	13.44 (-70.44%)	16.29 (-77.21%)
noms propres	5.36 (-79.01%)	8.16 (-80.61%)	5.24 (-77.89%)	1.68 (-63.71%)	5.16 (-75.14%)	11.40 (-74.93%)	26.51 (-62.92%)
trigrammes	3.02 (-88.17%)	15.45 (-63.30%)	4.05 (-82.91%)	0.43 (-90.71%)	5.40 (-73.98%)	8.18 (-82.01%)	8.92 (-87.52%)

Tableau 1. *Performances du SRI pour chaque information linguistique manipulée*

documents pour pouvoir évaluer l'apport intrinsèque de chaque type de représentation et cela pénalise évidemment certaines représentations très spécialisées comme les trigrammmes ou les noms propres. Ce n'est pas non plus problématique dans l'optique de leur couplage à condition que chaque index, quelles que soient ses performances, retrouve de façon complémentaire des documents pertinents. Dans le cas contraire où ces informations ne permettent pas au SRI de récupérer des documents différents de ceux retrouvés par un index de racines par exemple, il apparaîtrait inutile de les exploiter en RI. Cette idée justifie donc l'étude des relations entre les divers index. Cette étude fait l'objet de la section suivante.

4. Analyse des corrélations entre les informations linguistiques

Afin d'évaluer l'intérêt de coupler diverses informations linguistiques au sein d'un SRI, nous nous intéressons à l'étude des liens susceptibles d'exister entre ces informations, en cherchant à établir si ces dernières sont complémentaires ou au contraire redondantes. Pour cela, nous procédons en section 4.1 à une analyse approfondie des corrélations entre les différentes listes de documents retournées par le SRI. Pour mettre

en perspective la façon dont elles sont liées les unes aux autres, nous proposons ensuite d'établir en section 4.2 des classes d'informations linguistiques construites en fonction de leur comportement sur les performances[2].

4.1. *Analyse des corrélations entre listes de documents pertinents*

Cette série d'expérimentations se déroule en trois étapes. La première expérience (section 4.1.1) vise à analyser les relations entre les différents index uniquement à partir de leur capacité à retrouver (ou non) les documents pertinents. En s'appuyant sur la liste des documents qui devraient idéalement être ramenés par le SRI pour une requête donnée, il s'agit d'examiner si deux index ont un comportement similaire ou différent pour retrouver ou ne pas retrouver les documents de cette liste. À la suite des résultats obtenus, deux points nécessitent d'être approfondis par le biais d'expériences complémentaires. Le premier (*cf.* section 4.1.2) concerne les cas où les index semblent se comporter de manière similaire pour retrouver (ou non) les mêmes documents pertinents (*i.e.* les informations linguistiques semblent redondantes). Nous souhaitons déterminer si les informations linguistiques sont liées parce qu'elles permettent de retourner toutes deux les documents pertinents ou au contraire parce qu'elles ne les retrouvent pas. Le second point (*cf.* section 4.1.3) s'intéresse aux cas où les index ont un comportement différent dans la récupération des documents pertinents (ils sont considérés alors comme complémentaires). Il s'agit alors de vérifier la nature exacte de cette complémentarité.

4.1.1. *Analyse des similarités entre index*

Cette expérience s'appuie, pour analyser les relations entre les diverses informations linguistiques prises en compte, uniquement sur leur capacité à retrouver ou non, par le biais du SRI, les documents pertinents de la collection pour une requête donnée. Pour cela, nous proposons d'utiliser une méthode binaire simple qui consiste à évaluer si, à partir de la liste des documents pertinents pour chaque requête fournie avec la collection TIPSTER, les index (évalués par paires) permettent au SRI de retrouver ou non ces documents pertinents. Pour une requête donnée, un document est jugé retrouvé par un index s'il est proposé dans les 20, 100 ou 1000 premières réponses selon les expériences réalisées.

À partir de la liste des documents jugés *a priori* pertinents pour une requête donnée et pour toutes les paires d'informations linguistiques prises en compte, nous notons 1 si le premier index (nommé $index_1$) de la paire (*resp.* le second index de la paire noté $index_2$) retrouve le document pertinent et 0 sinon. Nous procédons de cette façon pour l'ensemble des documents de la liste, des requêtes et des paires d'index.

Nous cherchons ensuite à évaluer, pour chaque paire d'informations linguistiques étudiée, la similarité entre les résultats obtenus par $index_1$ et $index_2$. Autrement dit, nous souhaitons savoir si les documents pertinents retrouvés par le premier index sont

2. Nous proposons ici une synthèse des principales conclusions qui ressortent des diverses expériences réalisées. Pour une description plus détaillée de ces études, et pour d'autres analyses plus classiques de corrélation entre index sur le même jeu de données, se reporter à (Moreau, 2006).

les mêmes que ceux retournés par le second, et réciproquement. Pour juger de la ressemblance de ces deux listes, nous utilisons l'indice de similarité suivant :

$$sim = 1 - \frac{\sum_{i=1}^{n} (index1_i - index2_i)^2}{n}$$

où n est le nombre de paires de valeurs analysées, et où $index1_i$ et $index2_i$ représentent les valeurs (dans notre cas binaires) retournées par l'$index_1$ et l'$index_2$ pour un document pertinent i. Cet indice de similarité, construit à partir la distance euclidienne normalisée des deux index, permet de mesurer le taux de réponses identiques entre les deux index. Un taux proche de 100% signifie que les deux index d'une paire retrouvent (ou ne retrouvent pas), par le biais du SRI, les mêmes documents pertinents. Leurs listes de résultats respectives sont donc similaires. Un taux proche de 0 implique que les deux index sont complémentaires puisque les documents pertinents retrouvés par l'un ne sont pas retournés par l'autre (et inversement). La figure 2 (colonne 2, cas a, b ou c selon le nombre de premières réponses prises en compte) représente la moyenne (pour les 50 requêtes) des taux de réponses identiques obtenues pour chaque paire d'index[3].

L'analyse des différents taux obtenus montre que les paires d'index qui produisent des résultats très similaires (dont le taux de similarité est supérieur à 80-90% selon les expériences) sont celles qui combinent des informations de même niveau de langue, *i.e.* des connaissances uniquement d'ordre syntaxique (*e.g.* les paires *groupes nominaux-trigrammes*, *bigrammes-termes complexes...*), morphologique (*e.g.* le couple *lemmes-racines* qui retourne des résultats identiques dans plus de 90% des cas pour les 3 expériences réalisées) ou sémantique (*e.g.* les paires *synonymes-mots reliés*, *étiquettes sémantiques-synonymes...*). La combinaison d'informations morphologiques et sémantiques (*e.g.* les couples *lemmes-synonymes*, *racines-mots reliés*, *lemmes-étiquettes sémantiques*) semble également produire des résultats assez identiques puisque le taux de similarité des listes de résultats de ces paires d'index avoisine les 85%. Parmi les taux de similarité les plus bas, on trouve essentiellement des couples qui combinent des informations syntaxiques avec des connaissances morphologiques (*e.g.* les paires *lemmes-trigrammes*, *racines-groupes nominaux...*) ou bien avec des informations sémantiques (*e.g.* les couples *trigrammes-mots reliés*, *groupes nominaux-synonymes...*).

Au vu de cette première expérience, il semblerait donc plus intéressant en RI de combiner ce dernier type de connaissances (*i.e.* des informations morphologiques ou sémantiques avec des informations syntaxiques), qui semblent avoir un comportement complémentaire pour retrouver (ou non) les documents pertinents, que de coupler des informations d'un seul niveau de langue ou des connaissances morpho-sémantiques dont les résultats produits sont souvent redondants. Cette première expérience nécessite cependant d'être complétée. L'expérience suivante propose de déterminer si les couples d'index qui semblent redondants le sont parce qu'ils retrouvent tous deux des documents pertinents ou, au contraire, parce qu'ils échouent à les retrouver.

3. Pour des raisons de lisibilité, nous n'avons pas listé dans cette figure les résultats obtenus par toutes les paires d'index. L'intégralité des résultats obtenus se trouve dans (Moreau, 2006).

4.1.2. *Analyse des cas de redondances*

Pour les cas où les résultats renvoyés par deux index sont similaires ou presque, nous complétons l'expérience précédente et mesurons le taux de documents pertinents effectivement retrouvés par les deux index, et le taux de ceux non retrouvés. Les résultats obtenus sont présentés en figure 2 (colonne 3, cas a, b et c). Pour une liste de résultats de 1000 documents (cas c) par exemple, ce tableau se lit de la manière suivante : pour le premier couple d'index *bigrammes-groupes nominaux*, les listes de résultats retournées par ces deux index sont similaires à 74.91% (col. 2, cas c). Lorsque ces deux index renvoient les mêmes résultats, pour seulement 15.92% des cas (col. 3, cas c), les documents pertinents ont effectivement été retrouvés par les deux index.

À partir de ces données, nous distinguons principalement deux groupes : les paires qui combinent uniquement des informations syntaxiques (*e.g.* les couples *groupes nominaux-trigrammes, bigrammes-trigrammes...*) et les paires qui manipulent des informations morpho-sémantiques (voire seulement morphologiques ou uniquement sémantiques), telles que les couples *lemmes-synonymes, racines-synonymes, lemmes-racines, lemmes-étiquettes sémantiques, synonymes-mots reliés*. Pour le premier groupe, il apparaît nettement que, pour notre collection, les informations syntaxiques, lorsqu'elles sont couplées, ne permettent au SRI de retrouver que très peu de documents pertinents (le couple le plus performant est la paire *bigrammes-termes complexes* qui, sur les 83.03% de résultats identiques que ces deux index fournissent (pour le cas c), ramène seulement 24.52% de documents pertinents). Inversement, les résultats obtenus par la combinaison de connaissances morphologiques, sémantiques ou morpho-sémantiques semblent plus positifs puisqu'ils correspondent en grande partie à la récupération de documents pertinents.

4.1.3. *Analyse des cas de complémentarité entre index*

Nous examinons maintenant les cas où deux index ont un pourcentage élevé de réponses différentes lorsque l'on analyse leur capacité à retrouver les documents pertinents. Nous cherchons à déterminer si ces index se complètent et retrouvent tous deux des documents pertinents différents, ou si ce pourcentage est simplement lié au fait qu'un seul des deux index est efficace pour retrouver les bons documents.

Pour cela, nous proposons de mesurer un indice de complémentarité entre les deux index, obtenu en calculant, pour les cas où les listes de résultats sont différentes, le taux de documents pertinents retrouvés par le premier index et non retrouvés par le second et le taux de documents pertinents ramenés par le second index et non par le premier. Les résultats obtenus sont répertoriés dans le tableau de la figure 2, colonne 4 (cas a, b et c). Ils représentent la contribution de chacun des index d'une paire pour retrouver les documents pertinents. Deux index sont donc considérés comme complémentaires si les deux taux présents dans la colonne 4 (cas a, b ou c) sont plus ou moins proches de 50%. Le fait qu'un des deux taux de cette colonne soit élevé et l'autre faible (comme par exemple pour le couple *lemmes-trigrammes*, colonne 4) signifie que pour les cas où les listes de résultats fournies par les deux index sont différentes, seul l'un des deux index est performant pour retrouver les documents pertinents (pour ce couple, seul l'index des lemmes est efficace). Ce cas ne répond pas à nos objectifs, puisque

paire d'index	col. 2			col. 3			col. 4		
	taux (%) de résultats **identiques**			taux (%) de documents pertinents **retrouvés** par les 2 index			taux (%) de documents pertinents retrouvés **uniquement** par le **1er** index/par le **2nd** index		
	liste 20 docs **(a)**	liste 100 docs **(b)**	liste 1000 docs **(c)**	liste 20 docs **(a)**	liste 100 docs **(b)**	liste 1000 docs **(c)**	liste 20 docs **(a)**	liste 100 docs **(b)**	liste 1000 docs **(c)**
bigrammes + g. nominaux	91,05	82,59	74,91	2,22	8,70	15,92	53 / 47	76 / 24	94 / 6
bigrammes + mots reliés	80,22	65,70	56,69	2,56	19,26	52,77	24 / 76	23 / 77	13 / 87
bigrammes + synonymes	78,36	63,18	53,79	3,01	21,14	59,74	20 / 80	19 / 81	7 / 93
bigrammes + t. complexes	92,72	87,54	83,03	2,81	12,85	24,52	57 / 43	73 / 27	89 / 11
bigrammes + trigrammes	95,01	82,46	68,06	2,15	3,70	5,37	95 / 5	99 / 1	100 / /0
bigrammes+étiq sémantiq.	79,46	65,43	57,18	2,48	20,47	53,36	23 / 77	20 / 80	12 / 88
étiq. gram. + g. nominaux	85,64	71,56	55,54	3,04	10,70	21,32	75 / 25	87 / 13	96 / 4
étiq. gram. + mots reliés	90,06	82,61	78,34	11,39	32,80	64,26	31 / 69	30 / 70	20 / 80
étiq. gram. + synonymes	90,13	81,62	75,37	12,80	34,99	69,69	18 / 82	21 / 79	9 / 91
étiq. gram. + t. complexes	84,74	71,06	61,39	2,26	12,66	31,26	75 / 25	81 / 19	92 / 8
g. nominaux + mots reliés	82,42	66,96	43,61	3,48	13,14	28,30	19 / 81	8 / 92	2 / 98
g. nominaux + synonymes	81,10	64,19	37,95	4,26	14,60	34,72	15 / 85	6 / 94	1 / 99
g. nominaux + trigrammes	92,64	86,89	84,38	0,62	0,90	0,92	77 / 23	81 / 19	81 / 19
lemmes + bigrammes	78,76	62,39	51,95	3,66	23,84	62,79	82 / 18	85 / 15	94 / 6
lemmes + étiq. gram.	90,93	82,96	75,78	13,49	37,54	71,44	88 / 12	93 / 7	98 / 2
lemmes + mots reliés	92,25	85,87	84,71	16,0	41,97	76,81	70 / 30	77 / 23	83 / 17
lemmes + racines	93,96	93,24	94,41	19,27	46,35	79,94	35 / 65	56 / 44	41 / 59
lemmes + synonymes	93,36	88,46	91,79	17,78	44,42	78,74	55 / 45	67 / 33	67 / 33
lemmes + t. complexes	79,15	61,41	43,03	3,23	18,70	50,09	85 / 15	92 / 8	99 / 1
lemmes + trigrammes	79,18	54,42	24,97	1,09	4,28	13,02	93 / 7	98 / 2	99 / 1
lemmes + étiq. sémantiq.	90,93	87,78	86,15	15,88	43,14	76,77	64 / 36	75 / 25	84 / 16
noms propres+ synonymes	81,81	60,55	40,17	3,22	10,16	43,86	7 / 93	6 / 94	4 / 96
racines + bigrammes	76,93	62,41	50,68	3,71	23,24	64,05	83 / 17	84 / 16	94 / 6
racines + g. nominaux	78,96	60,84	33,62	4,55	15,17	38,36	88 / 12	94 / 6	99 / 1
racines + mots reliés	89,99	85,34	83,48	16,17	41,47	77,78	75 / 25	74 / 26	84 / 16
racines + noms propres	79,75	60,49	39,06	3,53	12,66	48,51	95 / 5	98 / 2	98 / 2
racines + synonymes	92,08	86,78	89,31	18,29	43,87	80,07	65 / 35	62 / 38	68 / 32
racines + t. complexes	77,69	61,21	41,28	3,49	17,97	51,25	86 / 14	91 / 9	98 / 2
racines + trigrammes	77,37	55,04	24,23	1,08	4,10	13,86	94 / 6	98 / 2	99 / 1
racines + étiq. sémantiq.	89,46	86,04	84,02	16,31	42,56	78,02	70 / 30	69 / 31	83 / 17
synonymes + mots reliés	91,62	84,97	85,43	15,45	39,63	74,94	65 / 35	63 / 37	75 / 25
t. complexes+g. nominaux	92,92	89,08	85,17	2,62	8,45	12,27	47 / 53	65 / 35	80 / 20
t. complexes + synonymes	80,03	64,30	45,57	3,34	17,13	47,00	15 / 85	10 /90	2 / 98
t. simples + bigrammes	77,64	63,79	55,50	3,33	23,24	56,32	81 / 19	85 / 15	90 / 10
t. simples + g. nominaux	79,42	61,83	40,55	4,02	15,14	31,43	85 / 15	95 / 5	99 / 1
t. simples + lemmes	92,66	90,64	88,84	18,16	45,84	77,74	54 / 46	42 / 58	22 / 78
t. simples + mots reliés	90,00	83,03	81,98	15,48	40,79	73,88	68 / 32	68 / 32	61 / 39
t. simples + noms propres	80,97	60,86	45,25	3,46	12,28	40,72	94 / 6	97 / 3	97 / 3
t. simples + racines	91,34	88,65	87,23	18,67	45,32	78,80	43 / 57	47 / 53	22 / 78
t. simples + synonymes	91,09	84,58	85,04	17,26	43,28	77,33	57 / 43	58 / 42	38 / 62
t. simples + t . complexes	78,68	62,33	46,82	3,28	17,95	43,38	85 / 15	91 / 9	96 / 4
t. simples + trigrammes	78,64	55,88	31,25	1,08	4,13	10,41	93 / 7	98 / 2	99 / 1
t. simples + étiq. sémantiq.	88,43	84,49	82,33	15,21	41,98	74,20	63 / 37	65 / 35	59 / 41
trigrammes + mots reliés	81,91	60,67	33,52	0,80	2,60	7,32	9 / 91	4 / 96	2 / 98
trigrammes +noms propres	95,64	90,61	80,26	0,95	1,37	2,56	32 / 68	22 / 78	9 / 91
étiq. sémantiq.+mots reliés	89,60	83,72	81,08	13,62	38,18	72,19	53 / 47	55 / 45	52 / 48
étiq. sémantiq+synonymes	89,37	84,23	85,34	14,92	40,58	75,38	41 / 59	43 / 57	27 / 73

Figure 2. *Taux de résultats identiques (colonne 2, cas a, b et c), proportion de documents pertinents identiques retrouvés par paire d'index (col. 3 (a, b et c)) et taux de documents pertinents retrouvés par un seul des deux index (col. 4 (a, b et c)).*

nous cherchons uniquement à repérer les informations linguistiques dont le couplage permet au SRI de retrouver davantage de documents pertinents.

D'après les résultats obtenus, il apparaît d'une manière générale que le couplage d'informations linguistiques n'est pas véritablement efficace dans la perspective d'améliorer les performances des SRI puisque très peu d'index sont complémentaires pour retrouver des documents pertinents. Les couples d'informations linguistiques qui ont un taux proche de 50% concernent les index dont les listes de documents sont plus fréquemment identiques que différentes (les taux de résultats différents ne dépassent pas les 18%). Nous pouvons cependant dresser un certain nombre d'observations, suite à cette étude, des cas de complémentarité des informations linguistiques.

Pour la combinaison de connaissances appartenant au même niveau de langue, il apparaît que le couplage d'informations morphologiques peut parfois permettre au SRI de retrouver davantage de documents pertinents. En effet, bien que ces informations semblent le plus souvent retourner des résultats identiques, on constate une véritable complémentarité de ces connaissances lorsque les documents pertinents retrouvés sont différents (pour moins de 20%). Ainsi, pour le couple *lemmes-racines* par exemple, les racines permettent de retrouver dans 59% des cas (colonne 4, cas c) des documents non retrouvés par les lemmes, et réciproquement. Un examen manuel des résultats montre que la prise en compte des lemmes est intéressante lorsque les formes des mots sont irrégulières (*e.g.* les formes *giv* et *gav*). De manière inverse, les racines sont plus performantes pour mettre en correspondance des formes appartenant à des catégories grammaticales différentes (*e.g.retrieval* et *retrieve*). Pour le couplage de connaissances sémantiques, des cas de complémentarité intéressants peuvent également être observés. Les informations de synonymie et de liens morpho-sémantiques (mots reliés) issues de WORDNET sont ainsi souvent complémentaires. Les mots reliés permettent d'étendre les mots des textes en utilisant des informations sémantiques inter-catégorielles, à la différence des synonymes qui sont souvent plus nombreux mais limités à une seule catégorie grammaticale.

En ce qui concerne la combinaison de connaissances appartenant à différents niveaux de langue, les informations morphologiques et sémantiques apparaissent comme les connaissances les plus intéressantes à combiner du point de vue complémentarité. Bien que les taux ne soient pas très élevés, le recours conjoint à des index morphologiques et aux synonymes ou aux mots reliés permet ainsi de retrouver plus de documents pertinents que chaque index pris individuellement. Leur couplage peut constituer une piste d'autant plus intéressante à explorer que les résultats individuels de ces index sont plutôt bons par rapport aux traditionnels termes simples et peuvent certainement encore être améliorés. L'intérêt en RI d'exploiter conjointement des informations morphologiques et syntaxiques d'une part ou des informations syntaxiques et sémantiques (*e.g.* pour les paires *trigrammes-mots reliés, termes complexes-synonymes*...) d'autre part n'a pas été véritablement démontré dans nos expériences, très peu de cas de complémentarité ayant été réellement observés.

4.2. *Application d'une classification ascendante hiérarchique*

Nous étudions à présent de manière simultanée les relations entre tous les index pour avoir un aperçu plus global de la corrélation entre l'ensemble des connaissances prises en compte. Pour cela, nous proposons d'appliquer un algorithme de classifica-

tion ascendante hiérarchique sur les listes de résultats produites par chaque index afin d'essayer de former des classes d'index en fonction des documents pertinents (et de leur positions respectives dans la liste des résultats) qu'ils permettent de retrouver. Les données sur lesquelles s'applique cet algorithme sont construites à partir de la liste des documents qui doivent idéalement être retrouvés par le SRI pour une requête donnée. Pour chacun de ces documents pertinents, on regarde à quel rang chaque index pris individuellement a retrouvé ce document. On procède de la sorte pour tous les documents de la liste et pour tous les index. Lorsqu'un index n'a pas retrouvé le document pertinent, un rang fictif lui est attribué, qui correspond au rang maximal des documents retrouvés par le SRI pour cet index. On obtient à la suite de ce traitement une matrice de données sur laquelle est appliqué l'algorithme de classification ascendante hiérarchique, utilisé avec comme critères la distance euclidienne et l'agrégation selon la méthode de Ward[4]. Le dendrogramme obtenu est présenté en figure 3.

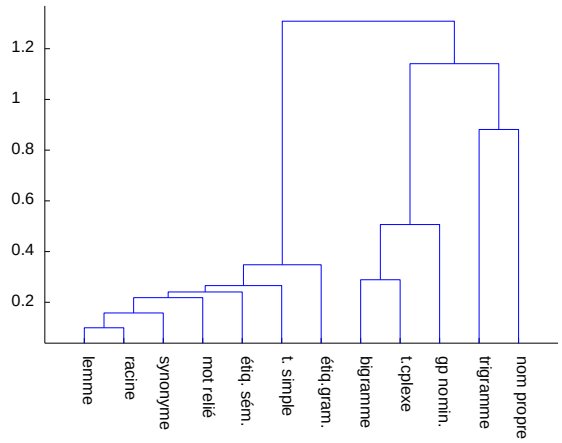

Figure 3. *Classification ascendante hiérarchique des 12 index*

Deux groupes se dessinent nettement. Le premier réunit toutes les informations appartenant aux niveaux morphologique et sémantique de la langue. Au sein de ce groupe, les informations de lemmes, racines, synonymes, mots reliés et étiquettes sémantiques semblent particulièrement liées les unes aux autres, apparaissant en premier à côté des termes simples. Le second groupe rassemble les informations syntaxiques (auxquelles viennent s'ajouter les noms propres). Les résultats sont en accord avec les performances individuelles des index : la première classe rassemble les informations linguistiques qui permettent au SRI de ramener le plus de documents pertinents ; la seconde regroupe les index qui ont le moins d'impact sur les performances des systèmes.

En conclusion, s'il existe des relations de complémentarité entre les informations linguistiques pour retrouver les documents pertinents, elles sont plutôt présentes soit à

4. Pour le paramétrage de l'algorithme, plusieurs autres techniques d'agrégation et mesures de similarités (corrélation de rang...) ont également été expérimentées, laissant toutefois apparaître des résultats très similaires à ceux présentés ici.

travers la combinaison d'informations morphologiques et sémantiques, soit au sein du couplage d'informations d'un seul niveau de langue (morphologique ou sémantique).

5. Conclusion

L'objectif de cet article était double : mesurer dans un cadre homogène l'apport en RI de multiples connaissances linguistiques, et évaluer la pertinence de leur combinaison au sein d'un même SRI. Pour cela, à partir d'une plate-forme de test conçue pour intégrer en parallèle au sein d'un SRI diverses représentations linguistiques des documents et requêtes, nous avons évalué l'apport individuel d'informations linguistiques de différents niveaux de langue et analysé, par le biais d'une étude originale des corrélations, leurs relations pour retrouver des documents pertinents.

Nos expérimentations conduisent à un certain nombre de remarques. Du point de vue de leur impact individuel, les résultats obtenus ont montré, contrairement aux travaux de l'état de l'art, l'impact positif et tranché de certaines connaissances linguistiques en particulier morphologiques et sémantiques. Du point de vue de leur couplage, il ressort que ces connaissances, bien que souvent redondantes, peuvent dans certains cas être complémentaires. Ce constat s'applique néanmoins dans nos expériences uniquement à la combinaison d'informations morphologiques, sémantiques ou morpho-sémantiques.

Bien que les résultats obtenus soient dépendants de la collection de données utilisée, et qu'on ne puisse par conséquent affirmer que les connaissances morphologiques et sémantiques soient toujours les informations linguistiques les plus intéressantes à exploiter en RI, nos travaux ont permis de mettre en place une méthodologie d'évaluation des interactions entre ces connaissances qui présente l'avantage d'être ré-utilisable sur d'autres collections – permettant ainsi de valider ou non les résultats observés – et répétable pour tester d'autres combinaisons d'informations linguistiques.

Plus généralement, ces analyses ont permis de mettre précisément en évidence la façon dont se comportent les informations linguistiques les unes par rapport aux autres pour retrouver des documents pertinents en RI. Elles montrent qu'on ne peut se contenter, si l'on souhaite concevoir un système qui couple des informations linguistiques, de multiplier les connaissances sans prendre en compte leur impact individuel ni leurs relations.

Plusieurs perspectives sont envisagées à la suite de ce travail. Il serait tout d'abord intéressant de réitérer ces expériences en prenant en compte d'autres sortes d'informations linguistiques (de meilleure qualité, plus riches sémantiquement...). En effet, l'exploitation de connaissances plus pertinentes d'un point de vue linguistique permettrait d'évaluer si les résultats obtenus sont liés à la qualité des représentations proposées ou à la façon de les utiliser en RI. Ensuite, ces analyses ayant permis de valider l'intérêt de combiner certaines informations linguistiques en RI, il convient de concevoir une méthode capable de trouver la meilleure façon de les coupler dans un SRI afin d'exploiter de manière optimale leur efficacité. Moreau (2006) a proposé, à partir d'un système d'apprentissage supervisé basé sur les réseaux de neurones, une première méthode de combinaison prenant en compte la plate-forme, les 12 index et des informations sur les requêtes.

6. Bibliographie

Claveau V., Sébillot P., « Extension de requêtes par lien sémantique nom-verbe acquis sur corpus », *Actes de la 11ième conférence annuelle sur le Traitement automatique des langues naturelles, TALN'04*, Fès, Maroc, 2004.

Fagan J. L., Experiments in Automatic Phrase Indexing for Document Retrieval : A Comparison of Syntactic and Non-Syntactic Methods, Thèse de doctorat, Université de Cornell, New-York, États-Unis, 1987.

Friburger N., Maurel D., « Textual Similarity Based on Proper Names », *Proceedings of the 25th ACM International Conference on Research and Development in Information Retrieval, SIGIR*, Tampere, Finlande, 2002.

Gaussier E., Grefenstette G., Hull D., Roux C., « Recherche d'information en français et traitement automatique des langues », *Traitement automatique des langues*, vol. 41, n° 2, p. 473-493, 2000.

Haddad H., Extraction et impact des connaissances sur les performances des systèmes de recherche d'information, Thèse de doctorat, Université Joseph Fourier, Grenoble, France, 2002.

Jacquemin C., « FASTR : A Unification-Based Front-End to Automatic Indexing », *Proceedings of the 4th International Conference : Recherche d'Informations Assistée par Ordinateur, RIAO 94*, New-York, États-Unis, 1994.

Kilgarriff A., Palmer M., « Special Issue on Senseval », *Computers and the Humanities*, 2000.

Moreau F., Revisiter le couplage traitement automatique des langues et recherche d'information, Thèse de doctorat, Université de Rennes 1, Rennes, France, 2006.

Pedersen T., Patwardhan S., Michelizzi J., « WORDNET : :SIMILARITY - Measuring the Relatedness of Concepts », *Proceedings of the 19th National Conference on Artificial Intelligence (Intelligent Systems Demonstration), AAAI-04*, San Jose, États-Unis, 2004.

Porter M. F., « An Algorithm for Suffix Stripping », *Program*, vol. 14, n° 3, p. 130-137, 1980.

Ramshaw L., Marcus M., « Text Chunking Using Transformation-Based Learning », *Proceedings of the 3rd ACL Workshop on Very Large Corpora*, Somerset, États-Unis, 1995.

Robertson S. E., Walker S., Hancock-Beaulieu M., « Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive », *Proceedings of the 7th Text Retrieval Conference, TREC-7*, Gaithersburg, États-Unis, 1998.

Sanderson M., Word Sense Disambiguation and Information Retrieval, Thèse de doctorat, Université de Glasgow, Glasgow, Écosse, 1997.

Savoy J., Morphologie et recherche d'information, Rapport technique, Institut interfacultaire d'informatique, Université de Neuchâtel, Suisse, 2002.

Strzalkowski T., Guthrie L., Karlgren J., Leistensnider J., Lin F., Carballo J. P., Straszheim T., Wang J., « Natural Language Information Retrieval : TREC-5 Report. », *Proceedings of the 5th International Conference on Text Retrieval (TREC)*, Gaithersburg, États-Unis, 1996.

Voorhees E., « Using WORDNET for Text Retrieval », *in* C. Fellbaum (ed.), WORDNET *: An Electronic Lexical Database*, The MIT Press, p. 285-303, 1998.

Zhai C., Tong X., Milic-Frayling N., Evans D. A., « Evaluation of Syntactic Phrase Indexing - CLARIT NLP Track Report », *Proceedings of the 5th Text Retrieval Conference, TREC-5*, Gaithersburg, États-Unis, 1997.