ACADÉMIE D'AIX-MARSEILLE
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

HDR

présentée à l'Université d'Avignon et des Pays de Vaucluse
pour obtenir le diplôme d'Habilitation à Diriger des Recherches

SPÉCIALITÉ : Informatique

Laboratoire d'Informatique d'Avignon (EA 931)

Contributions en faveur d'une meilleure personnalisation de la recherche d'informations

Applications à la tâche questions-réponses, à la recherche de documents audio et à l'accessibilité pour des personnes dyslexiques

par
Patrice BELLOT

Soutenue publiquement le 4 décembre 2008 devant un jury composé de :

M^{me}	Brigitte GRAU	Professeur, LIMSI-ENSIIE, Paris	Rapporteur
M.	Mohand BOUGHANEM	Professeur, IRIT, Toulouse	Rapporteur
M.	Jacques SAVOY	Professeur, Neuchâtel, Suisse	Rapporteur
M.	Jean-François BONASTRE	Professeur, LIA, Avignon	Examinateur
M.	Marc EL-BEZE	Professeur, LIA, Avignon	Examinateur
M.	Philippe BLACHE	Directeur de Recherche, LPL, Aix-en-Provence	Examinateur et Président du Jury

---

Remerciements

Les mots qui vont surgir savent de nous des choses que nous ignorons d'eux.

(R. Char - Chants de la Balandrane)

Les pages qui suivent présentent l'essentiel des travaux de Recherche que j'ai effectués au LIA depuis septembre 2000. Ils correspondent à un travail d'équipe et sont notamment l'œuvre, en respectant l'ordre chronologique, de Laurent Gillard, Benoît Favre et Laurianne Sitbon dont j'ai co-encadré les thèses de Doctorat avec Marc El-Bèze (Université d'Avignon et des Pays de Vaucluse), Jean-François Bonastre (Université d'Avignon et des Pays de Vaucluse) et Philippe Blache (CNRS et Université Aix-Marseille). J'espère que les quelques années passées en ma compagnie leur auront été aussi enrichissantes que leur présence et leur travail l'ont été pour moi. Il va de soi que je remercie tout autant M. El-Bèze, J.F. Bonastre et P. Blache pour m'avoir accordé leur confiance et leur soutien ainsi que pour leurs (très) précieux conseils. Si j'ai mentionné en premier lieu les doctorants, je n'oublie pas non plus ceux que j'ai suivis en DEA, en Master Recherche ou en Post-Docs, à savoir Christian Raymond, Nicolas Flavier, Pascal Pouchoulin, Mehdy Draoui, Inès Temou, Thierry Waszak, Marie-Laure Guénot, Mathieu Estratat et Rémi Lavalley.

Je tiens à remercier très chaleureusement Brigitte Grau, Mohand Boughanem et Jacques Savoy pour avoir accepté de rapporter sur ce Mémoire et pour leurs commentaires. J'adresse des remerciements tout particuliers à l'ensemble de mes collègues du LIA, de l'IUP GMI et de l'Université d'Avignon pour leur amitié et la qualité de travail (et de "non travail" !) qu'ils savent offrir.

Enfin, je ne peux terminer ces lignes sans remercier "ma petite famille" pour avoir supporté mes sautes d'humeur durant l'été dernier (surtout) et pour ma faible disponibilité. Merci à vous, Valérie, Anouk et Tristan. Merci enfin à Firminhac pour sa verdure et son silence matinal. Il se reconnaîtra.

A ma grand-mère, Hélène Bellot-Reinaudi, partie au printemps 2008.
A tous mes proches ; ils savent créer des brins d'éternité.

Résumé

Dans un article récent sur les enjeux de la recherche d'informations, Belkin (2008) rappelle que la question de la personnalisation est annoncée comme majeure depuis une vingtaine d'années mais que la plupart des tentatives vers une étude systématique ont échoué. S'il n'est pas question de prétendre que les évaluations TREC Interactive puis TREC Hard et TREC ciQA, pour ne mentionner qu'elles, n'ont abouti à rien dans ce domaine, force est de constater que l'utilisateur n'intervient que très peu dans les systèmes de recherche de l'Internet et qu'aussi bien des paradigmes d'évaluation que des stratégies efficaces restent encore à trouver.

Dans les différents chapitres qui constituent ce mémoire d'Habilitation, nous présentons trois directions vers une personnalisation de la recherche d'informations. La première correspond à une analyse du besoin en information d'un utilisateur qui permet de distinguer recherche documentaire et recherche de réponses précises sachant que celles-ci peuvent être des informations factuelles, des définitions ou des explications. Cette distinction correspond à celle entre Recherche d'informations (RI) et questions-réponses (QR) mais, pour être utilisables, les systèmes correspondants devront être fusionnés. En outre, il sera utile d'inciter l'utilisateur à changer ses habitudes pour laisser de côté les requêtes mots-clés et (re)venir à des requêtes en langue naturelle. Les solutions logicielles que nous avons développées ont été évaluées dans le cadre des campagnes TREC, CLEF et EQUER.

La seconde direction que nous avons suivie est celle de l'aide à la navigation dans de grandes bases documentaires mélangeant fichiers audio et textes. Elle consiste à définir une interface homme-machine permettant un survol chronologique, par l'exploitation de méthodes de reconnaissance de la parole, d'indexation sémantique (LSI), de segmentation thématique et de résumé automatique (campagne d'évaluation DUC), des documents de la collection. Les techniques d'indexation en jeu n'exploitent pas la totalité des traits propres à l'audio (prosodie, hésitations...) et cela fera l'objet de travaux futurs. L'objectif étant de parvenir à des systèmes multimodaux dans lesquels les documents audio ne sont pas *noyés* parmi des documents texte plus nombreux et plus verbeux.

La troisième direction consiste à prendre en compte la capacité de lecture et d'écriture d'un utilisateur dans le calcul du score de pertinence d'un document vis à vis d'une requête. Les avancées les plus récentes de la technique et de l'imagerie médicale nous offrent des modélisations plausibles de nos fonctionnements cognitifs dont nous

pouvons nous inspirer afin de simuler l'humain dans des domaines tels que le langage et la pensée. Nous nous sommes plus particulièrement intéressé aux modèles cognitifs de la lecture et à la tentative de les exploiter afin de définir des systèmes de recherche d'informations capables d'estimer l'effort nécessaire à la compréhension d'un document et d'être suffisamment robustes pour accepter des requêtes mal orthographiées. Les modèles de recherche d'informations usuels permettent d'ordonner des documents en fonction de la quantité d'informations qu'ils véhiculent vis à vis de ce que l'utilisateur a exprimé dans sa requête tout en tenant compte, dans le meilleur des cas, du taux de nouveautés apportées par rapport à d'autres documents déjà connus. Il s'agit d'une vision purement informationnelle de la pertinence posant l'hypothèse que plus le nombre d'informations nouvelles est grand, plus le document est susceptible d'intéresser l'utilisateur. Cela s'avère exact dans une certaine mesure mais ne tient pas compte du fait que les besoins sont différents suivant le niveau d'expertise de l'utilisateur : une personne novice dans un domaine sera certainement plus intéressée par un document de vulgarisation que par une étude approfondie, au vocabulaire et à la structure complexes. Cela est vrai à plus forte raison pour des personnes ayant des difficultés élevées de lecture tels les dylsexiques. Il s'agit alors de définir de nouvelles mesures prenant en compte cet aspect tout en offrant la possibilité de présenter d'abord les documents les plus "simples", les plus "lisibles".

La problématique de la personnalisation et de la prise en compte de l'utilisateur en recherche d'informations renvoie naturellement à celle, bien plus large, des fondements du traitement automatique des langues, au croisement de la linguistique et de l'informatique, toutes deux rejointes par la psycholinguistique et la psychologie cognitive pour l'étude des comportements individuels, les neurosciences pour l'étude des racines physiologiques du langage mais aussi par la sémiologie pour des analyses globales des usages et des significations. Ce croisement pluridisciplinaire est un enjeu majeur des années à venir si l'on veut aller au-delà, pour paraphraser K. Sparck-Jones, de la seule étude permettant d'espérer (et encore ne s'agit-il que d'un espoir sans même être convaincu de la significativité des gains) grappiller quelques points de précision en recherche ad-hoc.

Il va de soi que les recherches présentées correspondent à un travail d'équipe. Elles sont ainsi l'œuvre des activités conduites au LIA depuis septembre 2000, et notamment, en respectant l'ordre chronologique, celles de Laurent Gillard, Benoît Favre et Laurianne Sitbon dont j'ai co-encadrées les thèses de Doctorat avec Marc El-Bèze (Université d'Avignon et des Pays de Vaucluse), Jean-François Bonastre (Université d'Avignon et des Pays de Vaucluse) et Philippe Blache (CNRS et Université Aix-Marseille).

Table des matières

I Questions-réponses, segmentation et détection de plagiats **15**

1 Autour de la tâche questions-réponses **17**
1.1 Analyse de questions en langue naturelle 18
1.1.1 Classification de questions 19
1.1.2 Prédiction de la difficulté d'une question 23
1.1.3 Identification de questions complexes au sein de courriers électroniques 25
1.1.4 La recherche de questions similaires 27
1.2 Questions-réponses et Recherche d'informations 29
1.2.1 Les principaux modèles de recherche d'informations 29
1.2.2 Evaluation du module de recherche d'informations présent dans les moteurs QR 30
1.2.3 Quelques résultats 32
1.3 Segmentation automatique et filtrage de passages 33
1.3.1 Comparaison de différentes approches 34
1.3.2 Proposition d'une mesure de densité pour la recherche de passages 36
1.3.3 Evaluation de la densité pour questions-réponses 38
1.3.4 Perspectives dans la recherche de passages 39
1.4 Extraction de réponses pour des questions factuelles 40
1.5 Usage des bases de connaissance dans SQuaLIA 42
1.6 Perspectives 43
1.6.1 Questions-réponses, bases de connaissances et domaine de spécialité 44
1.6.2 Questions-réponses entre RI et TAL 49

2 Segmentation, enrichissement de requêtes et détection de plagiats **53**
2.1 Segmentation non supervisée par chaînes lexicales pondérées 53
2.1.1 Proposition à base de chaînes lexicales pondérées 54
2.1.2 Evaluation durant la campagne DEFT 2005 57
2.2 Segmentation pour l'expansion de requêtes 58
2.2.1 Expansion à partir de ressources externes. 58
2.2.2 Des arbres de décision non supervisés pour une segmentation thématique orientée requête 59

2.2.3 Proposition d'une méthode d'expansion à partir d'arbres de décision non supervisés . . . 62
2.3 Détection de plagiats : le projet PIITHIE . . . 64
2.3.1 Présentation . . . 64
2.3.2 Etat de l'art . . . 66
2.3.3 Détection de citations pour l'identification de plagiats . . . 68
2.3.4 Segmentation de textes pour la détection de plagiats . . . 72

II Recherche personnalisée : navigation et lisibilité 79

3 Recherche d'informations et résumé automatique audio 81
3.1 Recherche d'informations au sein de documents audio . . . 81
3.2 Navigation dans des bases audio par résumé automatique . . . 83
3.3 Une approche de résumé automatique audio orienté requête . . . 84
3.4 Expériences durant DUC 2006 . . . 87

4 Recherche d'informations, lisibilité et dyslexie 91
4.1 Introduction . . . 92
4.2 Modélisation cognitive de la lecture . . . 94
4.3 Critères pour estimer la difficulté de lecture d'un texte . . . 99
4.4 La dyslexie comme trouble du langage . . . 104
4.5 Robustesse de SQuaLIA face à des requêtes bruitées . . . 108
4.6 Hypothèses de réécriture de questions dysorthographiées . . . 110
4.6.1 Données recueillies . . . 111
4.6.2 Correction par phonétisation et retranscription . . . 112
4.7 Lisibilité . . . 116
4.7.1 Les mesures de lisibilité de Flesch . . . 117
4.7.2 Réordonnancement des documents trouvés selon la mesure de Flesch . . . 117
4.7.3 Proposition d'une mesure de lisibilité adaptée à la dyslexie . . . 118
4.8 Perspectives . . . 121
4.8.1 Fonction de score d'un document combinant pertinence et lisibilité 121
4.8.2 Un processus spécifique d'expansion par retour de pertinence... et de lisibilité . . . 123

5 Perspectives générales 127

III Annexes 135

6 Aide à l'apprentissage de la lecture 137

7 Recherche d'information suivant le modèle vectoriel 141
7.0.3 Mesures de similarité . . . 142
7.0.4 Pondération des entrées de l'index : les critères tf et idf . . . 142

8 Recherche d'information suivant le modèle probabiliste **145**
8.0.5 Le modèle de pertinence binaire 146
8.0.6 Estimation des paramètres 148

9 La méthode d'enrichissement de requêtes de Rocchio **155**

10 Mesures d'évaluation en recherche documentaire et en questions-réponses **157**
10.0.7 Mesures spécifiques à la tâche questions-réponses 158
10.0.8 Mesures de recherche documentaire spécifique à QR 160

11 De TREC à CLEF en passant par EQUER : synthèse des résultats **161**

12 Quelques méthodes de segmentation linéaire non supervisées **165**

13 Curriculum vitae **169**
13.1 Publications personnelles 172

Index **176**

Index des auteurs **178**

Liste des illustrations **187**

Liste des tableaux **189**

Bibliographie **191**

Introduction

Nos travaux se situent à la confluence des communautés scientifiques en Recherche d'Informations (RI) et en Traitement Automatique des Langues (TAL). Ils concernent des applications telles que la recherche documentaire, les moteurs questions-réponses, la segmentation thématique et le résumé automatique orienté requête. Privilégiant autant que possible les approches numériques, les méthodes que nous employons sont (relativement) indépendantes de la langue même si de nombreux traitements reposent sur des analyses de surface exploitant des ressources propres à une langue donnée (par exemple l'étiquetage morpho-syntaxique et la lemmatisation). Si nous travaillons le plus souvent dans le domaine général, nous nous intéressons depuis deux ans à des domaines de spécialité tels que la chimie organique, l'écophysiologie végétale ou les produits de la grande consommation. Cependant, nos résultats sur ces domaines de spécialité qui exploitent des ontologies formelles expertes et diverses sources de connaissance plus ou moins structurées, sont encore trop limités pour figurer dans ce mémoire autrement qu'en perspectives. À l'exception de quelques publications en marge avec nos activités récentes, l'ensemble des travaux publiés est évoqué dans ce document. Une liste complète de publications personnelles est malgré tout donnée en Annexes (p. 169).

Le chapitre 1 concerne la tâche questions-réponses et correspond pour l'essentiel à la thèse de doctorat de Laurent Gillard co-encadrée avec Marc El-Bèze, mais également au Master Recherche de Nicolas Flavier, à des parties des thèses de Laurianne Sitbon et de Jens Grivolla ainsi qu'à des travaux en collaboration avec Karine Lavenus. Le chapitre 2 concerne d'une part l'expansion de requêtes selon des arbres de décision non supervisés (Master Recherche de Christian Raymond), la segmentation thématique selon des chaînes lexicales pondérées (projet Technolangue OURAL) et d'autre part d'un projet en cours sur la détection de plagiats (projet ANR PIITHIE en collaboration avec le laboratoire LINA de l'Université de Nantes et les sociétés Syllabs, Sinequa et Advestigo) abordé sous ma direction par Thierry Waszak dans le cadre de son Master Recherche et par Marie-Laure Guénot et Mathieu Estratat, post-doctorants.

Dans la seconde partie, nous nous intéressons à une adaptation des systèmes de recherche d'informations orientée utilisateur. Le chapitre 3 est consacré au sujet de la recherche d'informations, du résumé automatique et de la navigation au sein de transcriptions de documents audio (thèse CIFRE de Benoît Favre avec la société Thalès, co-encadrée avec Jean-François Bonastre). Enfin, le chapitre 4 concerne le traitement automatique des langues pour certains handicaps langagiers et concerne plus spéci-

fiquement les modèles connexionnistes de la lecture et une adaptation de la recherche d'informations pour des utilisateurs dyslexiques. Il est le cœur de la thèse de Laurianne Sitbon, co-encadrée avec Philippe Blache. Ce chapitre est l'occasion de développer un peu plus largement les perspectives envisagées mais aussi d'ouvrir la problématique vers les sciences cognitives, l'intelligence artificielle et la neuropsychologie qui, en plus de la linguistique et de l'apprentissage automatique, peuvent jouer un rôle important dans le développement de systèmes informatiques adaptés.

Chaque chapitre comporte une introduction spécifique, et le tout se termine par des perspectives générales sur le traitement automatique des langues qui correspondent aux orientations que je souhaite suivre dans mes recherches futures. Le lecteur trouvera ici une synthèse des articles déjà publiés qui permet d'obtenir une vision que j'espère cohérente des propositions formulées et des résultats obtenus. Par rapport aux publications, l'étude bibliographique a été largement étendue de façon à permettre une mise en perspective plus large et, si possible, de dégager de nouvelles pistes à explorer.

Bref retour sur ma thèse de Doctorat

Ma thèse de doctorat qui a débuté en septembre 1996 et qui a été soutenue en janvier 2000 concernait la classification et la segmentation automatiques par des méthodes non supervisées comme aides à la recherche d'informations. En ce qui concerne la classification automatique de textes, nous avons élaboré un algorithme combinant classification hiérarchique partielle et méthode des « nuées dynamiques ». Cette méthode a été testée lors de l'évaluation TREC–7 (de Loupy et al., 1999). De nombreux tests et estimations empiriques ont été effectués sur les données des campagnes Amaryllis et TREC. Nous avons notamment mis en lumière les bénéfices obtenus lorsque le nombre de classes utilisées était choisi dynamiquement en fonction de certaines caractéristiques de la requête dont sa taille. Cette option a été présentée lors du congrès IEEE-ICSC'99 et publiée dans un numéro de LNCS (Bellot et El-Bèze, 2000b). Elle a apporté une solution nouvelle à un problème difficile. Cette méthode de classification est décrite plus largement dans deux articles parus dans les revues Traitement Automatique des Langues (Bellot et El-Bèze, 2001b) et Systèmes et Sécurité (Marteau et al., 1999).

Nous avons proposé un second algorithme de classification. Celui-ci emploie des arbres de décision non supervisés. Cet algorithme est original à plus d'un titre (généralement ces arbres sont utilisés après une phase d'apprentissage) et s'avère plus rapide que les méthodes de classification classiques (hiérarchiques ou de partitionnement). En outre, il offre la possibilité de représenter chaque classe par une expression booléenne composée de certains mots clés issus des documents classés. Une telle représentation permet de guider l'utilisateur dans son exploration des classes obtenues. Cette méthode a été testée durant la campagne Amaryllis'99. Finalement, une comparaison des résultats obtenus avec les arbres ou avec la combinaison hiérarchie et « nuées dynamiques » a été présentée lors du congrès RIAO'2000 (Bellot et El-Bèze, 2000a).

Nos différentes expérimentations ont souligné la nécessité de mener des recherches

sur des méthodes de segmentation thématique. Grâce à un découpage automatique, chaque passage extrait d'un document est vu comme une unité documentaire à part entière. Un premier algorithme basé sur des modèles markoviens a été présenté lors du congrès JST'97 (Bellot et El-Bèze, 1997). Nous avons ensuite mis en évidence le fait qu'une segmentation pouvait être obtenue à partir d'une classification (autrement dit, toute méthode de classification appliquée à des extraits de documents permet d'obtenir une segmentation de ces documents). Si deux passages contigus dans un texte se trouvent dans des classes thématiques différentes, une marque de segmentation est apposée entre eux. Notre méthode de classification mêlant hiérarchies et partitionnement est de complexité trop élevée pour pouvoir être appliquée à grande échelle sur tous les passages des textes (phrases ou paragraphes par exemple). Par contre, notre deuxième méthode, basée sur les arbres de décision, autorise la classification de toutes les phrases des 1 000 premiers documents trouvés lors d'une recherche documentaire en un temps de l'ordre d'une dizaine de secondes. Les résultats de son application à la segmentation sont présentés dans un article de la revue TSI (Bellot et El-Bèze, 2001a) mais aussi dans (Bellot, 2000b).

Première partie

Questions-réponses, segmentation et détection de plagiats

Chapitre 1

Autour de la tâche questions-réponses

Sommaire

1.1 Analyse de questions en langue naturelle 18
1.1.1 Classification de questions 19
1.1.2 Prédiction de la difficulté d'une question 23
1.1.3 Identification de questions complexes au sein de courriers électroniques 25
1.1.4 La recherche de questions similaires 27
1.2 Questions-réponses et Recherche d'informations 29
1.2.1 Les principaux modèles de recherche d'informations 29
1.2.2 Evaluation du module de recherche d'informations présent dans les moteurs QR 30
1.2.3 Quelques résultats 32
1.3 Segmentation automatique et filtrage de passages 33
1.3.1 Comparaison de différentes approches 34
1.3.2 Proposition d'une mesure de densité pour la recherche de passages 36
1.3.3 Evaluation de la densité pour questions-réponses 38
1.3.4 Perspectives dans la recherche de passages 39
1.4 Extraction de réponses pour des questions factuelles 40
1.5 Usage des bases de connaissance dans SQuaLIA 42
1.6 Perspectives 43
1.6.1 Questions-réponses, bases de connaissances et domaine de spécialité 44
1.6.2 Questions-réponses entre RI et TAL 49

NB. Ce chapitre correspond majoritairement aux travaux réalisés durant la thèse de Doctorat de Laurent Gillard, co-encadrée par Marc El-Bèze (LIA) et moi-même entre 2002 et 2007.

Les moteurs de questions-réponses actuels correspondent généralement à la concaténation de plusieurs systèmes indépendants. Parmi les plus importants, on trouve les analyseurs de questions (catégorisation, extraction de focus...), les analyseurs de textes (étiqueteurs d'entités nommées, étiqueteurs morpho-syntaxiques, analyseurs syntaxiques profonds...), les moteurs de recherche documentaire (extraction d'une liste de documents ou de passages depuis le corpus de recherche) et enfin les modules d'extraction de réponses candidates. Ceci correspond à une vision séquentielle de la recherche d'informations de type questions-réponses pour laquelle, à partir de l'analyse d'une question, on va réduire le champ de recherche depuis le corpus jusqu'à la réponse, en passant par un ensemble de documents, un ensemble de passages de ces documents et un ensemble de phrases extraites des passages. Ce découpage a une double origine : la préexistence de chacun des modules (évalués par l'intermédiaire de campagnes d'évaluation indépendantes : TREC, MUC...) et l'incapacité toujours réelle des techniques d'analyse profonde à opérer rapidement sur de grands corpus.

La vision minimaliste de la tâche questions-réponses correspond ainsi à une séquence d'opérations mises en œuvre par des systèmes logiciels éprouvés sur des textes de même nature mais dans des contextes différents. L'enjeu est alors double. Il consiste d'une part à vérifier, pour chacun des modules, que les approches les plus performantes dans un contexte autre que questions-réponses, sont les mêmes pour cette tâche et, d'autre part, parvenir à diminuer l'indépendance de ces modules afin de prendre en compte le problème dans sa globalité.

Un historique complet de la tâche questions-réponses pourra être trouvé dans (Grau et Chevallet, 2008; Maybury, 2004).

1.1 Analyse de questions en langue naturelle

L'analyse d'une question correspond généralement à l'application successive de plusieurs traitements. Le premier, optionnel, peut inclure une correction orthographique et/ou grammaticale de la question voire une normalisation. Par exemple, transformer la question TREC 885 "*Rotary engine cars were made by what company ?*" en "*What company were rotary engine cars made by ?*" permet de positionner le pronom interrogatif *what* en début de question, facilitant les traitements ultérieurs (Moldovan et al., 2003). Un autre traitement correspond à l'extraction des mots-clés et du *focus* (thème) de la question après suppression des mots outils et la reconnaissance de certaines contraintes (dates, lieux...) ou relations entre les mots clés. Cette étape permet d'aboutir à l'expression de la requête qui sera posée au module suivant : recherche de documents ou bien de passages. L'apport de réseaux sémantiques tels que Wordnet (Miller et al., 1990; Miller, 1995) peut s'avérer significatif aussi bien pour relier les termes de la question à des classes sémantiques et déterminer le type de réponse attendue que pour enrichir la question pour l'étape de recherche documentaire (Pasca et Harabagiu, 2001).

1.1.1 Classification de questions

Les étiqueteurs de questions mettent généralement en correspondance les questions avec les types d'entités nommées pouvant être recherchés par le moteur de questions-réponses : noms propres de personnes ou de lieux, dates, durées *etc*. L'étiquetage des questions correspond dans ce cas à une catégorisation en fonction des réponses attendues. Cependant, la réponse attendue peut ne pas être une entité nommée mais un objet textuel de structure complexe (explication, définition) ou au contraire très simple dans le cas binaire des questions booléennes ("oui/non")[1]. D'autres questions n'attendent pas une seule réponse mais plusieurs ("*Quels sont les trois derniers ministres de l'éducation en France ?*") et enfin, cas des questions "contextuelles", il peut n'être possible de répondre qu'en tenant compte de l'*historique i.e.* des questions précédemment posées au système. La figure 1.1 présente une hiérarchie des types de réponses possibles pour une question.

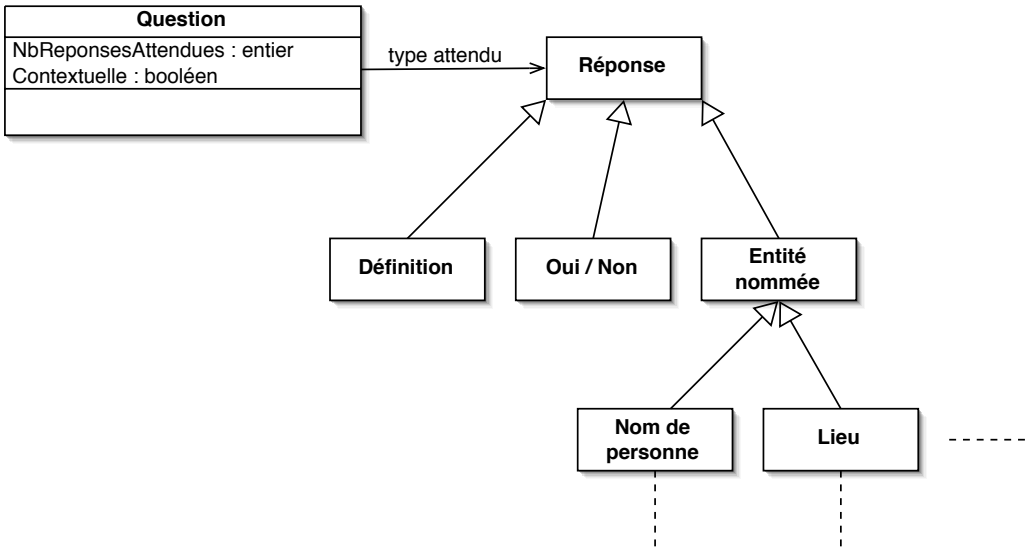


FIG. 1.1: *Processus de catégorisation d'une question précise. S'il s'agit d'une question de type "Liste", elle appellera plusieurs réponses, s'il s'agit de questions chaînées, la réponse devra être* contextuelle. *L'analyse de la question permet de déduire le type de réponse cherchée.*

D'une catégorisation basique à la hiérarchie de Sekine. La taxinomie des questions définie par Ittycheriah et Roukos (2002) essentiellement basée sur le pronom interrogatif présent dans les questions[2].

[1] Les travaux que nous avons effectués sur les questions booléennes ne sont pas repris ici. On pourra se reporter à (Gillard, 2007, chapitre 8) mais aussi à (Gillard et al., 2006c, 2007a) et (El-Bèze, 2006) *in* (Sabbah, 2006b).

[2] Cela n'est cependant pas toujours suffisant puisque des pronoms comme *quel* en français ou *what* en anglais peuvent correspondre à un grand nombre de types d'entités.

L'étude des articles publiés à l'issue des campagnes d'évaluation TREC montre que le nombre de catégories peut varier de 5 à plus d'une centaine. La version 2002 du système d'IBM (Ittycheriah et Roukos, 2002) n'emploie que cinq classes : *Name Expressions (person, organization, location, country...), Time Expressions (date, time...), Number Expressions (percent, money, ordinal, age, duration...), Earth Entities (weather, plant, animal, ...) and Human Entities (events, disease, company-role,...)*. D'autres systèmes exploitent un nombre supérieur de catégories : une cinquantaine avec deux niveaux hiérarchiques pour Li et Roth (2002) qui en étudient les distributions sur les questions TREC ou pour Clarke et al. (2003) : date, ville, température, saison... ainsi que des catégories liées explicitement à des unités (quantité, distance...)[3].

Méthodes de catégorisation des questions. La plupart des systèmes emploient des catégoriseurs basés sur des patrons lexicaux à partir de simples heuristiques sur des mots clés. Ainsi, Sutcliffe (2003) a classé correctement 425 des 500 questions de TREC 11 parmi 20 classes, et Plamondon et al. (2003) ont employé seulement 40 patrons pour correctement positionner 88 % des 492 questions de TREC 10 parmi 11 classes. En fonction du degré de généricité des patrons, leur nombre peut varier significativement : le système Qanda exploite ainsi une base de plusieurs milliers de syntagmes nominaux (Burger, 2003) déterminés à partir de Wordnet (méthode semi-automatique). Par ailleurs, Chang et al. (2003) dressent un dictionnaire de plusieurs centaines de mots auxquels correspondent des types d'entités nommées. Par exemple, `city` et `town` correspondent au type d'entité "CITY" tandis que `player` correspond au type "PERSON". Un étiquetage syntaxique de la question permet d'extraire le substantif qui aide à déterminer le type d'entité nommée attendue : ce substantif est soit le dernier du premier groupe nominal de la question (*Which past and present NFL players have...*) soit le dernier du second groupe nominal dans le cas où le premier se termine par un substantif *abstrait* (`type`, `kind`...).

Parmi les approches mentionnées dans la littérature, relevons (Wu et al., 2003) qui utilisent `LinkParser`[4] pour détecter le sujet, le prédicat et les éventuelles contraintes dans la question (*adverbial modifiers*). Par exemple, pour la question "*What book did Rachel Carson write in 1962 ?*", le sujet est `Rachel Carson`, le prédicat `wrote` et les modifieurs `in 1962`. Par la suite, cette analyse permet de rechercher dans les textes, par ordre de préférence, *Rachel Carson wrote in 1962*, *Rachel Carson wrote*, *Rachel Carson* et d'associer à chaque réponse potentielle un score dépendant du respect ou non des contraintes ainsi imposées (contraintes sur le sujet, le verbe ou l'année).

Parmi les autres contraintes éventuelles, on peut identifier :
- géographiques : *in France* ;
- temporelles : *in 1968* ou bien *before 1968* ;
- de précédence : *first woman in space* ;
- ....

[3] Une stratégie courante consiste à associer des ensembles de mots spécifiques à des types de questions qui seront ensuite utilisés pour enrichir la question.

[4] http://www.link.cs.cmu.edu/link/

Citons également Zhang et Lee (2003) qui établissent des listes d'expressions régulières génériques à partir des questions posées dans les différentes campagnes TREC en remplaçant toutes les occurrences d'entités nommées par le nom de leur classe ("*When was Albert Einstein born ?*" devient `When was <PERSON> born ?`). Ces expressions (il y en a autant que de questions) sont ensuite utilisées pour établir un modèle de langage lissé sur les bigrammes. L'étiquetage se fait ensuite en suivant une approche bayésienne sur des modèles de langue.

Catégories et stratégies de résolution. De manière complémentaire à la correspondance entre la question et le type de réponse cherchée, Lehnert (1978) a proposé une classification conceptuelle des questions (cause/conséquence, but, capacité, vérification, procédure, propriétés *etc.*) qui est destinée à une *stratégie* de résolution.

Certains systèmes proposent de déterminer simultanément le type de réponse cherchée et d'établir la stratégie qui peut y conduire. C'est le cas du système JAVELIN (Nyberg, 2003) qui catégorise la question selon la proposition de Graesser et al. (1992, chapitre 9), elle-même reprise de Lehnert (1978). Par exemple, pour les questions "*Who invented the paper clip*" et "*What did Vasco da Gama discover*", la catégorie est "*event-completion*" tandis que le type de réponse cherchée est "*proper-name*" pour la première question et "*object*" la deuxième. Il en va de même, ou presque, pour le système du LIMSI (Ferret et al., 2002) qui met en correspondance catégorie et patron de réponse : la "catégorie" est une *forme* de réponse. Ainsi pour la question "*When was Rosa Park born ?*", la catégorie de la question est "`When Be PN Born`".

La catégorisation des questions dans SQuaLIA. Pour notre participation à TREC 11 (Bellot et al., 2003), nous avons établi une hiérarchie à partir de l'étude manuelle des questions des années précédentes. La hiérarchie était composée de 31 catégories de niveau supérieur (acronyme, adresse, profession, couleur...), de 58 catégories de niveau 2 et de 24 catégories de niveau 3. Par exemple, "nom propre" a été subdivisé en 10 sous-catégories telle que "acteur", "musicien" ou "politicien". Pour catégoriser les questions, nous avons développé un classifieur symbolique à base de règles de réécriture (ce qui correspond à une certaine normalisation des questions afin de minimiser le nombre de patrons utilisés ensuite pour la catégorisation)[5] et de 156 patrons[6]. De manière à catégoriser plus efficacement des questions qui ne correspondaient à aucun des patrons, nous

[5] Une même question peut être posée selon de nombreuses variantes. Plutôt que d'essayer d'énumérer l'intégralité des variantes pour chaque type de réponse possible, nous commençons par normaliser les questions. Ainsi, la question *Quel est le nom de la ville qui...* équivaut à la question *Quelle est la ville qui...* ou encore *Comment s'appelle la ville qui... etc.* Naturellement, ces équivalences sont valables pour d'autres recherches que celles portant sur une `ville`. Les expressions : `quel est le nom de`, `quelle est de/à quelle`, `comment s'appelle`, `comment se nomme`, `quel nom est donné`, `comment nomme-t-on`, `citez une`, `citer une`, `à cause de quel`, `à bord de quel`, `grâce à quel`, `dans quel`, `par quel`, `avec quel`, `pour quel` sont toutes remplacées par `quel`. Ainsi *Quel est le nom de la ville* devient `Quel ville`.

[6] Ces 156 patrons sont représentatifs des questions des premières éditions de TREC et ne prétendent pas à une quelconque exhaustivité.

avons également mis en place un classifieur statistique basé sur des arbres de classification sémantique (Kuhn et De Mori, 1995) (voir figure 1.2) opérant aussi bien sur les mots (les lemmes) que sur les étiquettes morpho-syntaxiques ou sur les types d'entités nommées reconnues. En utilisant, comme base d'apprentissage, les 259 premières questions issues de la campagne TREC 10, nous obtenons une précision de 68,5 % pour la catégorisation des 150 autres questions de cette année là. La combinaison des deux approches symboliques (en amont) et numérique (dans le cas où l'approche symbolique ne donnait pas de réponse) a permis d'atteindre une précision de 80 % et d'indiquer, le cas échéant, plusieurs catégories possibles, chacune associée avec une probabilité.

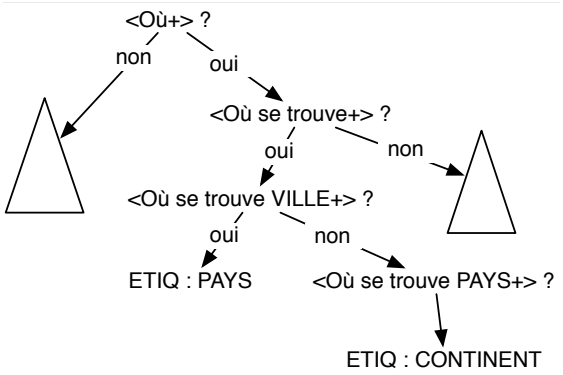


FIG. 1.2: *Exemple d'arbres de classification sémantique utilisé pour catégoriser une question. Si la question est de la forme : "Où se trouve VILLE ?", la catégorie de la réponse cherchée est PAYS (nom de pays).*

À partir de l'étude détaillée d'un grand nombre de questions effectuée avec Karine Lavenus (Lavenus et al., 2004a,b), nous avons enrichi la hiérarchie des types possibles en reprenant une grande partie des propositions de Sekine et al. (2002) et nous avons augmenté la base de patrons. Le tableau 1.1 montre les résultats du module de catégorisation sur les questions factuelles en français d' EQUER (comparables avec l'état de l'art) et CLEF 2006 sur celles en anglais et en français (Gillard et al., 2007b).

Langue	Correcte	Incorrecte	Aucune catégorie trouvée
Français (EQUER)	85 %	2 %	13 %
Français (CLEF 2006)	86 %	2 %	12 %
Anglais (CLEF 2006)	78 %	13 %	5 %

TAB. 1.1: *Evaluation du module de catégorisation sur les questions factuelles EQUER et sur les pistes FR-FR et EN-FR de CLEF2006.*

Questions définitoires. Ces questions ne sont pas traitées dans SQuaLIA selon la même architecture logicielle que les autres types de questions. À la place, Laurent Gillard a implémenté un composant qui, à partir du *focus* de la question, filtre l'ensemble des phrases du corpus qui le contient. Ensuite, des patrons d'extraction permettent d'obtenir un jeu de phrases plus restreint autour de schémas définitoires simples :

- "focus, [réponse], *"
- "[réponse], focus, *"
- "le$|$la [réponse] focus"...

Les [réponses] candidates sont ensuite ordonnées en fonction de leur fréquence dans le corpus en tenant compte de la distribution des mots qu'elles contiennent dans les autres réponses, de la nature du premier mot, des étiquettes morpho-syntaxiques... Ce composant "questions définitoires" est également en charge de l'identification et du développement d'acronymes. Cette approche fait le pari de la redondance informationnelle dans le corpus (on recherche la définition la *plus commune*) et a très bien fonctionné durant la campagne CLEF 2006 (tableaux p. 162).

1.1.2 Prédiction de la difficulté d'une question

La prédiction de la capacité à répondre à une question (qu'elle soit liée au système lui-même ou à l'absence de réponse dans le corpus) s'impose, afin de fournir au système un seuil d'admission de la question, et éventuellement d'entamer des procédures de reformulation, automatiques ou à l'aide d'un dialogue avec l'utilisateur (Jonsson et al., 2004). Afin de pouvoir prédire si l'on pourra répondre, le principal élément dont on dispose est le constat des résultats obtenus lors de campagnes d'évaluation. Outre les types de questions (factuelles, définitoires, *etc.*), une première piste qui découle de l'architecture même des systèmes, est d'utiliser le type de réponse attendu, qui est disponible dès la première étape de traitement. Cela permet de s'affranchir des autres étapes si on pense dès ce stade ne pas pouvoir répondre.

Les travaux de Lavenus et Lapalme (2002) prennent en compte une analyse en profondeur du type de question afin de déterminer, sinon la capacité à répondre, la difficulté de le faire. Le tableau 1.3 recense le pourcentage de bonnes réponses fournies par notre système de questions-réponses sur les questions factuelles de EQUER (voir p. 161, pour les résultats globaux). On constate que, pour la plupart des cas, une bonne réponse est obtenue pour environ 57 % des questions, ce qui ne permet pas de prendre une décision sur le niveau de difficulté en fonction du type de réponse cherchée.

Type de réponse cherchée	Nombre de questions correspondantes	Nombre de réponses correctes	% de réponses correctes
Lieu	76	43	56,6 %
Personne	102	62	60,8 %
Organisation	18	8	44,4 %
Date	41	24	58,5 %
Numérique	81	48	59,3 %

TAB. 1.2: *Capacité de SQuaLIA à répondre à certains types de questions factuelles (EQUER).*

Prédiction pour requêtes *ad-hoc* La prédiction de la capacité à répondre à des requêtes *ad hoc* est un axe de recherche fort. Il a fait l'objet d'un atelier lors de la conférence

internationale SIGIR en 2005 et de la tâche *Robust* de la campagne TREC (Voorhees Ellen, 2003). Les premiers critères de prédiction dégagés se basent sur des caractéristiques de la requête uniquement. C'est le cas de la méthode d'évaluation de la difficulté des requêtes que nous avions proposée (de Loupy et Bellot, 2000). Les travaux de Cronen-Townsend et al. (2002) se fondent sur un calcul du taux d'ambiguïté de la requête. Dans le même esprit, Mothe et Tanguy (2005) ont montré que les corrélations entre des caractéristiques linguistiques de la requête et la capacité des systèmes participant à TREC 3, 5, 6 et 7 se situent uniquement au niveau de la complexité syntaxique (distance entre les mots syntaxiquement liés) et de la polysémie, écartant ainsi l'utilisation de certains types de mots pour déterminer la difficulté d'une question (acronymes, noms propres, conjonctions, mots suffixés,...).

Plus récemment, l'utilisation des documents retournés a permis de dégager de nouvelles caractéristiques pour évaluer la difficulté. C'est le cas de Amati et al. (2004) qui étudient la répartition des mots de la requête dans les premiers documents retournés. Kwok (2005) utilise une régression à l'aide des SVM sur des critères appris sur les résultats de son système, qui sont la répartition moyenne de la fréquence des mots de la question et leurs idf. Cette piste a été étudiée au LIA par Jens Grivolla dans sa thèse de Doctorat : Grivolla et al. (2005) utilisent des classifieurs qui se fondent sur l'apprentissage de caractéristiques issues à la fois des questions et des documents retournés. Nous avons adapté cette dernière méthode à la prédiction de la capacité à répondre à une question précise (par opposition à la recherche *ad-hoc*). Nous voyons ce problème comme une tâche de classification en deux classes : les questions faciles et les questions difficiles. Naturellement, une classification plus fine pourrait être envisagée.

Classification des questions en 2 classes : faciles et difficiles

Les algorithmes d'apprentissage utilisés pour entraîner les classifieurs sont les arbres de décision, les machines à vecteurs supports (SVM) (Burges, 1998) et la régression logistique à l'aide de modèles à maximum d'entropie. Les deux classes définies sont : "facile" (un nombre de réponses correctes supérieur à 0) et "difficile" (pas de réponses correctes pour la classe).

Trois types d'attributs ont été utilisés :

- les mots des questions, leurs lemmes, filtrés ou non, les étiquettes morpho-syntaxiques et sémantiques (entités nommées) ainsi que le type de réponse attendue ;
- la cohésion lexicale entre les documents d'où sont extraits les passages et entre les passages eux-mêmes : scores de similarités entre les 5, 10, 15 et 20 premiers passages. Cette idée part du principe qu'une forte variabilité linguistique est corrélée avec un plus grand risque qu'une partie des documents ou passages utilisés conduisent à des réponses fausses. Nous avons intégré cette notion même si, dans le cas de questions-réponses, il arrive fréquemment que la réponse n'apparaisse que dans un tout petit nombre de documents ;
- les scores de densité qui ont permis de sélectionner les passages (section 1.3.2, p. 36), la moyenne sur les 5 à 50 premiers passages, l'écart type et la valeur à un rang donné.

Le tableau 1.3 montre les résultats des trois classifieurs selon une validation croisée (*10-fold cross-validation*. La première ligne correspond à une prédiction uniquement basée sur la distribution constatée des questions faciles et difficiles, c'est à dire qu'elle classe toutes les questions dans la classe d'apprentissage majoritaire, ici la classe des questions ayant reçu une réponse correcte. Cette classification est considérée comme la prédiction de base.

***Objets* considérés**	**SVMs**	**Arbres de décision**	**Régression logistique**
Aucun (*baseline*)	51,6 %		
Questions	69 %	64,1 %	68,5 %
Passages	52,3 %	50,1 %	–
Documents	53,8 %	49,9 %	–
Tous	68,5 %	62,4 %	**71 %**

TAB. *1.3:* *Prédiction de la difficulté d'une question - Taux de réussite dans la prédiction de la capacité de SQuaLIA à répondre correctement aux questions EQUER.*

Les résultats montrent que l'utilisation des seules caractéristiques des questions, quelle que soit la méthode de classification, est assez efficace (ligne "*Questions*" dans le tableau). De plus la classification avec les SVM donne de très bons résultats bien supérieurs à la *baseline*. On peut supposer que l'apport très pauvre, voire négatif, des attributs reliés aux similarités entre documents et passages, et aux scores de densité des passages, est dû au fait que ce sont justement ces mesures qui sont utilisées par le système pour extraire les réponses. Si ces indices étaient fiables, il ne serait pas utile de chercher à estimer la capacité du système à répondre : le score parlerait de lui-même.

Pour plus d'informations sur cette section, voir (Gillard et al., 2007a; Sitbon, 2007; Sitbon et al., 2007).

1.1.3 Identification de questions complexes au sein de courriers électroniques

Afin de pouvoir être utilisé dans des conditions réelles, un moteur de questions-réponses doit être suffisamment robuste pour prendre correctement en considération les questions posées par des utilisateurs. L'aspect de ce problème lié à l'orthographe même des questions constitue l'une des parties de la thèse de Laurianne Sitbon (section 4.5, p. 108). Dans le cadre du Master Recherche de Nicolas Flavier (Flavier, 2006), nous nous sommes intéressé au cas où le besoin en information est complexe et ne peut être réduit en une seule phrase interrogative. De très nombreuses applications sont concernées par des questions où les problèmes doivent être contextualisés : centres d'appels, centres de S.A.V., consultation de bases de F.A.Q. (*Frequently Asked Questions*) ou de listes de diffusion, routage automatique de courriers électroniques... Grâce à une collaboration avec l'association Coridys[7] (*Coordination des intervenants auprès des personnes souffrant de dysfonctionnements neuropsychologiques*), nous avons pu disposer d'un corpus

[7] http://www.coridys.asso.fr/

de 92 courriels anonymisés provenant de personnes à la recherche d'informations générales sur tel ou tel trouble du langage ou bien demandant des réponses très précises à des besoins clairement définis. La figure 1.3 montre un exemple de courriel.

bonsoir,
Mon fils âgé de 9 ans est dyslexique, dysorthographique phonologique sévère. Il est actuellement en CM1 et il n'a jamais redoublé. Il voit son orthophoniste deux fois par semaine. Et il va chez une neuropédiatre au CHU de Rouen 1 fois par an, ainsi qu'une orthophoniste et un psychologue.
Je voudrais savoir si on peut mettre en place un PIS pour sa rentrée en CM2 car à l'heure actuelle il est vraiment en retard au niveau de l'orhographe écrit (non oral).
Merci pour l'aide que vous pourrez m'apporter.
J'habite à Montivilliers près du Havre dans le département 76.

FIG. 1.3: *Exemple de courrier électonique issu du corpus constitué avec l'aide de l'association Coridys et utilisé pour l'identification de questions. Une question booléenne (en gras) est identifiable.*

Afin de pouvoir répondre automatiquement à ce type de question complexe, l'on se doit d'identifier, au sein des courriels, les phrases définissant le contexte ainsi que la ou les questions posées. Nous avons effectué manuellement une correction orthographique puis nous avons annoté chaque phrase selon qu'il s'agissait d'une question ou de son contexte (119 questions pour 617 phrases contextuelles). On distingue plusieurs types de questions :

1. classique : *Quelles sont les différences entre ces établissements et une CLISS pour trouble du langage écrit et oral ?*
2. indirecte : *Je me demande aussi si le passage en classe ordinaire apres plusieurs années en ecole specialisée est envisageable.* ou : *Nous aimerions savoir si le personnel de ce centre est formé pour accueillir les enfants atteints de troubles tels que ceux de notre fils.*
3. fausses questions (au sens où elles n'appellent pas vraiment de réponse de la part de la personne qui les reçoit) : *Il apparaît que rien n'a été institué et qu'on ne sait pas ce que vont devenir ces enfants à la fin de l'année : retour dans leur collège de secteur ? continuation de ce dispositif en cinquième ?*

Nous avons cherché à identifier automatiquement les questions grâce à plusieurs méthodes d'apprentissage automatique implémentées dans l'environnement WEKA[8] (Witten et Frank, 2005) : J48 (arbres de décision basés sur C4.5 Quinlan, 1992), AdaBoostM1 (*Adaptive Boosting* associé avec un seul classifieur J48 fonctionnant de manière itérative, Meir et Ratsch, 2003), BayesNet (à base de réseaux bayésiens, Friedman et

[8] http://www.cs.waikato.ac.nz/ml/weka/.

Mots	Méthode	Rappel	Précision	F-mesure
non lemmatisés	J48	0,647	0,794	0,713
	AdaBoost	0,647	0,794	0,713
	BayesNet	0,462	0,743	0,570
	SMO (SVM)	**0,714**	**0,817**	**0,762**
lemmatisés	J48	0,597	0,67	0,631
	AdaBoost	0,672	0,734	0,702
	BayesNet	0,487	0,586	0,532
	SMO (SVM)	0,647	0,688	0,667

TAB. *1.4:* *Evaluation de l'identification automatique de questions dans des courriers électroniques à partir de quatre méthodes d'apprentissage automatique. Les meilleurs résultats sont obtenus sans lemmatisation en utilisant la méthode SMO à base de machine à vecteurs supports.*

Goldzsmidt, 1996) et SMO (*Sequential Minimal Optimization* basé sur des machines à vecteurs supports).

Le tableau 1.4 montre les résultats obtenus par validation croisée : même si les différences sont faibles, la meilleure performance est atteinte sur la version non lemmatisée du corpus. Le classifieur basé sur les SVMs a obtenu une F-mesure de 0,762 (supérieure de 0,059 à celle du classifieur utilisant une méthode de *boosting*). Sur la version lemmatisée, la méthode de *boosting* utilisant J48 a obtenu les meilleures performances avec une F-mesure de 0,702 légèrement supérieure de 0,035 à celle du classifieur basé sur les SVMs.

La prise en compte des étiquettes morpho-syntaxiques parmi les attributs de classification a fortement dégradé ces résultats. Cela ne fait que confirmer, s'il en était besoin, que le fait qu'une phrase contienne ou non un verbe, un substantif ou un nom propre n'est pas un critère permettant d'identifier une question. Si les pronoms sont un indice important, encore faut-il distinguer, par exemple pour *qui*, le pronom interrogatif du pronom relatif.

Ces résultats, ainsi que ceux concernant la recherche de questions similaires (voir 1.1.4) sont présentés dans (Flavier et Bellot, 2007).

1.1.4 La recherche de questions similaires

La recherche de questions similaires et, plus largement, la gestion de l'historique d'utilisation, constitue une piste de recherche intéressante (voir à ce sujet Poibeau et Vilnat, 2008). Durant la campagne TREC-9, le jeu de questions de test contenait 193 variantes de 54 questions (Voorhees, 2000)[9]. Il peut s'agir de variantes plus ou moins éloignées comme par exemple « *What is the moral status of human cloning ?* » et « *What are the ethical issues for human cloning ?* » mais aussi de reformulations telles que « *What attracts tourists in Reims ?* » et « *What are tourist attractions in Reims ?* ». Malheureusement,

[9] disponibles à l'adresse `http://trec.nist.gov/data/qa/t9_qadata.html`.

peu de travaux spécifiques ont pris cette caractéristique en compte durant la campagne d'évaluation. Notons tout de même Harabagiu et al. (2001) qui vérifient, pour chaque nouvelle question, s'il existe dans l'historique une ou plusieurs reformulations en calculant une similarité avec les questions précédemment posées à partir des couples (mot, étiquette morpho-syntaxique du mot). Si c'est le cas, la réponse associée à la classe des reformulations trouvées est retournée. Dans le cas contraire, la question est traitée de manière classique. Cependant, pour être réellement opérantes, de telles approches nécessitent probablement une analyse en profondeur des questions et une bonne prise en compte du contexte. Pensons à des questions telles que "*Où se trouve la tour Eiffel ?*", "*À quel endroit peut-on voir la tour Eiffel ?*" et "*Dans quelle ville se trouve la tour Eiffel ?*". S'il s'agit de trois formulations que l'on peut estimer très proches les unes des autres, il sera possible de répondre "*France*" (ou bien "*Europe*" pour un utilisateur résidant dans un pays très éloigné) aux deux premières mais pas à la dernière où "*Paris*" est la seule réponse acceptable.

Depuis, Tomuro et Lytinen (2004) ont proposé de construire un système de questions-réponses à partir de bases de FAQs. Ils ont employé conjointement plusieurs critères pour mesurer l'appariement d'une question avec une autre : nombre de mots communs et cosinus, distances entre les mots par rapport aux synsets de Wordnet et identification du type de réponse cherchée. Dans une même optique, ils ont défini des patrons de réécriture permettant de « normaliser » les questions par extraction de leur forme canonique grâce à l'usage d'un analyseur syntaxique : passage de la voie passive à la voie active, réécritures de formes interrogatives similaires par l'usage d'un pronom interrogatif unique *etc.* (Lytinen et Tomuro, 2002; Tomuro, 2003). Une manière de réduire l'espace de recherche de questions similaires consiste à identifier le type de réponse cherchée. Il s'agit alors de ne calculer de similarités avec des réécritures candidates que dans le cas où elles attendent un type de réponse similaire.

Dans le cadre de son Master Recherche, Nicolas Flavier a choisi de se concentrer sur des mesures de similarité classiques (taux de recouvrement, cosinus avec pondération $tf.idf$[10]). Une mesure de similarité a été calculée pour toutes les questions prises deux à deux puis ces scores ont été ordonnés de manière décroissante et une *moyenne de rang inverse* estimée[11]. Ensuite, les paires de questions ayant une similarité supérieure à un seuil ont été retenues pour être classées hiérarchiquement et permettre alors la

[10] Les composantes idf ont été calculées à partir du corpus Frantext de l'ATILF.

[11] Par exemple, si nous avons, pour 4 questions a, b, c et d et pour une mesure de similarité sim donnée :

$$sim(a,b) > sim(a,d) > sim(a,c)$$
$$sim(b,c) > sim(b,a) > sim(b,d)$$
$$sim(c,a) > sim(c,b) > sim(c,b)$$
$$sim(d,a) > sim(d,b) > sim(d,c)$$

Alors, pour a, le rang de b est 1, celui de c est 3 et celui de d est 2, et l'on obtient la *moyenne de rang inverse* sim_u :

$$sim_u(a,b) = 1/2 \cdot (1/1 + 1/2) = 0,75$$
$$sim_u(a,c) = 1/2 \cdot (1/3 + 1/1) = 0,67...$$

constitution de classes d'équivalences (rééécritures).

Sur le jeu des questions de la piste QA de TREC-9, l'évaluation des classes s'est faite à partir de plusieurs indices : la proportion de paires correctes trouvées au cours de l'appariement, la F-Mesure globale, l'indice de Rand et le taux d'exactitude. Les tests sur les jeux de questions de TREC 9 montrent que l'appariement utilisant une mesure de type cosinus avec $tf.idf$ permet de retrouver 49,3 % des réécritures avec une F-mesure relativement faible (0,29) ou bien un rappel de 33,6 % avec une F-mesure de 0,48. La figure 1.4 montre un exemple de l'application de cette technique d'appariement sur les courriels reçus par l'association Coridys après identification des questions (section 1.1.3).

– *Bonjour, à votre connaissance existe-il des centres ou des organismes - en région parisienne - dédiés à la dysorthographie adulte ?*
– *Mais où puis-je renseigner concernant les adultes car je trouve des noms et adresses de pédopsychiatre, spécialiste des enfants, adolescents mais rien concernant les adultes.*
– *Bonjour, j'aimerais savoir si il existe sur Paris ou sa région des centres d'aide aux adultes dyslexiques.*
– *Bonjour avez-vous de la documentation pour les adultes dyslexiques ?*

FIG. 1.4: *Exemple de questions extraites automatiquement de courriers électoniques qui ont été regroupées en une seule classe.*

Il est certain que des approches d'analyse plus profondes doivent être mis en place pour aller au-delà de ces premiers résultats. Nos travaux dans le cadre du projet ANR PIITHIE sur la détection de plagiats vont dans ce sens (chapitre 2.3, p. 64).

1.2 Questions-réponses et Recherche d'informations

Une différence essentielle entre la recherche d'informations (RI) et questions-réponses réside dans la nature de la requête : dans un cas, RI, il s'agit d'une description de l'information recherchée (qui contient bien souvent les mots des documents qui intéressent l'utilisateur, i.e. une forme de *réponse*) et dans l'autre cas d'une question qui ne contient justement pas les mots recherchés (la réponse). Le problème sous-jacent est alors : "*A quel point a-t-on besoin d'approches spécifiques de recherche d'informations pour questions-réponses ?*" (Bellot et Boughanem, 2008).

1.2.1 Les principaux modèles de recherche d'informations

Les modèles de recherche d'informations peuvent intervenir durant au moins deux des étapes d'un processus de questions-réponses : la recherche de documents pouvant

contenir une réponse et la recherche de passages au sein de ces documents (Cui et al., 2005; Kaszkiel et Zobel, 1997; Tellex et al., 2003).

La recherche documentaire en texte intégral se décompose quant à elle en au moins deux parties, précédées par une série de pré-traitements (parmi lesquels nous pouvons trouver un étiquetage morpho-syntaxique, une lemmatisation *etc.*) :

1. l'indexation des documents de la collection complète : l'indexation est un processus permettant d'extraire à partir d'un document les unités d'indexation (termes) ciblées (mots, couples de mots, syntagmes *etc.*). Cette étape produit pour chaque document une liste d'entrées caractérisant son contenu. Ces entrées sont alors regroupées dans une structure appelée fichier inverse, dans laquelle chaque entrée est reliée à la liste des documents dans lesquels elle apparaît, associée à son poids dans chacun d'eux (fréquence d'apparition, pondérations de type TF.IDF, ...) ;
2. la recherche de documents proprement dite, à partir d'une requête exprimée sous forme booléenne ou bien écrite en langage naturel : il s'agit de mesurer, grâce aux informations enregistrées dans l'index, la ressemblance – le score – de chaque document vis à vis de la requête. Les documents trouvés sont ordonnés en fonction de leur score et sont proposés à l'utilisateur ou, dans le cas d'un moteur de questions-réponses, au module de traitement suivant : segmenteur pour isoler plus finement des parties de document en fonction des questions ou, directement, extracteur de réponses. Notons que le segmenteur peut lui-même être un moteur de recherche documentaire où les unités recherchées ne sont plus les documents mais des paragraphes ou bien des séries de phrases contiguës.

Un modèle de RI a pour but de fournir une formalisation du processus de recherche d'informations. Il doit accomplir plusieurs rôles dont le plus important est de fournir un cadre théorique pour la modélisation de la mesure de pertinence. De nombreux modèles ont été proposés[12] depuis les années 1960 parmi lesquels citons le modèle booléen, le modèle vectoriel, le modèle probabiliste (Robertson et Sparck-Jones, 1976), le modèle booléen étendu (Salton et al., 1983), le modèle vectoriel étendu basé sur l'indexation sémantique latente (LSI - Deerwester et al., 1990), le modèle probabiliste à base de modèles de langage (Ponte et Croft, 1998; Boughanem et al., 2004a) et le modèle connexionniste (Kwok, 1989; Belew, 1989; Kwok, 1995; Boughanem et Soulé-Dupuy, 1997).

1.2.2 Evaluation du module de recherche d'informations présent dans les moteurs QR

Il faut différencier l'évaluation intrinsèque de l'étape de RI dans questions-réponses (les documents trouvés contiennent-ils ou non une réponse correcte dans le contexte de la question ?) de l'évaluation globale de la tâche (entre deux systèmes de RI, lequel

[12]Une description détaillée de quelques uns de ces modèles est donnée en Annexes : p.141 pour le modèle vectoriel, p.145 pour le modèle probabiliste. Sinon le lecteur peut se reporter par exemple à Baeza-Yates (1999); Savoy (2003) ou bien au chapitre que M. Boughanem et moi-même avons écrit dans l'ouvrage collectif sur "la recherche d'informations précises" dirigé par B. Grau et J.P. Chevallet : Bellot et Boughanem (2008).

permet au final de mieux répondre aux questions ?). Mais posons en premier lieu le problème du référentiel d'évaluation.

Evaluation stricte *vs.* évaluation tolérante

Si la question du référentiel sur lequel évaluer les systèmes était déjà cruciale en RI, elle l'est d'autant plus en questions-réponses où c'est un couple qui doit être trouvé : réponse correcte et documents supports dans lesquels une expression équivalente de la réponse est présente et apparaît dans un contexte identique à celui de la question (exemple de réponse : 2005 ; exemple de phrase où une information équivalente est trouvée : "*5 ans après l'an 2000*"). Le document support permet de vérifier la capacité du système à puiser les réponses dans les documents (par opposition à une base de données pré-construite) et à minimiser la chance d'obtenir par hasard la réponse. Dans un système réel, le document support permet à l'utilisateur de comprendre pourquoi une réponse lui est fournie. Construire un référentiel fiable en vue d'une évaluation automatique d'un système questions-réponses correspondrait non seulement à énumérer toutes les manières d'exprimer la réponse cherchée mais également à identifier tous les documents de la collection qui en contiennent l'expression en contexte. Notons qu'une *bonne réponse* n'est pas toujours une *réponse vraie* et peut être une réponse approximative. Il suffit de songer à des questions telles que "*Quel est le nombre d'habitants en Europe ?*" ou encore "*Quelle est la distance séparant Brest de Marseille ?*" pour deviner qu'une certaine marge d'erreur est tolérable.

On appelle "évaluation stricte" une évaluation où le référentiel est constitué de couples (réponse correcte – généralement une expression régulière –, document support) où les deux composantes sont prises en compte. On appelle "évaluation tolérante" (*lenient*) une évaluation où les réponses du système sont uniquement comparées aux expressions régulières du référentiel sans tenir compte du document support.

Nota Bene. La plupart des articles discutant de questions-réponses en dehors du cadre strict des campagnes d'évaluation utilisent ces deux modes d'évaluation comme un pis aller à défaut d'une évaluation manuelle très coûteuse. Malheureusement, plusieurs études, dont la nôtre — voir ci-dessous —, ont montré le peu de fiabilité que l'on pouvait accorder aussi bien aux évaluations strictes que tolérantes dans leur capacité à comparer *a posteriori* différentes approches (Lin, 2005; Voorhees, 2003). Plus loin, nous présentons les résultats officiels que nous avons obtenus au cours de campagnes d'évaluation mais également des résultats évalués par nos propres moyens. Ils sont donc à considérer avec les précautions nécessaires et, au-delà des chiffres bruts, c'est la tendance générale qu'il est important de retenir. L'importance de la significativité ou non des améliorations ou des dégradations observées doit être une préoccupation constante comme cela a été souligné et largement expérimenté par Savoy (2006); Abdou et Savoy (2007).

À l'issue de la campagne d'évaluation EQUER, nous avons travaillé avec l'aide des organisateurs[13] afin de mesurer l'impact d'une évaluation automatique par rap-

[13] Un grand merci à Christelle Ayache (ELDA) pour le temps qu'elle y a consacré.

port à une évaluation manuelle classique. Durant sa thèse, Laurent Gillard a constitué un ensemble de patrons représentatifs à 97,5 % des réponses correctes trouvées par les participants à partir des 19 708 réponses jugées. Après quelques modifications sur le fonctionnement de notre système censées l'améliorer, une évaluation automatique sur la référence précédemment élaborée a montré une chute significative des performances. Une évaluation manuelle, dans des conditions strictement identiques à l'évaluation officielle, a par contre souligné une amélioration nette de nos résultats. Cela prouve une nouvelle fois l'absolu manque de fiabilité d'une post-évaluation automatique en questions-réponses du moins tant que le nombre de systèmes, et donc d'expressions des réponses correctes dans les textes, à partir desquels les références sont construites, n'augmente pas très significativement. Ces résultats ont été présentés dans (Gillard et al., 2006b).

1.2.3 Quelques résultats

L'une des premières questions que l'on doit poser concerne la qualité du filtrage effectué par le module de recherche documentaire : les documents retenus contiennent-ils la bonne réponse à la question posée ? Pour la campagne d'évaluation TREC-9 (Voorhees, 2000), NIST et AT&T fournissaient pour chaque question la liste des 50 premiers documents trouvés par le moteur de recherche vectoriel SMART (Buckley, 1985) dans une collection de 979 000 documents. Cela permettait aux équipes participantes de ne pas avoir à travailler sur un module de recherche documentaire pour se concentrer sur l'extraction d'information. Prager et al. (1999) disposaient de leur propre moteur de recherche documentaire au sein de GuruQA[14], lui-même basé sur Guru (Brown et Chong, 1997). Ils ont ainsi pu comparer les résultats à partir des deux moteurs, SMART et Guru. Pour les deux systèmes, le MRDR valait 0,49 *i.e.* qu'en moyenne le premier document contenant la bonne réponse figurait en position 2. Avec GuruQA, une bonne réponse se trouvait parmi les 50 premiers documents pour 542 questions et parmi les 200 premiers documents pour 576 questions. Les résultats globaux sont sensiblement identiques avec SMART même si les documents trouvés ne sont pas toujours les mêmes : parmi les 682 documents contenant une bonne réponse et trouvés par l'un ou l'autre des systèmes (50 documents par question furent retenus), seuls 483 l'ont été par les deux systèmes simultanément. Ces différences se traduisent par des écarts de performance suivant les types de question. Il est cependant probable que ce ne soit pas tant les moteurs qui expliquent ces écarts que l'enrichissement des questions avec des informations sémantiques (type de la réponse cherchée, type des entités nommées rencontrées) qui a pu être réalisé sur les questions posées à GuruQA mais pas à AT&T SMART.

Pour la campagne d'évaluation TREC-2001 (Voorhees, 2001), NIST fournissait pour chaque question la liste des 1 000 premiers documents trouvés par le moteur de re-

[14]Les questions étaient posées au moteur GuruQA sous une forme lemmatisée et enrichies par le type de réponse cherchée (avec, en parallèle, un étiquetage en entités nommées des documents du corpus) ainsi que par des synonymes. Le moteur retrouvait alors des passages de une, deux ou trois phrases en leur associant un score *ad-hoc* basé sur les occurrences communes avec la question. Seul le meilleur passage étant retenu pour chaque document, cela induisait directement une liste de documents pouvant être comparée avec celle trouvée par AT&T avec SMART.

cherche PRISE sans garantie qu'ils contiennent les bonnes réponses (à comparer avec seulement 50 documents pour TREC-9 avec le moteur AT&T SMART). D'après l'analyse conduite par Litkowski (2002), une bonne réponse se trouvait dans les 10 premiers documents trouvés pour 311 des 500 questions de TREC-10. Pour 26 autres questions il fallait chercher entre le 11^e et le 20^e document et pour 32 autres questions entre le 21^e et le 50^e document. Parmi les 122 questions n'ayant pas de bonne réponse dans les 50 premiers documents, 49 n'avaient pas de réponse du tout parmi les 979 000 documents de la collection. Autrement dit, PRISE a échoué à classer des bons documents pour 73 questions sur 500 et il n'était guère utile d'aller chercher des réponses au-delà du 10^e document.

Toujours sur les données de TREC-2001, Brill et al. (2001) ont relevé que seulement 37 questions ont leur réponse dans au moins 25 documents différents de la collection et 138 questions dans au moins 10 documents différents. Cela implique que les stratégies de résolution des questions mettant en œuvre la redondance de l'information doivent être différentes sur des collections fermées comme celle de TREC et des corpus ouverts tels que le Web.

En exploitant le moteur de recherche Lucene (Cutting) pour la piste questions-réponses de CLEF 2006 (Gillard et al., 2007b), nous avons obtenu pour 38 des 156 questions factuelles, un document en rang 1 incluant la bonne réponse. Pour ces 38 questions, le score moyen du premier document retourné par Lucene est de 0,94 (écart-type 0,082) alors que le score moyen sur les 156 questions est de 0,92 (écart-type 0,118). La différence entre les scores obtenus ne permet pas d'en faire un critère décisif pour la sélection de la réponse.

De manière générale, la connaissance des performances d'un système de recherche documentaire aide à déterminer le seuil au-delà duquel il est inutile — ou trop risqué — d'aller chercher des réponses dans les documents mais il s'agit d'un problème ouvert.

1.3 Segmentation automatique et filtrage de passages

Le filtrage des passages de documents est l'étape qui suit habituellement la recherche documentaire au sein d'un processus questions-réponses. Cette étape peut se faire de manière non supervisée en employant des approches *ad-hoc* comme nous allons le détailler mais aussi en suivant des techniques d'apprentissage automatique de fonction d'ordonnancement (Usunier et al., 2008a).

La plupart des méthodes de segmentation sont destinées à la reconnaissance de frontières thématiques pour la catégorisation et pour la création automatique de résumés. Pourtant, grâce à une segmentation, *en fonction de la requête*, des documents rapportés par un moteur de recherche, certains d'entre eux, pertinents mais mal positionnés, peuvent être remontés en début de liste et ainsi récupérés plus facilement par l'utilisateur. Cela a été montré avec plus ou moins de succès entre autres dans (Salton et al., 1993; Callan, 1994; Kaszkiel et Zobel, 1997; Bellot et El-Bèze, 2001a; Liu et Croft, 2002; Wang et Si, 2008). Un tel procédé peut permettre d'augmenter la précision du système

de recherche sans diminuer le niveau global de rappel puisqu'il ne s'agit que de réordonner la liste des documents trouvés par le système et non d'en créer une nouvelle. Une segmentation thématique autorise enfin de présenter à l'utilisateur seulement la partie du document susceptible de l'intéresser.

La recherche de passage est une étape importante en questions-réponses puisqu'elle permet non seulement de réduire le champ d'extraction de la réponse à certaines parties des documents trouvés mais également de justifier aux yeux de l'utilisateur la réponse par un extrait de document (Lin et al., 2003).

Dans sa forme la plus simple, la recherche de passages correspond à l'extraction d'un certain nombre de phrases contiguës formant un bloc ayant une similarité particulièrement élevée avec la question. La similarité entre une question et un passage peut se calculer de la même manière que pour un document mais la plus petite taille du texte rend les mesures classiques moins performantes. Cela dit, il est intéressant de relever, à la vue des résultats de la campagne TREC-8, que la recherche de passages suffit à elle seule pour une tâche de questions réponses dans laquelle l'utilisateur ne demande pas une réponse précise mais se contente d'un extrait de texte de 250 caractères (Singhal et al., 2000; Voorhees et Harman, 2005).

1.3.1 Comparaison de différentes approches

Tellex et al. (2003) ont comparé huit méthodes de segmentation appliquées à la tâche questions-réponses sur les données de la campagne TREC-10 :

- *méthode 1* : classe les passages candidats en fonction du nombre de mots ou bien de stems qu'ils ont en commun avec la question ;
- *méthode 2* : calcule une similarité de type Okapi/BM25 (voir page 152) entre la question et les passages selon une fenêtre glissante sur les documents ;
- *méthode 3* : emploie une variante de celle employée dans le système MultiText (Clarke et al., 2000). Elle calcule des scores de densité en favorisant les passages courts contenant des mots rares dans les documents (idf élevées). Les passages retenus doivent commencer et se terminer par des mots de la question ;
- *méthode 4* : emploie une méthode proposée dans (Ittycheriah et al., 2001) qui associe à chaque passage potentiel une combinaison linéaire de scores calculés en fonction des poids (idf) des mots communs avec la question ou dont un synonyme est trouvé dans Wordnet mais aussi des poids des mots qui sont dans la question mais pas dans le passage, du nombre de mots communs et du nombre de mots communs adjacents à la fois dans le passage et dans la question ;
- *méthode 5* : inspirée par le système SiteQ (Lee et al., 2001), elle combine le score de plusieurs phrases adjacentes afin de trouver le meilleur passage (la longueur optimale trouvée est de trois phrases) ;
- *méthode 6* : emploie la méthode du système de l'Université d'Alicante (Vicedo et al., 2001) qui calcule le cosinus entre la question et les passages candidats. Les chercheurs d'Alicante trouvent une longueur de passage optimale de 20 phrases alors que Tellex et al. aboutissent à 6 phrases et nous à 3 sur CLEF 2006 (Gillard et al., 2007b) ;

– *méthode 7* : emploie la méthode de l'Université de Sud Californie (Hovy et al., 2001) qui ordonne les phrases en différenciant les noms propres des autres mots dans le calcul des scores mais aussi les mots trouvés en commun avec la requête (graphie identique) des mots dont seuls les stems sont en commun ;
– *méthode 8* : correspond à la fusion des résultats obtenus par les méthodes 4, 5 et 7 en fonction des rangs de chaque passage retenu et du nombre de passages extraits de chaque document par chacune des trois méthodes initiales.

Les expériences effectuées à partir de ces huit méthodes ont été réalisées d'une part avec le moteur de recherche documentaire Prise (les requêtes étaient les questions telles quelles), d'autre part avec le moteur Lucene en mode booléen (conjonction des mots des questions après élimination des mots outils) et enfin à partir d'un oracle ne retenant que les documents connus comme contenant une réponse correcte aux questions.

Chaque méthode pouvait retourner jusqu'à vingt passages par question à partir des 200 premiers documents trouvés (seul le meilleur passage d'un document a été retenu). Pour une évaluation stricte, avec Prise, le MRR varie entre 0,189 (méthode 1) ou 0,242 (méthode 1 avec stems) et 0,358 (méthode 5). Le pourcentage de questions pour lesquelles le système n'a pas su trouver de bonne réponse entre les rangs 1 et 5 varie entre 52 % (méthode 1) ou 58,6 % (méthode 1 avec stemmes) et 39,6 % (méthode 4). Avec Lucene, le MRR varie entre 0,271 (méthode 1) ou 0,25 (méthode 1 avec stemmes) et 0,354 (méthode 3) et le pourcentage de questions « incorrectes » entre 49,4 % (méthode 1) ou 52,6 % (méthode 1 avec stemmes) et 48 % (méthode 5).

Les différences entre les résultats obtenus par chacune des méthodes en employant le moteur de recherche Prise se sont montré statistiquement significatives contrairement à celles observées en employant le moteur Lucene. Les conclusions de Tellex et al. sont que l'emploi de méthodes de recherche de passages (plutôt que de recherche de documents seules) avec Lucene permet de mieux identifier les bonnes réponses (amélioration de la précision, hausse du MRR) tandis qu'il permet de répondre à plus de questions avec Prise (amélioration du rappel). Ils remarquent également que la méthode 4 n'est pas sensible au moteur de recherche documentaire employé en amont de la recherche de passages contrairement à la méthode 1.

Au final, les scores obtenus à partir de Lucene soulignent qu'une approche booléenne obtient d'aussi bonnes performances qu'une approche numérique avec Prise et que la prise en compte de la densité d'occurrences des mots de la question dans les passages (méthode 4) et de la différenciation des mots suivant leur nature syntaxique (méthode 7) conduisent à des améliorations significatives.

Ordonnancement des phrases. Au-delà de la définition de scores permettant de sélectionner des passages de documents, quelques auteurs ont proposé des mesures pour travailler sur les phrases elles-mêmes. Parmi eux, Radev et al. (2002, 2005) calculent, pour chaque phrase des documents trouvés, une combinaison linéaire des poids de chaque unigramme, bigramme et trigramme de la question dans la phrase. Ce modèle a donné de meilleurs résultats qu'une variante de la mesure Okapi (p. 152) appliquée aux documents sur TREC-8. Pour 132 questions sur 200, la réponse correcte se trouve

parmi les 20 premières phrases trouvées.

1.3.2 Proposition d'une mesure de densité pour la recherche de passages

Nous avons proposé (Gillard et al., 2005) de définir une notion de densité inspirée par les fonctions de score citées dans (Tellex et al., 2003) et dans (Ittycheriah et al., 2001). Comme elles, elle permet de choisir un passage en fonction des mots qu'il a en commun avec une question mais elle tient compte, en outre, des phrases adjacentes dans le document et de la présence ou non du type de réponse recherchée.

Nous avons défini un ensemble d'« objets caractéristiques » o_i, extraits d'une question q donnée afin d'aboutir à une reformulation enrichie de cette question. Cet ensemble est constitué des lemmes des mots (après élimination des mots outils), des types d'entités nommées (EN) présentes – noms propres, dates, lieux *etc.* –, et du/des type(s) de réponse attendue lorsque celle-ci est une entité nommée. L'idée est de favoriser les passages qui ont non seulement des traits en commun avec la question mais qui contiennent en outre des mots susceptibles d'être la réponse à la question. Cela illustre l'intégration possible dans le processus de RI de caractéristiques de la tâche questions-réponses (ici le type de réponse attendue). La forme enrichie de la question est appelée *requête* q' dans ce qui suit.

Nous proposons alors de définir une mesure qui identifie, dans chaque document, l'objet caractéristique autour duquel "gravitent", au plus près, le plus grand nombre d'autres objets caractéristiques de la requête.

Pour chacune des occurrences o_w des « objets caractéristiques » w rencontrés, à l'intérieur de chaque document, une distance moyenne $\mu(o_w)$, évaluée en « nombre d'objets », est calculée entre l'occurrence courante o_w et celles des autres objets caractéristiques de la requête, ou de leur plus proche occurrence en cas de présences multiples, au sein du document (voir les figures 1.5 et 1.6). Un **score de densité** Δ est attribué à chaque occurrence de chaque objet caractéristique du document selon la définition suivante :

Définition 1 (Densité)

$$\Delta(o_w, d) = \frac{\log\Big(1 + \mu(o_w) + p \cdot \big(|w| - |w, d|\big)\Big)}{|w|} \geqslant 0 \tag{1.1}$$

avec d un document, o_w une occurrence d'un objet caractéristique w (lemme, type d'entité nommée, type de réponse attendue), $p \geqslant 0$ une pénalité fixée empiriquement, $|w|$ le nombre des objets caractéristiques différents dans la requête q' et $|w, d|$ le nombre des objets caractéristiques appartenant à q' et présents dans d.

La densité d'une occurrence est ainsi estimée en fonction de la distance qui le sépare des autres objets caractéristiques, du nombre de ces objets dans la question, et enfin, du nombre d'objets communs entre la question et le document. La pénalité, fixée empiriquement, a pour rôle de plus ou moins favoriser une forte proximité de quelques

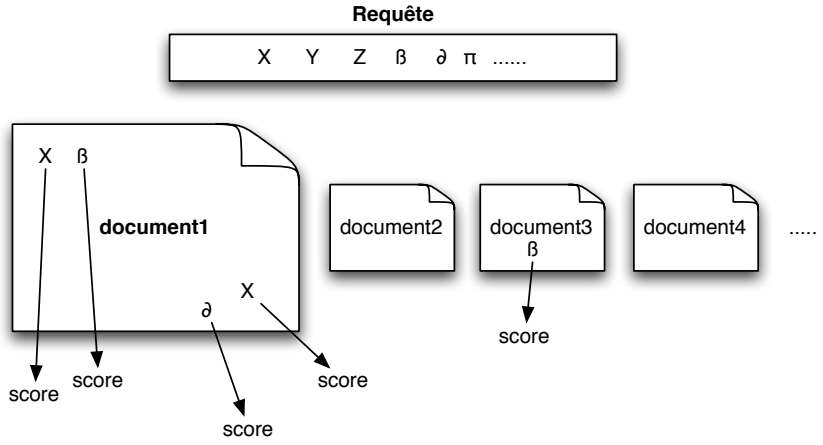


FIG. 1.5: *Un score de densité est calculé pour chaque occurrence de chaque objet caractéristique (mots, entités nommées, catégorie sémantique recherchée... tels que X ou β) de la question enrichie (*requête *q') au sein de chaque document.*

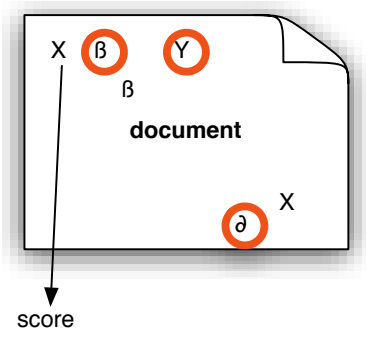


FIG. 1.6: *Le score d'un objet caractéristique (ici X) est déterminé en fonction de la position des occurrences des autres objets caractéristiques (ici β, Y et δ) de la requête q'.*

objets communs avec la requête par rapport à une proximité plus faible d'un plus grand nombre d'objets communs. Au contraire, lorsque tous les objets de la requête sont trouvés, la pénalité ne doit pas intervenir.

Un score σ est ensuite attribué à chaque phrase S comme étant le meilleur score[15] obtenu par une occurrence d'un objet caractéristique qu'elle contient :

$$\sigma(S,d) = \max_{o_w \in S} \Delta(o_w, d) \tag{1.2}$$

Chaque phrase est ensuite étendue à un « passage » constitué de la phrase qui la précède et de la phrase qui la suit (lorsqu'elles existent). Cela permet de compenser

[15] Une autre stratégie consisterait à considérer non pas le max mais la somme ou la moyenne des scores ainsi que le rapport entre la densité dans les documents et celle dans la question. En effet, et ce d'autant plus que cette dernière est longue, il paraît souhaitable de ne pas favoriser des passages qui contiennent des mots de la question (trop) éloignés dans cette dernière.

	Combinaison *cosine/okapi*		**Densité Δ pour segments d'une phrase**		**Densité Δ pour segments de 3 phrases**	
	Evaluation		Evaluation		Evaluation	
Rang	Stricte	Tolérante	Stricte	Tolérante	Stricte	Tolérante
1	0,380	0,425	0,397	0,455	0,435	0,512
5	0,477	0,525	0,480	0,538	0,512	0,585
10	0,485	0,532	0,487	0,547	0,520	0,593
50	0,493	0,538	0,493	0,552	0,526	0,598

TAB. *1.5:* *Evaluation sur les questions factuelles de la campagne EQueR de la densité par rapport à une combinaison cosinus/okapi — mesure MRR.*

quelque peu une éventuelle perte de contexte. Le score d'un passage est celui de sa phrase centrale.

1.3.3 Evaluation de la densité pour questions-réponses

Dans (Gillard et al., 2006a, 2007a, 2008) nous avons comparé, sur un sous-ensemble du corpus en français de la campagne Technolangue-EQUER, différentes méthodes de filtrage et d'extraction d'une réponse candidate dans le cadre de notre système de questions-réponses SQuaLIA[16].

Pour les 400 questions factuelles d'EQueR, nous avons évalué si les passages sélectionnés contenaient ou non les réponses recherchées. Dans le tableau 1.5, la colonne "Cosine-Okapi" correspond à la méthode que nous avons employée pour notre participation à TREC 11 (Bellot et al., 2003) (il s'agit d'une combinaison linéaire empirique entre un cosinus et une mesure de type *okapi*) et la dernière colonne "Densité sur 3 phrases" reprend les paramètres utilisés lors de notre participation à EQueR. Les valeurs numériques présentées dans ce tableau correspondent à la "Moyenne de l'inverse du rang" (MRR) des réponses contenues dans les meilleurs passages classés suivant l'ordre des différents scores, cela pour différents rangs.

Dans le cadre d'une évaluation stricte, et pour un unique passage accepté comme solution, la précision de la densité est meilleure de plus de 20 % que celle d'une combinaison Cosine+Okapi (cas où les segments ne font qu'une seule phrase). Par ailleurs, pour des segments d'au plus 3 phrases, notre meilleure valeur de MRR est pratiquement atteinte dès le rang 5. En allant jusqu'au rang 1 000, en évaluation tolérante, le MRR atteindrait 0,5988 au lieu de 0,5981 pour le rang 50, montrant l'inutilité de considérer plus de quelques dizaines de segments lors de l'extraction de la réponse. Notons qu'un test de Wilcoxon-Mann-Whitney permet de rejeter l'hypothèse que les séries d'observations, Cosine+Okapi et densité, sont issues de la même distribution avec un risque inférieur à 0,05.

[16] L'architecture du système de questions-réponses que nous avons développé avec Laurent Gillard et Marc El-Bèze est présenté dans la figure 1.7.

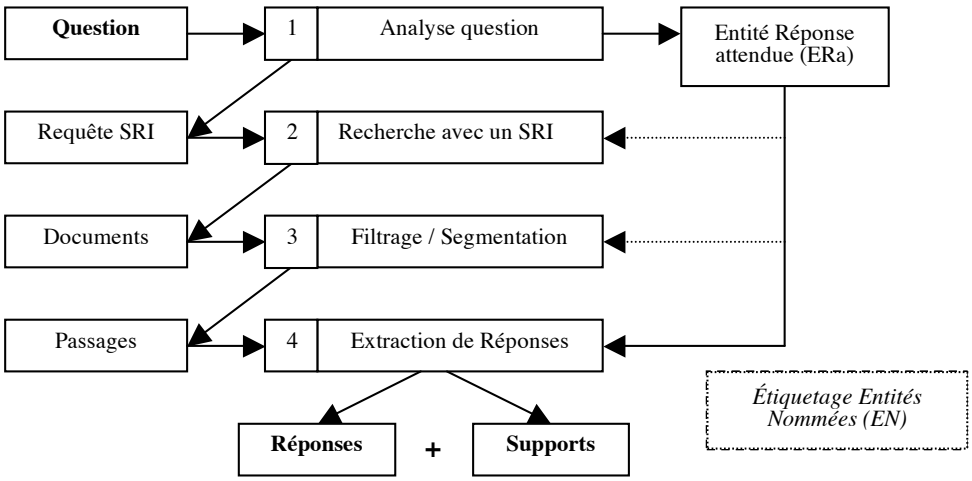


FIG. 1.7: *Le système de questions-réponses SQuaLIA présente une architecture en chaîne classique avec 4 principaux modules pour l'analyse de la question, la recherche de documents (SRI), la recherche de passages et enfin pour l'extraction de la ou des réponses. Le type de réponse attendue est exploité par les 3 derniers modules.*

En revanche, cette évaluation ne présume pas de la capacité du système à extraire les bonnes réponses lors de l'étape suivante comme nous le verrons dans la section 1.4. Une autre évaluation a été réalisée dans le cadre de la campagne de résumé automatique orienté requête DUC 2006 (voir section 3.4, p. 87) et durant la piste TREC-Enterprise en 2008 (à paraître).

1.3.4 Perspectives dans la recherche de passages

À notre connaissance, et contrairement à l'un de ses usages en recherche documentaire *ad-hoc*, la recherche de passages n'est pas utilisée en questions-réponses pour remettre en cause l'ordre des documents trouvés à l'étape précédente. La prise en compte de similarités locales permettant pourtant à certaines informations de ne pas être noyées dans la globalité du document et les réponses à des questions pouvant se trouver dans des documents dont la thématique générale est très éloignée de celle de la question, ce processus pourrait s'avérer payant. Une telle rétroaction (recherche de documents puis recherche de passages aboutissant au ré-ordonnancement des documents trouvés puis de nouveau recherche de passages) mériterait d'être expérimentée. Notons toutefois que (Rasolofo et Savoy, 2003) ont proposé de combiner une mesure de densité — qui tient compte des paires de mots sans nécessiter qu'ils soient adjacents — avec une similarité de type Okapi sur le document entier afin d'obtenir le score d'un passage.

Une autre perspective concerne l'une des critiques que l'on peut formuler à l'encontre des approches venant d'être présentées. En effet, ces dernières ne tiennent pas compte de l'ordre des mots dans les phrases et encore moins de leurs rôles syntaxiques et sémantiques. Plusieurs auteurs ont par exemple proposé d'utiliser des modèles de

langage probabilistes afin de tenir compte des relations pouvant exister entre les mots d'une question et favoriser ainsi les documents et les passages dans lesquels des relations identiques apparaissent (Gao et al., 2004).

Par exemple, dans la question extraite de la campagne d'évaluation francophone EQUER (Ayache et al., 2005) : «*Quand a été construite la maison d'arrêt de Fleury-Mérogis ?*», l'interrogation concerne la maison d'arrêt située à Fleury-Mérogis et non pas celle d'une autre ville. Cette propriété devrait être retrouvée dans les passages retenus. À défaut d'écarter ceux qui ne semblent pas vérifier cette contrainte de localisation (processus qui sera confié ultérieurement au module d'extraction de réponses), il faut privilégier ceux pour lesquels cela semble fort probable (proximité des mots *Fleury-Mérogis*, *maison d'arrêt* et d'une date dans le passage). Ce constat a été formulé dans Cui et al. (2005), où est montré l'apport significatif de cette approche par rapport aux méthodes uniquement lexicales.

1.4 Extraction de réponses pour des questions factuelles

Pour les questions factuelles, les réponses candidates sont *a priori* les entités nommées du type recherché qui se trouvent dans les premiers passages sélectionnés. Nous avons défini un score de *compacité* (Gillard et al., 2007a) permettant d'ordonner ces réponses candidates en fonction des occurrences des mots de la question dans leur voisinage, ce qui correspond à rechercher une pseudo-réécriture affirmative de la question incluant la réponse à l'intérieur des passages[17].

Le score de *compacité* est inspiré du CWS ("*Confidence Weighted Score*") (Voorhees, 2002) et de la notion de *précision*. L'idée est de considérer chaque occurrence d'une réponse candidate comme un point zéro d'un repère puis la présence des mots de la question dans son voisinage comme s'il s'agissait d'objets à trouver. Les mots apparaissant dans son voisinage mais qui ne sont pas présents dans la question sont considérés, dans le même esprit, comme des objets incorrects. Un critère fondé seulement sur un calcul reproduisant celui de la précision n'est pas satisfaisant : si une question contient n mots et qu'un seul est présent dans un passage, il aura tendance à attribuer un score plus important à un passage contenant un seul mot à côté de la réponse candidate qu'à un passage contenant outre ce mot (à la même position) d'autres mots de la question plus éloignés. Pour compenser cela, il suffit de modifier le critère de précision moyenne pour y introduire une partie de notion de *rappel* en tenant compte du nombre de mots différents de la question.

Définition 2 (Compacité) *Soient q' la question enrichie comme précédemment — $q' = \{o_i\}$, ensemble des objets caractéristiques — (section 1.3.2, p. 36), o_i un objet caractéristique (le plus souvent un mot), R_j une réponse candidate (une entité nommée), $\pi(o_i, R_j)$ la précision d'un*

[17] Ce qui revient à rechercher pour la question "*Quelle est la capitale de l'Ecosse ?*", une phrase de la forme "*la capitale de l'Ecosse est* [Nom de ville]" tout en autorisant certaines variations autour de cette réponse "triviale" par des insertions, inversions ou omissions même si certaines d'entre elles peuvent entraîner le système dans une mauvaise direction (par exemple l'insertion de la négation dans la phrase "*La capitale de l'Ecosse* n'*est* pas [Nom de ville]".

Filtrage du passage par	Sélection des réponses par	Evaluation			
		Stricte		Tolérante	
Cosinus $tf.idf$ / okapi	Mots communs	91	37 %	104	40 %
	Compacité	158	63 %	174	67 %
	Cosinus et compacité	172	69 %	185	71 %
Densité	Mots communs	90	37 %	104	41 %
	Compacité	159	65 %	175	69 %
	Densité et compacité	166	68 %	179	70 %

TAB. 1.6: *Evaluation sur les questions factuelles de la campagne EQueR en combinant ou non le score de compacité utilisé pour l'extraction de la réponse avec les scores des passages selon trois méthodes (le nombre de mots communs entre la question et le passage, cosinus $tf.idf$/okapi ou bien densité Δ). Le filtrage des passages est réalisé à partir de cosinus $tf.idf$/okapi ou bien à partir de la densité Δ. Les valeurs numériques correspondent au nombre de bonnes réponses (sur 400) et au pourcentage de bonnes réponses.*

o_i dans une fenêtre centrée sur R_j et d'un rayon ρ égale à la distance entre o_i et R_j. Avec $\mathbb{O}$ l'ensemble des $o_k \in q'$ à l'intérieur de cette fenêtre :

$$\pi(o_i, R_j) = \frac{|\mathbb{O}|}{2\rho + 1} \tag{1.3}$$

La compacité de R_j est :

$$compacité(R_j) = \frac{\sum_{o_i \in q'} \pi(o_i, R_j)}{|q'|} \tag{1.4}$$

Le tableau 1.4 montre que malgré les meilleures performances intrinsèques de la densité sur la similarité (tableau 1.5), nous obtenons au final des performances similaires en conservant la sélection de passages engendrée par la combinaison cosinus/okapi avec $tf.idf$ ou celle produite par la densité Δ (nous trouvons 158 bonnes réponses selon une évaluation stricte avec le cosinus et 159 avec la densité). Par contre, il semble plus profitable de combiner linéairement, au final, le score de compacité avec le cosinus plutôt qu'avec la densité pour choisir la réponse. Cela peut s'expliquer par le fait que la mesure de densité Δ est de nature plus proche de la mesure de compacité employée pour extraire les réponses des passages qu'une similarité de type cosinus. Nous trouvons 172 bonnes réponses selon une évaluation stricte en combinant avec le cosinus au lieu de 166 en combinant avec la densité. Dans tous les cas, une combinaison avec les scores des passages est pus performante que le score de compacité seul (158 ou 159 bonnes réponses sur 400 questions). Ces résultats sont assez naturels étant donnée la proximité des éléments pris en compte dans les scores de densité et compacité. Une version probabiliste de la mesure compacité a ensuite été proposée et présentée dans (Gillard et al., 2007b) qui a en outre étudié la complémentarité des différentes mesures.

1.5 Usage des bases de connaissance dans SQuaLIA

Au-delà de l'intérêt potentiel d'utiliser une base de connaissances (voir section suivante) pour répondre à des questions factuelles simples dont les réponses sont stables dans le temps plutôt que d'aller les chercher dans des textes[18], nous avons constitué une amorce de base de connaissances pour nos participations aux campagnes d'évaluation. L'intérêt était double. Il s'agissait d'une part d'exploiter ces connaissances pour tenter de distinguer une bonne réponse candidate (automatiquement extraite des documents du corpus) d'une mauvaise. Dans nos expériences, nous nous sommes limités à vérifier si la base contenait ou non la réponse à la question posée (par exemple la capitale d'un pays) et, si c'était le cas, le système sélectionnait cette réponse parmi les réponses candidates. Une base de connaissances pourrait être utilisée de manière intermédiaire en servant d'appui pour trouver la réponse à une question. Par exemple si l'on demande "*Combien y a-t-il de millions d'habitants à Bangkok ?*" (question EQUER GF134), la connaissance du fait que "Bangkok" désigne une ville de Thaïlande permet non seulement de *situer* la recherche mais aussi de borner la valeur à retrouver.

D'autre part, ces connaissances nous ont été utiles pour constituer des bases de patrons morpho-syntaxiques. L'idée étant de capturer le contexte d'apparition d'une information d'un type particulier dans un texte afin de pouvoir ultérieurement répondre à une question du même ordre. Cela aboutit à une liste de séquences que l'on pourra factoriser, et, en fonction des fréquences de leurs occurrences, déduire un ensemble d'expressions supports mettant en relation une personne et, par exemple, son année de naissance :

- `"(Madame|Monsieur) [...] est née en [année]"`
- `"La date de naissance de [...] est [année]"`
- `"[année], année où naquit [...]"` *etc.*

Lors de notre participation à TREC-11 (2002), le choix du contenu des bases de connaissances avait été établi d'après un examen rapide de l'ensemble des questions des campagnes TREC précédentes, en fonction :

- de la constance des thématiques d'interrogation, pour d'obtenir un taux de couverture suffisant sur l'ensemble des précédentes campagnes TREC-QA (en tablant sur la stabilité de ce taux d'une campagne à l'autre) ;
- du fait que la réponse attendue mette en jeu des associations entre entités nommées (par exemple : pays/capitale, inventeur/date/invention).

[18] Si le risque de se tromper est *a priori* moins grand d'utiliser une base de connaissances — indépendamment de l'exactitude des informations présentes — plutôt que de procéder à une fouille de textes, cela n'est ni sans risque ni forcément plus simple. La polysémie des questions, y compris celle des noms propres, peut entraîner des réponses inadéquates si la couverture de la base n'est pas suffisante (par exemple un même acronyme utilisé par deux organisations différentes). Voir Gillard et al. (2007a, p. 62) pour une discussion à ce sujet. Dans tous les cas, l'interrogation de bases de connaissances à partir de requêtes en langage naturel est un problème ouvert : détermination des couples (entités/valeurs) pertinents, choix de la bonne table à interroger...

Ainsi, une partie des bases de connaissances porte sur des lieux géographiques comme : des noms de cours d'eau et les pays traversés, des montagnes avec leur localisation et leurs points culminants, les capitales des pays ; mais aussi des personnages célèbres comme : les noms des différents prix Nobel par année et discipline, des noms d'écrivains accompagnés du titre de leurs œuvres, les noms des différents Dieux des panthéons grecs ou romains ; ou encore quelques (abréviation / définition). Le module tel qu'utilisé durant TREC 11 comprend 30 bases de connaissances thématiques et assure une couverture de l'ordre de 10 % à 12 % des questions des campagnes TREC 8 à TREC 11 (Gillard et al., 2003). Sur les 500 questions de la campagne TREC-11, 59 réponses pouvaient être obtenues depuis les bases de connaissances, mais 30 seulement ont pu être retrouvées dans des passages supports dont 24 ont donné lieu à une réponse correcte et justifiée.

Pour répondre aux questions de la campagne EQUER (2004), la couverture a été augmentée puisque le système disposait alors de 83 bases de connaissances. Ainsi, à titre d'exemple, il existe 6 relations ayant pour sujet « les animaux ». Elles permettent de traiter les équivalences nom du mâle, de la femelle et du petit (76 entrées) mais également leur vitesse de déplacement (66 entrées), leur longévité (46) ; le nom de leur cri (122), et la durée de gestation (30). À titre de comparaison, dans la version anglaise, seule la première base était présente. Enfin, la base « archives » était constituée des couples questions-réponses des précédentes campagnes CLEF en français. Sur un potentiel de 79 réponses effectivement présentes dans les bases de connaissances, 60 de ces réponses ont pu entrer en concordance avec les passages sélectionnés par le système et 53 l'ont été de manière correcte et justifiée. Parmi ces 53 réponses, 39 avaient été retrouvées *sans* l'usage des bases de connaissances. Autrement dit le gain par rapport à l'approche générique de notre système s'élève à 3,2 % sur les 433 questions de la campagne (voir Gillard et al., 2007a).

Jeu de questions (année)	**2004**	**2005**	**2006**
questions couvertes (/200)	51	26	24
patterns trouvés dans les passages	48	–	11
bonnes réponses	40	–	11

TAB. 1.7: *Taux de couverture des bases de connaissance sur les questions en français de CLEF (Gillard et al., 2007b). Des expressions régulières sont associées à chaque* connaissance *de la base afin de permettre l'extraction de phrases supports dans les passages des documents trouvés.*

1.6 Perspectives

La figure 1.8 souligne les progressions de notre système de questions-réponses au fur et à mesure des évaluations (% de bonnes réponses). SQuaLIA a répondu correctement à 52 questions sur 500 lors de TREC-11, à 139 sur 464 lors d' EQUER (puis à 235 sur 464 après corrections et une 2^e évaluation *in situ*) et à 93 sur 200 sur de CLEF 2006 sur la tâche FR-FR. Si les meilleurs systèmes se situent aux alentours de 80 %, ils sont très peu nombreux à atteindre de telles performances. En outre, sur le français, un seul

système a obtenu de meilleurs résultats que le notre lors des campagnes EQUER 2004 et CLEF 2006.

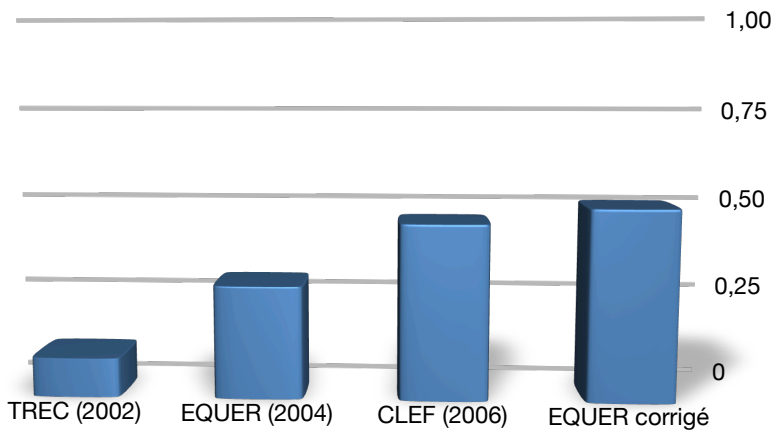


FIG. 1.8: *Progression des résultats de SQuaLIA entre 2002 et 2007 (% de bonnes réponses) sur les données des campagnes TREC, EQUER et CLEF FR-FR.*

Dans ce qui suit, nous évoquons deux des pistes qui nous semblent être les plus intéressantes pour des travaux futurs, à savoir l'intégration de connaissances linguistiques plus poussées et le développement de modèles numériques adaptés à la tâche d'une part et l'exploitation de bases de connaissances (ontologies, bases de données structurées ou non) pour des domaines de spécialité d'autre part.

1.6.1 Questions-réponses, bases de connaissances et domaine de spécialité

Un court état de l'art

Le fait de ne considérer qu'un seul domaine d'application, aussi vaste soit-il, offre l'avantage de pouvoir plus facilement disposer de bases de connaissances et de nomenclatures (Rinaldi et al., 2003) que dans un cadre universel tout en diminuant la difficulté de leur utilisation grâce à un niveau d'ambiguïté inférieur. Par ailleurs, les volumes de données sont théoriquement moindres, favorisant *a priori* les approches symboliques de Traitement Automatique des Langues par rapport aux approches numériques. Malgré tout, ce sont souvent plusieurs millions de documents qui restent à considérer. Un autre aspect à prendre en compte est le formatage des textes qui obéit plus souvent à des règles strictes, aussi bien pour le texte lui-même que par la présence quasi-systématique de formules mathématiques, de tableaux ou de références bibliographiques (ces dernières étant autant de liens entre les documents tout comme des hyperliens au sein de pages Web). Au final, le traitement de documents spécialisés est au moins aussi difficile que celui de documents dans des domaines ouverts (Zweigenbaum, 2003) ne serait-ce qu'à cause des problèmes de normalisation terminologique. Plusieurs auteurs soulignent en effet le rôle central de la terminologie en recherche d'information spécialisé en faisant le parallèle entre les termes dans un domaine spécifique et les entités

nommées en domaine non spécialisé : leur reconnaissance et leur utilisation est indispensable. En outre, dans un domaine spécialisé, l'analyse des besoins en information à partir de textes et formulés par des experts montre que les questions procédurales les plus complexes (du type « comment… », « pourquoi… ») sont plus nombreuses que les questions factuelles pour lesquelles les réponses sont facilement accessibles dans des bases de données.

La grande majorité des campagnes d'évaluation internationales concernent des domaines non spécialisés. À notre connaissance, hormis, pour la recherche *ad-hoc*, la piste Genomics de TREC, la piste Enterprise de TREC depuis 2006, la seule campagne d'évaluation de systèmes de questions-réponses spécialisée a eu lieu durant le projet Technolangue Equer[19] : la collection de textes était composée d'articles scientifiques et de recommandations de bonne pratique médicale (Jacquemart et Zweigenbaum, 2003; Malaisé et al., 2005; Zweigenbaum, 2005).

L'utilisation de bases de connaissance pour questions-réponses fait l'objet d'un nombre relativement faible de publications par rapport à ce que l'on pourrait attendre. Citons tout de même les thèses de Doctorat de Moriceau (2007) qui a essentiellement travaillé sur des données numériques et des propriétés de base de certains objets et de Embarek (2008) qui exploite des ressources terminologiques ou ontologiques dans le domaine médical. On peut aussi se reporter au chapitre de Ferret et Zweigenbaum (2008) dans (Grau et Chevallet, 2008). À l'inverse, Lin (2007a) discute de l'opportunité d'établir ou d'exploiter des bases de connaissances sachant l'exploitation possible qui peut être réalisée de la redondance informationnelle sur le web.

La notion de *connaissances* est souvent prise au sens large et peut être distinguée de la notion d'*informations* qui correspond à un ensemble de faits ou de caractéristiques connues d'un objet ou d'un être. Les connaissances peuvent alors désigner des ressources de variations linguistiques, des chaînes de déduction, des schémas syntaxiques *etc.* Rinaldi et al. (2004) discutent d'une extension vers un domaine spécialisé du moteur de recherche ExtrAns (Molla et al., 2000), en l'illustrant autour d'un corpus de manuels de maintenance d'un Airbus A320. Ils indiquent que l'utilisation d'une ressource sémantique telle que Wordnet a plutôt tendance à dégrader les résultats de la recherche qu'à l'améliorer (cela peut s'expliquer notamment par le fait que le sens de certains termes dans le domaine spécialisé n'est pas le même que dans le cadre général à partir duquel Wordnet a été construit). Enfin, l'existence de terminologies riches ne les rend pas moins complexes à utiliser et de nombreux experts n'hésitent pas à s'en écarter lors de l'interrogation du système.

Historiquement, les projets InfoSleuth[20] et TSIMMIS[21] ont ouvert la voie vers la fusion et l'agrégation d'informations issues de documents semi-structurés ou de bases de données. Le projet européen MESH[22] (*Multimedia Semantic Syndication for Enhan-*

[19] Sans être pour autant une évaluation spécialisée, la piste questions-réponses de CLEF 2007 présentait un certain nombre de questions liées à une même thématique.

[20] http://www.argreenhouse.com/InfoSleuth/

[21] http://infolab.stanford.edu/tsimmis/

[22] http://www.mesh-ip.eu/

ced News Services) concerne lui l'extraction d'informations et le résumé automatique personnalisé à partir de multiples sources d'actualités. En France, le projet RNTL 2006 DaFOE4App[23] s'attache à développer une plateforme technique de création et d'exploitation d'ontologies avec trois applications dans des domaines médicaux, d'indexation patrimoniale ou d'images satellitaires. D'autres projets tels WebContent[24] (RNTL 2005), e-WOK HUB[25] (RNTL 2005) ou VODEL[26] (RNTL 2005) concernent plus spécialement le Web sémantique.

Utilisation des ressources terminologiques et des ontologies dans les outils d'accès à l'information. Il s'agit d'une approche traditionnelle puisque les documentalistes gèrent depuis longtemps des thesaurus pour classer les documents et proposer des plans de classements aux utilisateurs. L'indexation de *PubMed* repose ainsi sur le thesaurus hiérarchique d'UMLS. À partir de là, de très nombreux travaux ont cherché à exploiter des ontologies ou des modèles conceptuels pour guider la recherche d'information. Selon les cas, on s'appuie sur l'ontologie pour proposer de reformuler la requête (généralisation ou spécialisation), pour modéliser le profil de l'utilisateur (Gauch et al., 2003) et affiner la recherche en tenant compte du contexte (Hernandez et al., 2006), pour proposer une cartographie sémantique des documents retrouvés et ainsi aider le lecteur à affiner sa requête (Hearst, 1999). Dans l'interface des moteurs de recherche spécialisés proposés dans le projet européen ALVIS, la prise en compte des types sémantiques lors de l'indexation permet de proposer à l'utilisateur un index conceptuel dans lequel il peut se repérer. C'est la même idée qui été reprise dans les systèmes d'aide à la navigation documentaire (Mondary et al., 2007). Pour une synthèse des méthodes d'utilisation et de construction de bases de connaissances pour la recherche d'informations, voir par exemple (Bruandet et Chevallet, 2003), et pour une présentation du modèle vectoriel DSIR incluant des connaissances sémantiques, se reporter à (Besançon et al., 2003).

Extraction terminologique à partir de corpus. De nombreuses expériences ont été menées utilisant des outils spécifiques (extracteurs de candidats-termes, système d'apprentissage de patrons lexicaux-syntaxiques, population d'ontologies (Bourigault et al., 2001). Staab et Maedche (2001) ont créé *Text-to-Onto* qui propose d'extraire des connaissances à partir de textes en combinant plusieurs outils de TAL dont la fouille de textes. La plateforme *Terminae* est à la fois une méthode et un outil pour construire des terminologies et/ou des ontologies à partir d'un corpus reflétant le domaine et l'application visés (Aussenac-Gilles et al., 2000). La méthode a été élaborée en collaboration entre le LIPN et l'IRIT. L'hypothèse linguistique est que la signification des termes est spécifique d'un domaine et peut être inférée à travers leurs usages dans le corpus. L'environnement permet de dépouiller les résultats des extracteurs de candidats termes comme *Syntex* (Bourigault et al., 2004) ou *Yatea* (Aubin et Hamon, 2006), d'explorer un corpus

[23] http://dafoe4app.fr/
[24] http://www.webcontent.fr
[25] http://www-sop.inria.fr/edelweiss/projects/ewok/
[26] http://vodel.insa-rouen.fr/

à l'aide de patrons lexicaux et/ou syntaxiques, de déterminer d'éventuels synonymes grâce à l'outil *SynoTerm* (Hamon et Nazarenko, 2001).

Reconnaissance des entités nommées. Il s'agit d'un point essentiel pour toute tâche d'extraction d'informations. Cette reconnaissance a été réalisée de façon satisfaisante pour des textes journalistiques anglais avec une précision et un rappel autour de 90 % lors des dernières conférences MUC (*Message Understanding Conference*) en 1998. Ces systèmes distinguent deux étapes dans la reconnaissance des EN : l'identification et la délimitation à droite et à gauche de l'EN, et sa catégorisation (voir l'état de l'art des techniques existantes dans (Daille et Morin, 2000)). Poibeau (2003) distingue trois types de systèmes de reconnaissance d'entités nommées : les systèmes à base de règles linguistiques, ceux à base d'apprentissage automatique et les systèmes mixtes.

Enrichissement de requêtes à partir de ressources externes. Il peut s'agir tout d'abord d'ajouter dans la requête les différentes écritures possibles des noms propres et des acronymes, de corriger (ajouter ?) certaines erreurs d'orthographe ou tout simplement d'ajouter/remplacer les caractères non standards (accents, majuscules). Ces traitements, parfois effectués sur les corpus eux-mêmes durant l'indexation, ne sont pas sans danger : ajout d'ambiguïtés, modification du sens de la requête, perte de précision de la recherche (éventuellement au profit d'un meilleur rappel). De nombreux auteurs ont proposé d'enrichir les requêtes à l'aide des informations sémantiques généralistes de Wordnet (Miller et al., 1990). Il n'a été malheureusement été constaté d'amélioration des résultats que dans des cas restreints (Baziz et al., 2005). Il est également montré qu'une amélioration par hyponymie améliore plus nettement les résultats que par synonymie. Une autre piste consiste à enrichir la requête avec des informations que l'on s'attend à trouver dans la réponse : par exemple les unités pour des questions appelant des réponses numériques (ajout de km par exemple pour une question portant sur une distance).

Le projet Ontofruit

Nous travaillons avec l'INRA d'Avignon sur un projet destiné à exploiter certaines connaissances expertes (Aventurier et al., 2008) dans la cadre d'un système de recherche documentaire et de questions-réponses pour un domaine scientifique spécifique, en l'occurrence l'écophysiologie végétale de l'arboriculture fruitière. La conduite de ce travail repose sur les compétences complémentaires d'un groupe constitué de chercheurs en écophysiologie, de documentalistes, d'ingénieurs en biométrie et d'informaticiens. La vocation première de ce travail est de répondre aux besoins exprimés par les chercheurs en matière de gestion et de représentation des connaissances. Il s'agit de prendre en compte sur le plan conceptuel la complexité des relations qui sous tendent la production des connaissances dans un domaine au contact de plusieurs disciplines (écophysiologie, génomique, modélisation...). L'apport du système réside dans une présentation contextualisée des données et des concepts qui permet l'accès au contenu notamment

par navigation interactive. Par ailleurs, le système en cours de développement à vocation à intégrer largement les connaissances produites dans le domaine, à partir, par exemple, de ressources bibliographiques indexées dans bases de données internationales.

Sur le plan scientifique, notre objectif est de proposer des modèles numériques efficaces permettant l'exploitation de plusieurs ontologies (ontologies de référence du domaine mais aussi ontologies propres à chaque utilisateur), sans pour autant chercher à les aligner, et tout en profitant des approches plein texte. À l'heure actuelle, le prototype développé permet d'annoter semi-automatiquement ou manuellement des documents (figure 1.9), de rechercher des concepts au sein d'ontologies et d'accéder au contenu des documents référencés (figure 1.10). Les systèmes de questions-réponses exploitant des raisonnements par inférence ont montré des performances de haut niveau dans les évaluations TREC (Moldovan et al., 2002, 2003). En domaine de spécialité, en l'occurrence le tourisme, un système de questions-réponses utilisant la logique et le raisonnement est présenté dans (Benamara et Saint Dizier, 2004). Ce type d'approche est très courant lorsqu'il s'agit d'exploiter des ontologies et des modèles de représentation de la connaissance. Néanmoins nous croyons qu'il est possible de s'affranchir, au moins en partie, des approches symboliques logiques ((Chevallet, 2004) *in* (Ihadjadene, 2004)) en faveur de modèles numériques tels que ceux qui semblent être à l'ouvrage dans le cerveau humain (Edelman, 2007). Les approches vectorielles ou connexionnistes (Boughanem et Tamine, 2004) déjà proposées pour l'intégration de connaissances dans les systèmes de recherche d'informations sont autant de pistes que nous allons exploiter désormais (Baziz et al., 2005; Castells et al., 2007; Baziz et al., 2007).

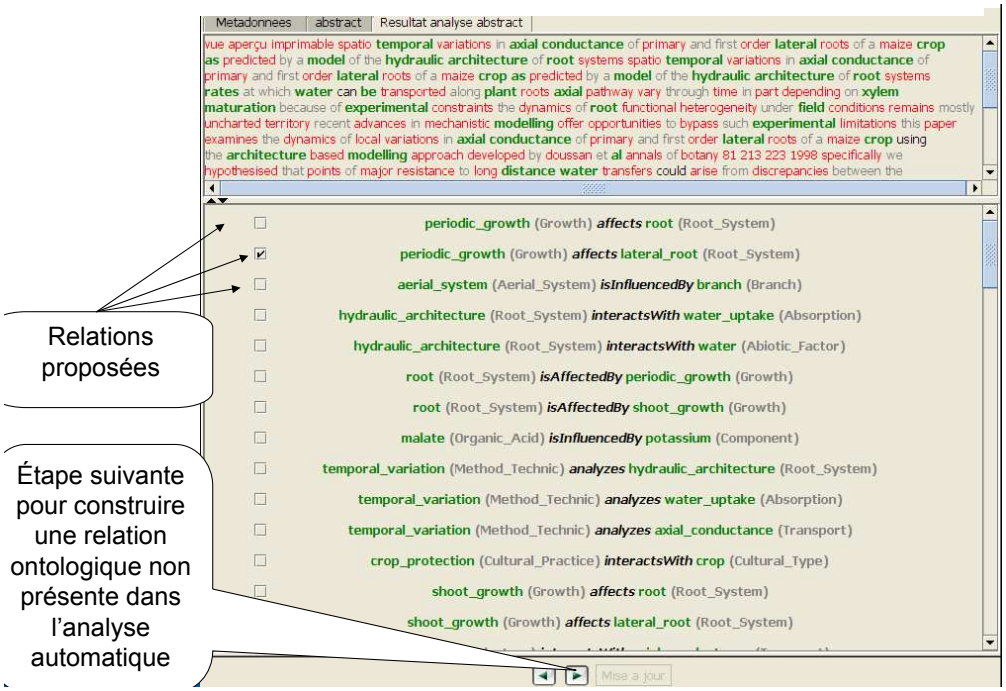


FIG. 1.9: *Le module d'annotations du projet Ontofruit développé avec l'INRA d'Avignon*

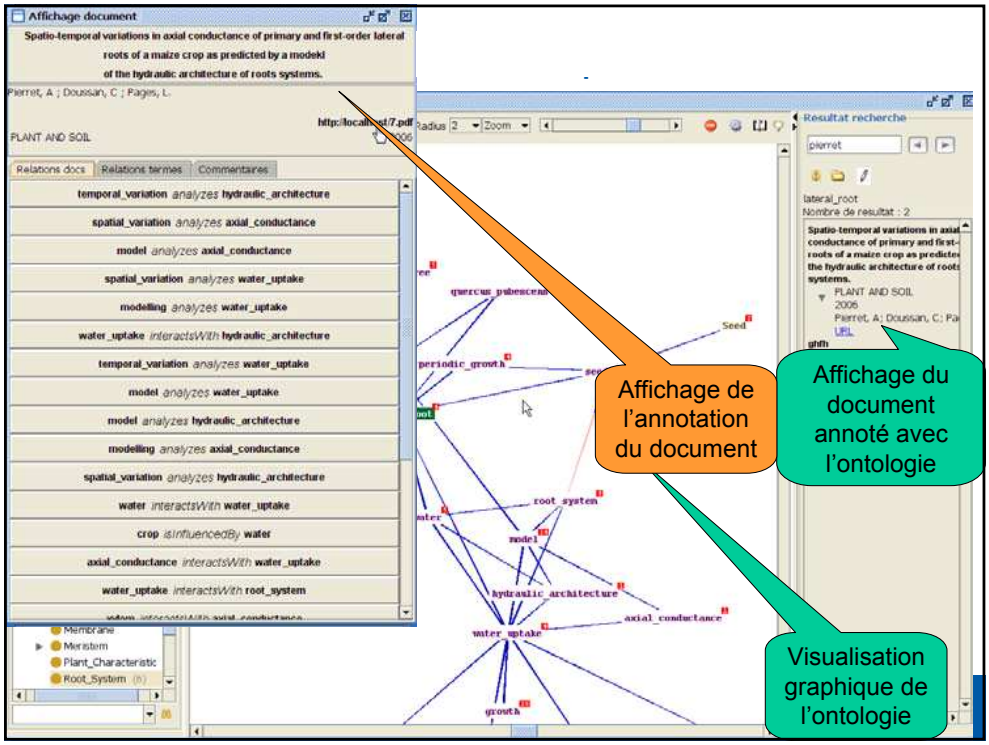


FIG. 1.10: *Le module de recherche du projet Ontofruit : accès par navigation dans l'ontologie (graphe en fond d'écran) ou bien par mot-clé.*

1.6.2 Questions-réponses entre RI et TAL

L'article que nous avons publié dans la Revue Française de Linguistique Appliquée au sujet de l'apport de la linguistique dans les systèmes de questions-réponses (Zweigenbaum et al., 2008) dresse la conclusion suivante[27] :

> [...] *Il ressort ainsi que du point de vue de l'ingénierie des systèmes de questions-réponses, là où des ressources linguistiques sont disponibles en taille et en précision suffisantes, les méthodes de type linguistique sont indiquées ; et qu'à l'inverse si des données vastes et redondantes permettent de trouver facilement la réponse, ou si elles permettent d'entraîner facilement un système à base d'apprentissage (Usunier et al., 2008b; Grau et Chevallet, 2008), on aurait tort de se passer de ces données et de ces méthodes. Les méthodes par apprentissage automatique fonctionnent cependant presque toutes après un premier niveau d'analyse linguistique symbolique, que ce soit pour segmenter le texte en mots, les lemmatiser ou même les étiqueter morphosyntaxiquement.*
>
> *La conclusion à laquelle on arrive est que les seules méthodes numériques ou par apprentissage restent limitées en dehors de situations où l'on a pléthore de données,*

[27] Pour des réponses concernant la même interrogation au sujet de la recherche d'informations, voir par exemple (Gaussier et al., 2003; Moreau et al., 2007).

> *et qu'il est donc très utile de préparer des ressources linguistiques (lexique, grammaire, cadres de sous-catégorisation, etc.). Mais on doit reconnaître en même temps que les approches linguistiques sont plus longues à mettre en place et peuvent être rejointes (voire dépassées) dans certaines situations, comme indiqué [...] à propos des questions non factuelles. Une voie différente peut alors consister à tourner les efforts linguistiques vers la caractérisation et l'annotation de données qui serviront à entraîner un système fonctionnant par apprentissage : un corpus annoté selon le niveau que l'on veut obtenir pour l'analyseur (catégories morphosyntaxiques, dépendances, etc.). À la limite, si la collection cible est très grande et redondante, il n'est plus nécessaire de faire autant d'efforts pour obtenir une analyse précise. Mais comme souvent, l'issue peut être dans la combinaison des méthodes, les approches linguistiques formant des parties sûres du traitement (voir par exemple Habert et Zweigenbaum, 2002), les méthodes numériques intervenant soit pour prendre en charge des parties dont la modélisation linguistique est absente ou incomplète, soit pour aider à acquérir les connaissances linguistiques nécessaires, soit encore pour aider à la prise de décision lorsqu'elle n'est pas fondée sur des conditions catégoriques auxquelles la modélisation par règles nous a habitués.*

Si les modèles de recherche d'informations (RI) habituels peuvent paraître suffisamment performants pour la tâche questions-réponses, des progrès restent souhaitables. Pour la partie sélection de documents, les approches « sac de mots » paraissent suffisantes — il est encore théoriquement possible de répondre à plus de 95 % des questions après filtrage des corpus à 1 000 documents par question —, mais la question se pose plus nettement dès que l'on arrive à la sélection de phrases. Sur les corpus en français de la campagne Technolangue-EQUER (Ayache et al., 2005), il n'est plus possible de répondre à environ 50 % des questions après avoir utilisé des méthodes d'extraction de passages de type « sac de mots » pures (Gillard et al., 2006a). En ce qui concerne la sélection de la réponse, des techniques basées sur la compacité des mots de la question autour de la réponse candidate permettent de répondre correctement à près de 40 % des questions factuelles EQUER sans ajout de connaissances linguistiques spécifiques ni de lexiques étendus (Gillard et al., 2006b). En moyenne, le deuxième document trouvé contient la bonne réponse mais celle-ci peut demeurer inaccessible au module d'extraction de réponse : il pourrait être avantageux d'aller la chercher plus loin, au sein de documents plus faciles. C'est bien là l'une des difficultés majeures de questions-réponses : le besoin d'utiliser des techniques d'extraction « en profondeur », techniques qui, lorsqu'elles sont disponibles et adaptées à l'expression des réponses candidates dans les documents de la collection, ne sont pas toujours applicables en des temps ou occupation mémoire raisonnables. Dans la pratique, limiter la recherche de la réponse aux tout premiers documents trouvés empêche d'utiliser la redondance comme critère d'extraction. À l'inverse, ne pas se limiter aux premiers documents peut entraîner la prise en considération d'un trop grand nombre de réponses potentielles et donc diminuer la précision du système.

Durant la campagne d'évaluation TREC 2006, une piste *complex interactive Question Answering* (ciQA) fut initiée (Dang et al., 2006). Les sujets de recherche (*topics*) contenaient deux champs : une question (« *What evidence is there for transport of drugs from Mexico to the U.S. ?* ») ainsi qu'une description textuelle décrivant le contexte de

la recherche ou des centres d'intérêt particuliers. À partir des questions, les systèmes devaient fournir des réponses, limitées en taille, accompagnées de formulaires Web libres, à transmettre aux assesseurs. Une fois remplis par ces derniers, les formulaires étaient retournés aux participants afin qu'ils fournissent des réponses affinées. À des fins de comparaisons, un système basique a considéré les questions telles quelles et les a soumis au moteur de recherche Lucene. Les phrases contenant au moins un mot de la question (mots outils exceptés) parmi les 20 premiers documents retournés ont ensuite été retenues. Le formulaire soumis aux assesseurs demandait alors de juger de la pertinence de chacune des phrases retenues (ce qui revenait à faire « remonter » des phrases ayant des scores plus faibles) et, au final, le système éliminait les phrases jugées « non pertinentes ». L'évaluation a montré que ce système de base arrivait en 3^e position parmi 13 alors que la meilleure soumission correspondait à un test manuel.

L'étude sur les données des campagnes TREC 2004 et 2005 conduite par Lin (2007b) montre que le moteur de recherche documentaire Lucene obtient en moyenne des résultats compétitifs sur les questions définitoires, ou qui appellent des réponses longues, par rapport aux autres moteurs de questions-réponses. La conclusion est que les techniques de traitement automatique des langues ont prouvé leur apport pour la tâche question-réponse factuelle mais ne sont pas encore vraiment à maturité pour les autres types de questions. Les approches traditionnelles de la RI s'en sortent donc plutôt bien (Kelly et Lin, 2007).

Il est souhaitable que le système de recherche d'informations soit capable de fournir l'ensemble des documents du corpus qui contiennent la réponse tout en respectant le contexte de la question. Dans cette optique, la marge de progression demeure réelle (des référentiels, utilisables sans une évaluation manuelle lourde, restent à construire) : amélioration de la précision et du rappel afin d'offrir plus de documents pertinents à explorer aux autres modules de la chaîne et variabilité de la granularité (du document aux passages). L'intégration des spécificités de questions-réponses dans les modèles de RI reste à réaliser avec comme double conséquence celle d'unifier, au moins en partie, les problématiques et, par ricochet, d'améliorer les performances des deux applications que sont la recherche documentaire et la recherche de réponses précises.

Chapitre 2

Segmentation, enrichissement de requêtes et détection de plagiats

Sommaire

2.1 Segmentation non supervisée par chaînes lexicales pondérées 53
2.1.1 Proposition à base de chaînes lexicales pondérées 54
2.1.2 Evaluation durant la campagne DEFT 2005 57
2.2 Segmentation pour l'expansion de requêtes 58
2.2.1 Expansion à partir de ressources externes. 58
2.2.2 Des arbres de décision non supervisés pour une segmentation thématique orientée requête 59
2.2.3 Proposition d'une méthode d'expansion à partir d'arbres de décision non supervisés 62
2.3 Détection de plagiats : le projet PIITHIE 64
2.3.1 Présentation 64
2.3.2 Etat de l'art 66
2.3.3 Détection de citations pour l'identification de plagiats 68
2.3.4 Segmentation de textes pour la détection de plagiats 72

2.1 Segmentation non supervisée par chaînes lexicales pondérées

Dans le cadre du projet Technolangue[1] AGILE-OURAL (2003-2005)[2], nous étions en charge de développer et d'évaluer un segmenteur thématique non supervisé de textes en français. Nous avons établi un état de l'art des méthodes de segmentation et d'évaluation de la segmentation que l'on retrouve en partie en Annexes et dans Sitbon et

[1] http://www.technolangue.net
[2] http://projetoural.org

Bellot (2004b). Parmi les outils que nous avons testés sur le français[3] figurent DotPlotting (Reynar, 2000), C99 (Choi, 2000), Text-Tiling (Hearst, 1997) et Segmenter (Kan et al., 1998). Les évaluations décrites dans Sitbon et Bellot (2004b) montrent que C99 est le segmenteur le plus efficace. Cependant, ce résultat n'est pas systématique et semble dépendre des thèmes abordés ou bien encore de la taille des documents. Plus récemment, le défi DEFT 2006 a permis une évaluation des systèmes de segmentation à partir d'un corpus de test où tous les titres et autres séparateurs entre documents avaient été ôtés.

2.1.1 Proposition à base de chaînes lexicales pondérées

Le principal reproche que l'on peut faire aux méthodes précédemment citées est qu'elles prennent en compte des paramètres dont les valeurs ont été déterminées de manière empirique. C'est notamment le cas des chaînes lexicales, pour lesquelles le *hiatus* (nombre maximal de phrases ou de mots séparant deux occurrences du mot de la chaîne lexicale considérée) joue un rôle prépondérant dans le calcul des frontières. Nous avons proposé plusieurs approches pour répondre à ce problème.

Estimation automatique du hiatus. Le hiatus peut être identique pour tous les mots et calculé en fonction de la répartition globale des mots dans les textes (distance moyenne entre deux occurrences par exemple). Il peut à l'inverse être défini spécifiquement pour chacun des mots. D'autres critères que la moyenne peuvent être pris en compte en considérant des variations locales de densité des occurrences. Nous avons ainsi proposé de calculer, pour chaque phrase, la fréquence locale de chaque mot définie en fonction de la distance séparant la dite phrase des autres occurrences du mot. Il s'agissait ensuite de procéder à des coupures des chaînes lexicales en fonction de la pente de la courbe représentant les successions de ces fréquences (figure 2.1). Néanmoins, outre le fait que cela ne conduisait qu'à remplacer un paramètre empirique — le hiatus —, par un autre — la pente pour déterminer le hiatus —, les évaluations conduites ont montré que nous ne parvenions pas à égaler les résultats obtenus avec les autres segmenteurs. Cela est essentiellement dû à un déséquilibre entre une analyse locale et globale (empan de la zone où apparaissent les mots) et à l'apparition de chaînes contenant peu d'occurrences au détriment d'autres chaînes plus denses.

Chaînes lexicales pondérées. Il s'agit de ne plus couper les chaînes (toutes les occurrences d'un mot sont alors reliées), mais plutôt de leur attribuer localement un *coefficient d'activité*, un poids local. Le coefficient d'activité des chaînes peut être évalué suivant la densité locale des occurrences du mot étudié en fonction d'une fenêtre de calcul en effectuant un lissage qui s'appuie sur l'ensemble du texte. Ensuite, à chaque frontière potentielle, au lieu de comptabiliser les débuts et fins de chaînes comme le fait Segmenter (Kan et al., 1998), on évalue les variations d'activité de l'ensemble des chaînes, et on

[3] Une adaptation des implémentations effectuées par Choi (2000) et que l'on peut trouver à l'adresse `http://www.cs.man.ac.uk/~mary/choif/software.html` a été réalisée pour l'occasion.

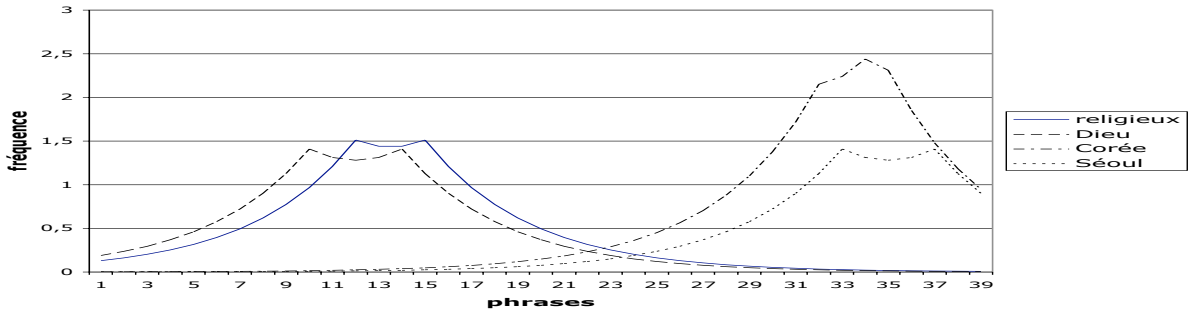


FIG. 2.1: *Exemple de lissage des fréquences locales d'apparition de 4 mots issus du corpus de test du projet Oural. Les premières phrases font apparaître une concomitance des mots "religieux" et "Dieu" tandis qu'après une transition aux alentours des phrases 21–23, ce sont les entités "Corée" et "Séoul" qui apparaissent simultanément. (figure issue de Sitbon et Bellot, 2005a)*

conserve les plus significatives. Galley et al. (2003) avaient déjà proposé une pondération des chaînes en fonction de la compacité et de la fréquence du mot considéré mais continuaient à utiliser un hiatus fixé empiriquement.

Définition 3 (Poids d'une chaîne lexicale) *Pour une position i correspondant à une occurrence d'un mot m, le poids π de la chaîne c correspondant à cette occurrence est défini par :*

$$\pi(c_m, i) = w_{pos_m} \times |m| \times \log \frac{|texte|}{|c|} \tag{2.1}$$

où w_{pos_m} est un poids associé à la catégorie morpho-syntaxique de m dans la chaîne c, $|m|$ désigne le nombre d'occurrences de m dans c, $|texte|$ la longueur du texte à segmenter en nombre total d'occurrences et $|c|$ la longueur de c.

Une fois l'ensemble des poids π calculés, nous avons choisi de déterminer les zones de rupture entre segments en calculant un score de similarité s entre chaque phrase successive A et B (Hearst, 1994) :

$$s(A,B) = \frac{\sum_m \text{score}(A,m) \times \sum_m \text{score}(B,m)}{\sqrt{\sum_m \text{score}(A,m) \times \sum_m \text{score}(B,m)}} \tag{2.2}$$

avec

$$\text{score}(X,m) = \max_{\iota \in X \pm 2 phrases} \pi(c_m, \iota) \tag{2.3}$$

Les frontières entre segments sont ensuite positionnées entre les paires de phrases A et B ayant les valeurs de similarités s les plus faibles, telles que :

$$s(A,B) < \mu + \frac{\sigma}{2} \tag{2.4}$$

avec μ et σ la moyenne arithmétique et la variance de l'ensemble des valeurs s obtenues.

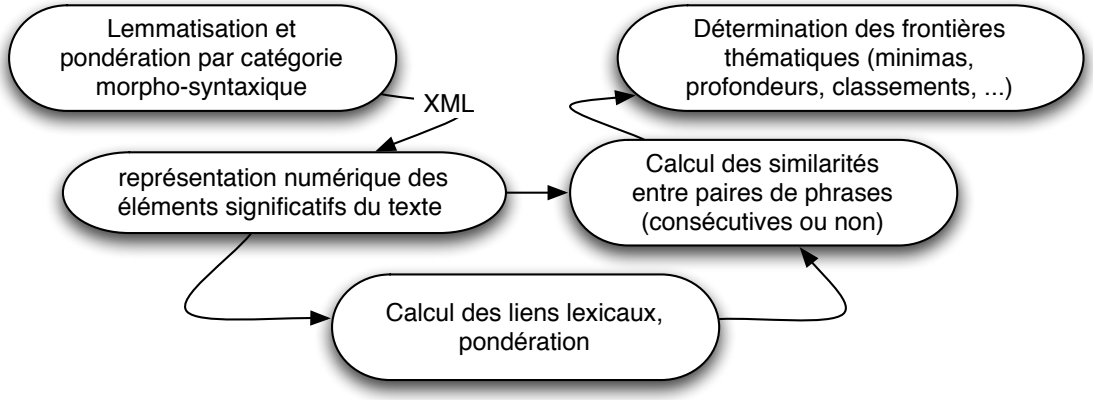


FIG. 2.2: *Schéma de fonctionnement général du segmenteur LIA_SEG développé dans le cadre du projet Technolange AGILE-OURAL.*

L'évaluation de la segmentation peut se faire de multiples façons (Sitbon et Bellot, 2005b, 2006) : les classiques rappel et précision ou leur combinaison, la mesure de Beeferman et al. (1997) *etc.*Nous avons privilégié la mesure d'évaluation *WindowDiff* (Pevzner et Hearst, 2002) qui se calcule en faisant glisser une fenêtre dans le texte segmenté et dans le texte de référence afin de mesurer la distance qui sépare les frontières trouvées des frontières réelles (l'objectif étant d'obtenir la plus petite valeur possible). Cette mesure présente l'avantage de ne pas être sensible aux tailles des segments et de mettre sur un pied d'égalité fausses alarmes et frontières manquées. Par contre, il s'agit d'une mesure qui n'est pas bornée, rendant de fait difficile toute comparaison automatique sur différents corpus.

Définition 4 (WindowDiff) *Soient i et j deux positions dans le texte x, N le nombre de phrases dans x, $b(x_i, x_j)$ le nombre de frontières entre i et j, ref le texte de référence et hyp le texte segmenté automatiquement. La mesure est définie comme suit :*

$$\text{WindowDiff}(rep, hyp) = \frac{1}{N-k} \sum_i \Big(|b(ref_i, ref_{i+k}) - b(hyp_i, hyp_{i+k})| \Big) \tag{2.5}$$

Le corpus de test que nous avons développé durant le projet Oural est constitué de 50 documents aux thématiques variées et qui sont construits à partir d'extraits du quotidien Le Monde. Sur ce corpus[4], la mesure *WindowDiff* montre une amélioration de la segmentation par rapport à un hiatus fixe et des chaînes lexicales non pondérées (tableau 2.1). Les résultats de cette étude ont été publiés lors de la conférence ACM SIGIR en 2007 (Sitbon et Bellot, 2007).

L'intégration de pondérations des mots en fonction de leur distribution en corpus (composante idf par exemple) est à envisager de même que des mesures d'association lexicales pour tenir compte de la cohérence interne des mots apparaissant dans une phrase comme proposé par Ferret (2002).

[4] Les documents ont été lemmatisés par l'étiqueteur morpho-syntaxique probabiliste ECSta (Spriet et El-Bèze, 1998).

	Sans pondération, hiatus=11 phrases (LCSeg)	Chaînes pondérées, longueur non limitée (LIASeg)
WindowDiff	0,3837	**0,3685**

TAB. 2.1: *Evaluation sur le corpus Oural de la segmentation à base de chaînes lexicales pondérées (LIASeg) par rapport aux chaînes non pondérées (LCSeg) — mesure WindowDiff (2.5). Le meilleur résultat est en gras.*

2.1.2 Evaluation durant la campagne DEFT 2005

Cette méthode de segmentation non supervisée a été évaluée dans le cadre de la campagne DEFT 2005 (Labadié et al., 2005) pour la tâche d'extraction de parties de discours politiques "intruses" : détection de phrases de F. Mitterrand au sein de discours de J. Chirac. Cette tâche pouvait être subdivisée en deux : d'une part la détection de ruptures (segmentation) et d'autre part l'identification de l'auteur réel (catégorisation). La figure 2.2, p. 56, schématise le fonctionnement du segmenteur LIA_SEG développé dans le cadre du projet Technolangue OURAL.

La partie "segmentation" a été évaluée selon un *F-score* comparant les frontières obtenues avec celles de la référence en recherchant les frontières les plus proches deux à deux. Si la frontière déterminée automatiquement est avant un changement Chirac vers Mitterrand ou après un passage Mitterrand vers Chirac, on considère que des phrases ont été rajoutées (à l'inverse, on suppose que des phrases ont été oubliées) :

$$\begin{aligned} FScore &= \frac{2 \times M_{corrects}}{M_{total_à_trouver} + M_{trouvés}} \\ &= \frac{2 \times (M_{total_à_trouver} - M_{oublis})}{(M_{total_à_trouver} + M_{ajouts} - M_{oublis}) + M_{total_à_trouver}} \end{aligned} \tag{2.6}$$

avec $M_{corrects}$ le nombre de frontières trouvées à la bonne position, $M_{total_à_trouver}$ le nombre total de frontières à trouver, M_{oublis} le nombre de frontières oubliées et M_{ajouts} le nombre de frontières ajoutées comme définies précédemment.

Pour évaluer uniquement la segmentation, nous nous plaçons dans le cas idéal où le catégoriseur intervenant après la segmentation serait capable d'identifier à coup sûr si le segment correspond à un discours de J. Chirac ou de F. Mitterrand. Nous obtenons alors un *F-score* de 0,91 avec un hiatus fixé empiriquement à 11 phrases et un *F-score* de 0,90 sans hiatus (chaînes lexicales non bornées). Dans les mêmes conditions, le *F-score* de la méthode C99 serait de 0,93. Les écarts entre ces 3 évaluations sont faibles et pour choisir l'une ou l'autre, on doit tenir compte des différences entre les tailles moyennes des segments obtenus : 6,7 phrases pour LIA_SEG et 3,9 pour C99. Notons qu'aussi bien LIA_SEG que C99 ont été conçus pour détecter des ruptures thématiques et non des ruptures stylistiques, ce qui aurait pu être plus adapté au défi DEFT 2005.

Pour plus de détails et de résultats concernant cette méthode de segmentation, se reporter aux papiers que nous avons publiés à l'issue du projet OURAL (Sitbon et Bellot, 2004b,a, 2005a, 2007).

2.2 Segmentation pour l'expansion de requêtes

Dans la suite de ma thèse de Doctorat (Bellot, 2000a), j'ai exploité une méthode de segmentation non supervisée pour l'expansion de requêtes. Cette section commence par un bref rappel des méthodes d'expansion avant de résumer l'approche de segmentation proposée durant ma thèse et de présenter les résultats obtenus depuis.

2.2.1 Expansion à partir de ressources externes.

Rocchio (1971) (voir en Annexes, p. 155) a publié le premier travail majeur sur l'enrichissement automatique de requêtes afin de pallier les phénomènes de synonymie et de polysémie. L'idée centrale était d'aller puiser dans les documents que l'on sait être pertinents les variantes des mots de la requête et dans les documents non pertinents les mots qu'il ne faut au contraire pas trouver. Ce processus « requête, expansion, nouvelle requête » peut être itéré. Dans une telle "boucle de rétroaction de pertinence", l'utilisateur choisit, dans la liste des documents trouvés, quelques documents pertinents (et le cas échéant, quelques documents non pertinents). Dans la pratique seuls les 10 ou 20 premiers documents demandent à être examinés et servent de base à l'expansion de la requête.

En ce qui concerne l'enrichissement à partir de ressources externes, il peut s'agir tout d'abord d'ajouter dans la requête les différentes écritures possibles des noms propres et des acronymes, de corriger certaines erreurs d'orthographe ou tout simplement d'ajouter ou de remplacer les caractères non standards (accents, majuscules).

Ces traitements, parfois effectués sur les corpus eux-mêmes durant l'indexation, ne sont pas sans danger : ajout d'ambiguïtés, modification du sens de la requête, perte de précision de la recherche (éventuellement au profit d'un meilleur rappel). De nombreux auteurs ont proposé d'enrichir les requêtes à l'aide des informations sémantiques de Wordnet (Miller et al., 1990) parmi lesquelles les synonymes. Il n'a malheureusement été constaté d'amélioration des résultats que dans des cas restreints (Baziz et al., 2005). Il est également montré qu'une amélioration par hyponymie améliore plus nettement les résultats que par synonymie. Une autre piste consiste à enrichir la requête avec des informations que l'on s'attend à trouver dans la réponse : par exemple les unités pour des questions appelant des réponses numériques (ajout des unités des valeurs numériques cherchées, tel que *km* par exemple pour une question portant sur une distance, *cf.* Monz, 2003).

Réécriture de requêtes pour questions-réponses. Dans le cadre de l'exploitation des moteurs de recherche sur le Web, Radev et al. (2001) ont cherché à déterminer automatiquement quelle est la meilleure écriture d'une question pour chaque moteur du Web visé. Leur approche s'appuie sur les modèles de traduction qui, dans ce cadre, «traduisent» une question en une requête à l'aide d'opérateurs simples (insertion, suppression, traduction, ajout de guillemets pour les expressions figées, d'opérateurs rendant le mot nécessaire ou bien au contraire indésirable) et d'une phase d'apprentissage automatique à partir de couples questions / réponses ad-hoc.

Pour la tâche questions-réponses, Monz (2003) a proposé plusieurs stratégies pour transformer la question en « requête » pondérée : l'idée est de profiter de variantes des questions et d'analyser les performances de chacune d'elles pour estimer le poids des mots à utiliser. La méthode d'apprentissage est basée sur les arbres de décision de type M5' qui sont présents dans l'environnement WEKA[5]. Si les résultats obtenus sont intéressants, le problème de la généralisation à des questions nouvelles reste posé faute de données d'apprentissage en quantité insuffisante.

Une autre stratégie de recherche consiste à transformer la question en une expression de ce que pourrait être la réponse cherchée (patrons de réécriture). Cette stratégie a été suivie par Brill et al. (2001) lors de la campagne TREC-2001 et dans un certain sens par Echihabi et Marcu (2003) qui établissent un modèle de transformation des phrases trouvées vers les questions et, ainsi, estiment une mesure de ressemblance (ils utilisent à cet effet le modèle canal-bruité déjà utilisé en traduction automatique ou en reconnaissance de la parole).

De notre côté, nous proposons une méthode fonctionnant à partir des résultats d'une segmentation, fonction de la requête, des documents trouvés.

2.2.2 Des arbres de décision non supervisés pour une segmentation thématique orientée requête

Les arbres de décision peuvent être utilisés pour répartir des textes en différentes classes que l'on supposera thématiques à partir du moment où l'on s'accorde sur l'hypothèse que la distribution des mots dans un texte est significative de sa thématique et que deux textes qui partagent des distributions similaires sont thématiquement proches[6]. Si en outre l'on suppose que, pour une requête donnée, les documents pertinents ont tendance à être proches les uns des autres et éloignés des non pertinents (ou encore : deux documents proches ont tendance à être pertinents si l'un d'entre eux est pertinent), la classification par arbres de décisions de l'ensemble des documents trouvés par un moteur de recherche à partir d'une requête conduira à l'apparition de deux types de classes/feuilles (au moins), celles des documents pertinents et celles des documents non pertinents. Dans le cas où la classification est opérée non pas sur les documents trouvés mais *sur les phrases* des documents trouvés, une segmentation des documents pourra être déduite. Une marque de segmentation sera apposée lorsque deux phrases contiguës d'un document se trouvent dans deux classes différentes.

Sur le plan méthodologique, l'idée forte de la méthode de segmentation de textes proposée durant ma thèse de Doctorat, se résume à dire que les arbres de décision tels que les arbres de classification sémantique (Kuhn et De Mori, 1995), peuvent être

[5] http://www.cs.waikato.ac.nz/ml/weka/

[6] Cette hypothèse n'est pas toujours vérifiée à cause des nombreux phénomènes bien connus liés à la variation terminologique. La quantité des mots pris en compte permet toutefois de minimiser le problème puisque le nombre de mots communs à deux documents est un facteur fortement indicatif de cohésion thématique permettant de masquer en partie les ambiguïtés créées par le phénomène d'homonymie. Le problème n'est généralement visible qu'en présence de documents courts.

utilisés de façon non supervisée. Cela est possible en appliquant, sur les données non étiquetées dont nous disposons, la même approche que celle employée classiquement sur des données d'apprentissage. Il faut pour cela utiliser comme critère de mesure de la pureté d'un nœud, non pas un critère de bonne ou mauvaise classification objective (nous n'avons pas d'exemples étiquetés) mais un critère qui, sans être toujours exact, s'est avéré fiable dans ses estimations. C'est le cas par exemple de la probabilité qu'une requête puisse être *générée*, au sens des modèles de langage bayésiens, par les individus — ici, des phrases — du nœud considéré.

Pour construire un arbre de décision, on doit définir un ensemble de questions portant sur les caractéristiques des individus, une règle pour déterminer la meilleure question à poser aux individus d'un nœud et un critère d'arrêt déterminant l'ensemble des feuilles de l'arbre.

Les "questions". Les mots employés dans les textes permettent de déterminer la ressemblance de ces derniers. Le but étant ici de regrouper les individus (les phrases) proches les uns des autres, on fait en sorte de placer dans la même feuille de l'arbre ceux qui ont en commun le plus grand nombre possible de mots. À chaque nœud de l'arbre, il s'agit donc de calculer quel est le mot x qui permet de subdiviser au mieux les individus qu'il contient. Une question Q_x de la forme : « *les individus contiennent-ils le mot x ?* » est posée à chaque individu. Elle subdivise un nœud N en deux nœuds fils N_{OUI} et N_{NON} qui comprennent respectivement les individus de N qui contiennent et qui ne contiennent pas le mot x. Dans le cas de questions simples, x désigne n'importe quelle entrée référencée dans l'index. Des questions « doubles » de la forme « *les individus contiennent-ils les mots x et y ?* » peuvent également être posées. En imposant que l'un des deux mots x ou y soit issu de la requête, l'utilisation de questions doubles permet une certaine désambiguïsation des termes de la requête. Par exemple, si la requête contient le mot `résistance`, des questions du type `(résistance, guerre)` et `(résistance, ohm)` permettent de départager deux interprétations possibles du mot *résistance*. Cette restriction permet en outre de limiter le nombre de questions possibles et de faire en sorte que cette réduction dépende des mots de la requête.

Calcul de l'entropie d'un nœud. La classification opérée par l'arbre de décision conduit à la partition d'un ensemble en deux classes, celle des phrases ayant répondu « oui » à la question et celle des phrases ayant répondu « non ». Le critère de qualité d'une classe dépend de la pertinence estimée des phrases qu'elle contient, sachant que l'objectif est d'obtenir des classes ayant le plus possible de phrases issues de documents pertinents et le moins possible de phrases issues de documents non pertinents. Une des manières d'estimer la pertinence d'une classe est d'évaluer la probabilité que la requête puisse être générée à partir des phrases contenues dans cette classe.

Plus les valeurs de probabilité sont proches les unes des autres, plus l'entropie est grande. Ici, la variable aléatoire choisie est discrète et a deux modalités : les phrases contenues dans la classe considérée sont pertinentes vis à vis de la requête ou ne le sont pas. Trouver la meilleure *question* pour subdiviser un nœud revient à déterminer quel

est le mot permettant de partitionner l'ensemble des phrases de telle sorte que, suivant que ce mot est contenu ou non dans les phrases, pour un des deux nouveaux nœuds créés, la probabilité de pertinence soit aussi élevée que possible, et, pour le second, aussi faible que possible. Cela revient à départager au mieux les phrases d'un nœud en deux sous-ensembles de phrases pertinentes d'un côté et de phrases non pertinentes de l'autre. Dans le cas idéal, la *question* doit permettre de dire « *j'avais un ensemble de phrases de pertinence globale indéterminée, j'ai maintenant deux ensembles de phrases pour lesquelles je peux décider si elles sont pertinentes ou non avec une probabilité nulle de me tromper* ».

L'entropie (2.8) permet de mesurer l'incertitude liée à la décision de pertinence ou de non pertinence des phrases d'un nœud. La meilleure question est celle qui permet la plus grande perte d'incertitude (formule 2.11 où ΔH désigne le gain en entropie, lui-même fonction de l'entropie moyenne des deux classes, formule 2.10). Pour cela, sont calculées pour chaque question q : l'entropie H liée au nœud considéré N et l'entropie moyenne des nœuds N_{OUI} et N_{NON} résultant de la subdivision entraînée par q. La question choisie est celle à laquelle correspond la plus grande baisse d'entropie (ou qui maximise ΔH 2.11).

À chaque nœud de l'arbre correspondent deux valeurs de probabilité. La valeur p correspond à la probabilité que la requête soit générée par les phrases S_i du nœud considéré. La seconde valeur (probabilité que la requête *ne* soit *pas* générée par les phrases) est naturellement le complément de la première, soit $1-p$. La probabilité p est calculée suivant un modèle de langage. En choisissant un modèle unigramme :

$$\begin{aligned} p\left(q = w_1, w_2, \cdots, w_n | \cup_{i/S_i \in N}\right) &= \prod_{j=1}^{n} p\left(w_j | \cup_{i/S_i \in N} S_i\right) \\ &= \frac{\prod_{j=1}^{n}\left(\sum_i Z(w_j, S_i)\right)}{|\cup_{i/S_i \in N S_i|^n}} \end{aligned} \tag{2.7}$$

avec q la requête, w_j le j^e mot de q, N un nœud de l'arbre, S_i une phrase contenue dans N et $Z(w_j, S_i)$ le nombre d'occurrences de w_j dans S_i. Les barres $|\ |$ indiquent le nombre total d'occurrences des mots dans l'ensemble des phrases.

De manière générale, l'entropie H d'une classe C est définie suivant les valeurs de probabilité des modalités e_i d'une variable aléatoire :

$$H_C = \sum_i -P(e_i) \cdot \log\left(P(e_i)\right) \tag{2.8}$$

L'entropie d'un nœud N est définie comme suit :

$$H_N = -p \log p - (1-p)\log(1-p) \tag{2.9}$$

Les entropies correspondant à chacun des deux nœuds fils issus du nœud N sont notées $H_{OUI,N}$ et $H_{NON,N}$ (ils contiennent respectivement les phrases ayant *répondu* "oui" et "non" à la *question* considérée).

L'entropie moyenne (2.10) correspondant à chacun de ces deux nœuds est fonction de la taille des individus migrant dans $H_{OUI,N}$ et $H_{NON,N}$ — la taille est exprimée par le nombre de segments contenant le mot — par rapport à celle de tous les individus du nœud père :

$$\overline{H_N} = \frac{|\cup_{i/S_i \in N_{OUI}} S_i|}{|\cup_{i/S_i \in N} S_i|} \cdot H_{OUI,N} + \frac{|\cup_{i/S_i \in N_{NON}} S_i|}{|\cup_{i/S_i \in N} S_i|} \cdot H_{NON,N} \tag{2.10}$$

Enfin, le gain en entropie associé à une *question* est égal à :

$$\Delta H = \overline{H_N} - \overline{H_N + 1} \tag{2.11}$$

Evaluation pour la recherche documentaire. Le lecteur pourra se reporter à ma thèse de Doctorat ainsi qu'à (Bellot et El-Bèze, 2000a, 2001a) pour des exemples de segmentation et des résultats obtenus sur les corpus et questions des campagnes d'évaluation en recherche documentaire Amaryllis organisées en 1997 et 1999 par l'AUF et l'INIST (Lespinasse et al., 1999). Disons cependant que, indépendamment de la possibilité d'isoler des segments et de permettre un accès plus rapide à l'information recherchée, la méthode a permis, en combinant les scores des documents avec les scores des segments, une amélioration légère mais significative de la précision dans les premiers documents.

2.2.3 Proposition d'une méthode d'expansion à partir d'arbres de décision non supervisés

Cette méthode de segmentation non supervisée peut apporter une réponse au problème de la réécriture et de l'enrichissement automatique de requêtes. Elle permet de surcroît de prendre en compte l'expression d'une certaine forme de négation dans le calcul de la pondération des mots d'une requête. Ce travail a fait l'objet du DEA de Christian Raymond (Raymond et al., 2002).

La classification, effectuée à l'aide d'arbres de décision non supervisés classe les phrases selon qu'elles contiennent ou non certains mots extraits automatiquement de l'ensemble des documents trouvés. Une expression booléenne peut ainsi être associée à chaque feuille (elle correspond à l'ensemble des mots qui par leur présence ou leur absence ont permis de positionner les phrases dans la feuille). Une sélection de la feuille la plus *proche* de la requête, au sens d'une mesure de similarité, permet d'étendre la requête avec l'expression booléenne qui lui correspond, voir figures 2.3 et 2.4. Dans le cas d'une expansion positive, les mots sont simplement rajoutés à la requête originelle, ce qui correspond à augmenter d'autant la composante tf. Dans le cas d'une expansion négative (correspondant au fait que les meilleurs documents trouvés ne contiennent pas le mot donné), cela conduit à diminuer le nombre d'occurrences du mot dans la requête originelle. Si le mot n'était pas présent, il est alors rajouté avec un pseudo nombre d'occurrences négatif. La fonction de score d'un document peut alors prendre des valeurs négatives :

$$tf(m, q') = tf(m, q) + \alpha \cdot tf(m, b) \tag{2.12}$$

avec tf le nombre d'occurrences utilisé pour le calcul des pondérations des mots, m un mot, q la requête d'origine, b l'expression booléenne utilisée pour l'expansion, q' la requête étendue et α un coefficient entier.

Après enrichissement, un score de similarité est calculé exactement de la même manière que pour la requête initiale. En ce sens, nous avons proposé une approche similaire à celle qui l'a été par la suite par Robertson et al. (2004) pour la pondération BM25, préférant agir sur les poids plutôt que sur la fonction de score. Une alternative aurait en effet été de combiner plusieurs scores requête/document, celui obtenu avec la requête initiale, celui calculé avec l'enrichissement positif et enfin celui correspondant à l'enrichissement négatif (voir une discussion à ce sujet, p. 121).

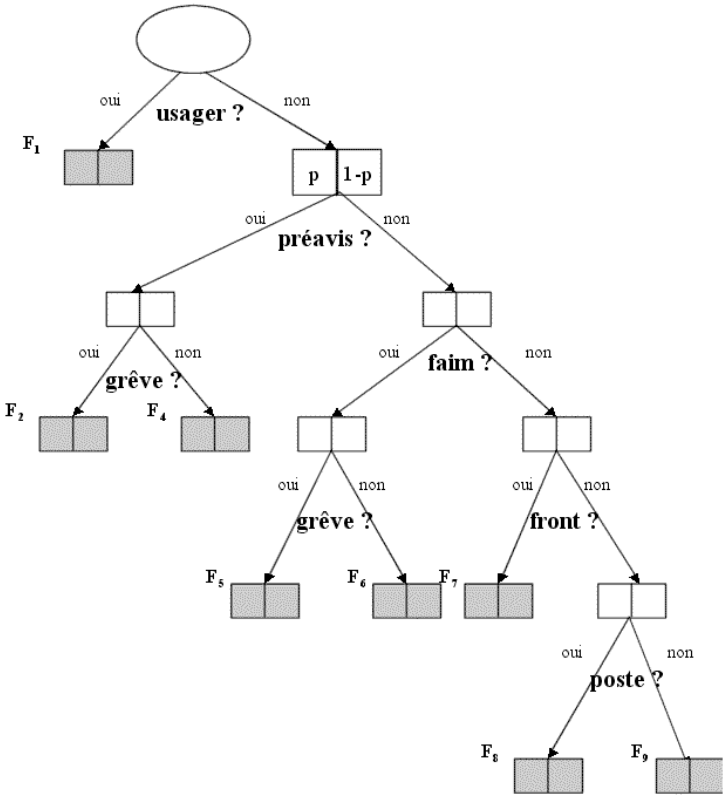


FIG. 2.3: *Exemple d'arbre de décision utilisé pour l'expansion de requête. De tels arbres sont construits à partir d'une segmentation non supervisée des documents trouvés avec les formes initiales des requêtes. Un score de pertinence est associé à chaque feuille selon une fonction de similarité entre les phrases qu'elle contient et la requête. Une expression booléenne développée depuis la racine de l'arbre jusqu'à la feuille ayant le score le plus élevé permet l'expansion de la requête.*

Sur les corpus d'articles journalistiques de la campagne Amaryllis (Lespinasse et al., 1999), la restriction à un enrichissement positif a permis d'améliorer la précision des 5 et 10 premiers documents de manière significative (figure 2.5) aussi bien pour un ré-

dumping social grêve représentant personnel acquis social pouvoir argent exploitation profit ***préavis faim***	attentat conspiration terroriste violence urbain secte de+le Davidiens Emeutes sanglant complot terroriste ***attentat trade***

FIG. 2.4: *Exemples d'expansions obtenues à partir d'arbres de décision non supervisés. Les mots en gras sont les mots rajoutés, ici sur 2 questions de la campagne d'évaluation Amaryllis en recherche documentaire (Lespinasse et al., 1999).*

ordonnacement des documents trouvés initialement que pour une nouvelle recherche. Cependant, nous ne sommes pas (encore) parvenus à établir une méthode de sélection automatique de la meilleure feuille de l'arbre, et ces résultats n'ont pu être obtenus qu'en exploitant la liste des documents à trouver (le référentiel). Ainsi, l'expansion testée correspond au cas idéal où une boucle de rétro-action manuelle conduirait à juger l'ensemble des documents trouvés.

2.3 Détection de plagiats : le projet PIITHIE

NB. Cette section correspond aux travaux réalisés durant le projet ANR-RNTL Piithie depuis février 2007 : Master Recherche de Thierry Waszak et post-docs de Marie-Laure Guénot et Mathieu Estratat sous ma direction.

2.3.1 Présentation

La notion de propriété intellectuelle et de droit d'auteur souffre d'attaques répétées. Si l'on parle très souvent de la diffusion illégale d'œuvres (en particulier musicales et cinématographiques), il est un problème d'une autre nature mais tout aussi important : le plagiat. La réutilisation non consentie d'un texte sans citer la source a toujours existé mais a pris une autre dimension avec l'avènement du Web où la perception de la propriété par les internautes est sensiblement affaiblie par la facilité de copie sans coût des contenus digitaux et la facilité technique à réaliser des pages web à partir d'agrégats d'autres pages. La notion même d'auteur est remise en cause dans la pratique et nous conduit à définir des méthodes de suivi, ne serait-ce que pour tenter de vérifier la véracité et l'origine d'une information. Deux buts applicatifs sont ainsi visés :

- la détection de plagiats de textes ;
- le suivi d'impact.

Outre le LIA, les partenaires de ce projet commencé en 2007 et se terminant en 2009 sont :

- la société Advestigo qui a mis au point un outil de détection de plagiat pouvant analyser à la fois des flux de données et chercher directement sur Internet, dans des sites non ciblés au départ ;
- le Laboratoire d'Informatique de Nantes Atlantique (LINA) ;
- les société Syllabs et Sinequa, porteuses du projet.

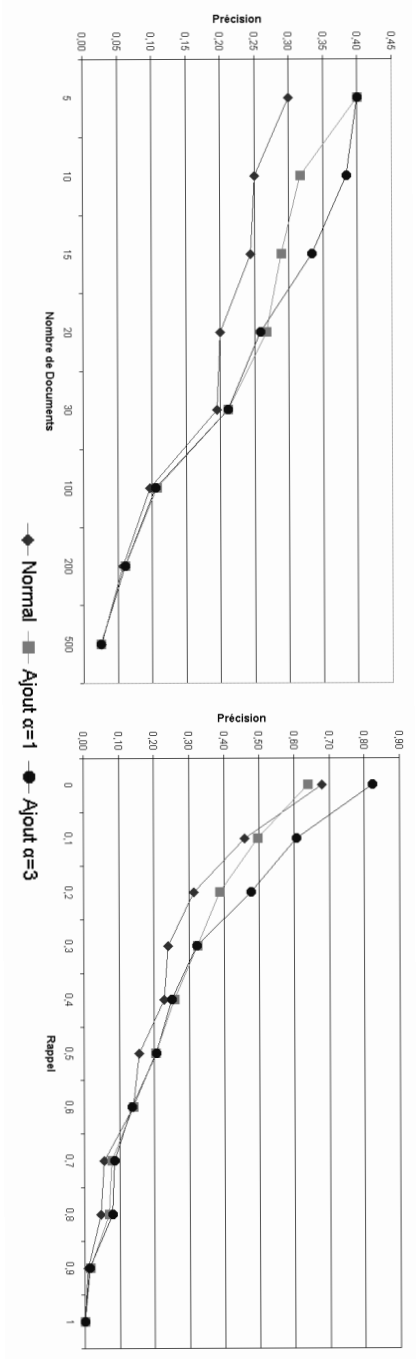


FIG. 2.5: *Courbes de précision et de rappel/précision obtenues avant et après enrichissement sur les 7 requêtes de la campagnes Amaryllis pour lesquelles la meilleure feuille de l'arbre de décision ne correspondait pas qu'à une expansion négative. Tests avec* $\alpha = 1$ *ou* $\alpha = 4$.

Le principe général du système de détection de plagiats est le suivant : un texte original est extrait d'une base ou d'un flux valide. Des opérations de segmentation du

texte (LIA), de calcul d'empreinte lexicale et sémantico-discursives (LIA/LINA) sont effectuées afin de calculer l'empreinte du document original. À partir de cette empreinte sont extraites des requêtes qui sont soumises à des moteurs de recherche sur Internet (LINA). Des documents, considérés comme des plagiats potentiels sont alors rapportés et un processus similaire à celui effectué sur le document original est effectué : segmentation et calcul d'empreintes. Pour le calcul d'un score de plagiat, les citations (reprises avec mention de l'origine) sont détectées et l'empreinte du plagiat potentiel et celle du document original sont comparées. Si le score est supérieur à un seuil, des techniques d'alignement et de visualisation spécifiques sont employées pour soumettre le plagiat potentiel à un spécialiste qui valide ou invalide l'étiquette de plagiat sur ce document (LIA).

2.3.2 Etat de l'art

Des méthodes de mesures de similarités entre codes sources informatiques sont utilisées depuis de nombreuses années pour la détection de plagiat (Samuelson, 1994; Wise, 1996). Cependant, les applications sur des textes en langue naturelle sont une activité assez récente et d'une grande complexité. Citons toutefois le projet METER[7] (*MEasuring TExt Reuse*) qui s'est proposé de recenser l'ensemble des algorithmes permettant d'effectuer un calcul de similarité entre documents pour la détection automatique de plagiat (Clough et al., 2002). Les différentes méthodes de mesure de similarité reposent toutes sur un calcul (plus ou moins complexe) de fréquences d'apparition de mots ou d'ensembles contigus de mots. Ainsi l'utilisation des n-grammes (Schleimer et al., 2003), de détection de sous chaînes communes ou de phrases communes sont largement utilisées.

Les autres techniques couramment utilisées sont les suivantes :

- alignement de phrase. Les techniques d'alignement de phrases ont été développées la décennie passée et de nombreux algorithmes ont été suggérés qui permettent la détection d'un ensemble de phrases communes et donc de déterminer une similarité entre deux textes ;
- recherche de sous chaînes communes. Il s'agit d'identifier les plus longues sous-chaînes communes entre les textes suspects et le document original. D'autres méthodes, plus sophistiquées, utilisant des arbres pour la recherche de segments identiques (ou proches) sont également mises en œuvre avec succès (Pereira et Ziviani, 2003). Nous formulerons des propositions en ce sens en section 2.3.4, p. 72 ;

De manière générale, détection de plagiat et suivi d'impact sont très liés car il s'agit avant tout d'être capable de détecter qu'une même information est évoquée dans différents textes. En ce sens, les travaux menés dans le cadre du programme TDT nous intéressent particulièrement.

La piste *Topic Detection and Tracking* (TDT). Elle faisait partie du programme DARPA *Translingual Information Detection, Extraction and Summarization* (TIDES 1) destiné à per-

[7] (ttp://www.dcs.shef.ac.uk/nlp/meter/

mettre à des anglophones natifs de retrouver de l'information dans des textes écrits dans d'autres langues que l'anglais (Allan, 2002). La piste s'organisait autour de cinq tâches : la segmentation thématique d'un article de presse, la détection d'un événement nouveau, la classification thématique des dépêches au fur et à mesure de leur réception (de manière hiérarchique pour TDT-2004) et le suivi d'informations (trouver des articles liés à une information donnée mais aussi décider si oui ou non deux informations traitent du même sujet). L'ensemble de ces tâches était effectué sur des textes en différentes langues : anglais, arabe, chinois-mandarin et les corpus disponibles par l'intermédiaire du Linguistic Data Consortium (LDC). Avec la fin du programme TIDES-1 en 2005, la campagne TDT-2004 est la dernière en date. Depuis, les évaluations MSE 2005 (*Multilingual Summarization Evaluation*), puis DUC et TAC ont en quelque sorte pris le relais.

Pour l'ensemble de ces tâches, il est habituel de représenter les textes par une empreinte lexicale (un vecteur ou une distribution de probabilités) qui est ensuite comparée à d'autres. Si la différence entre une nouvelle empreinte et chacune des précédentes est supérieure à une valeur seuil alors cette nouvelle empreinte correspond à une *nouveauté* (Allan et al., 2000). Les calculs de similarités entre ces empreintes sont utilisés pour la détection de liens entre documents ou pour le suivi en utilisant des approches numériques telles que les *k plus proches voisins* (Yang et al., 2000). De nombreux modèles probabilistes ont été développés à partir de ceux employés en recherche documentaire ou en filtrage de documents[8] (Zhang et Callan, 2004) : modèles de Markov cachés, modèles de langage (Elsayed et al., 2004), modèles à maximum d'entropie, arbres de décision... Une tendance actuelle consiste en l'utilisation conjointe de différentes approches et à l'apprentissage automatique des coefficients de confiance accordés à chaque module en jeu.

L'aspect multilingue a été introduit lors de TDT 1999. À cette occasion, Leek et al. (2002) ont montré qu'une simple traduction mot à mot en employant des logiciels en ligne grand public, permettait d'obtenir des résultats proches de ceux constatés sur la langue d'origine : les mots les plus difficiles à traduire du fait de leur polysémie ne sont pas les mots les plus utiles pour les tâches de TDT. Ceci confirme à quel point, la détection d'entités nommées est une étape essentielle (Kumaran et Allan, 2004). La traduction d'entités nommées n'est pas une tâche simple du fait qu'il s'agit de termes généralement absents des dictionnaires. Néanmoins de nombreuses entités nommées ne réclament pas de réelle traduction du moins dans les langues employant des alphabets presque identiques. Al-Onaizan et Knight (2001) proposent une approche de traduction arabe-anglais à base de translittération produisant plusieurs traductions candidates ensuite sélectionnées en utilisant le Web comme suggéré par Grefenstette (1999). En ce qui concerne le vocabulaire général, Schultz et Liberman (2002) conseillent de n'utiliser que des dictionnaires bilingues de tailles très réduites (environ 7000 termes) et assez facilement constructibles manuellement. Malgré tout, des adaptations propres aux différentes langues ont montré leur intérêt notamment en ce qui concerne les normalisations des pondérations des termes (Allan et al., 2002).

[8] *cf.* les actes de TDT-2004, http://www.nist.gov/speech/tests/tdt/tdt2004/workshop.htm

Les expérimentations conduites durant TDT s'insèrent aussi dans la recherche de liens (*link mining*) qui est un domaine à l'intersection de la fouille de textes (*web mining*), de l'analyse de graphes (*graph mining*), de la classification automatique et de l'analyse de liens explicites (*link analysis*) (Getoor et Diehl, 2005; Senator, 2005). Les principales caractéristiques du Web résident à la fois dans l'organisation hypertextuelle des pages qui le composent et par son aspect dynamique. Le premier point est largement exploité depuis des années par les moteurs de recherche du Web qui exploitent la structure en graphe du réseau pour en identifier les nœuds les plus représentatifs (*authorities*). Ainsi, ce n'est pas seulement le strict contenu du texte qui permet de rechercher de l'information mais aussi sa localisation, ses liens, par rapport à l'ensemble des autres textes disponibles. L'intégration de ces deux types de données, le contenu d'une part et les liens explicites et implicites d'autre part (et plus largement le traitement de données semi-structurées) a entraîné une remise en question de nombreuses approches en fouille de données (Senator, 2005). Cela a par exemple été réalisé pour des tâches en catégorisation automatique (Oh et al., 2000) et en détection de groupes (Gibson et al., 1998).

Quelle que soit l'approche choisie pour estimer un score de plagiat entre deux documents, les citations sont un type particulier de reprise qui doit être pris en compte. Elles sont à la fois un indice de plagiat potentiel (les citations sont rarement modifiées dans un texte plagié) et de non plagiat (les citations sont le plus souvent légales). Leur détection est un problème à part entière comme nous allons le voir dans la section suivante.

2.3.3 Détection de citations pour l'identification de plagiats

La détection automatique de citations est une problématique relativement peu explorée dans la littérature alors qu'elle peut avoir un fort impact au sein d'applications qui s'intéressent aux opinions exprimées dans des documents, journaux ou blogs : suivi d'impact (Dave et al., 2003), résumé automatique (Stoyanov et Cardie, 2006), questions-réponses (Somasundaran et al., 2007)... Par opposition aux travaux reposant sur des ressources lexicales importantes pour mesurer la subjectivité et sur des analyses syntaxiques profondes pour déterminer les différentes propositions au sein d'un discours, nous avons souhaité nous restreindre au repérage des débuts et fins de citations ainsi qu'à l'identification de son auteur. En suivant l'approche décrite par Lucas et Nadine (2004) exploitant essentiellement des marques de surfaces de nature typographique, morphologique et positionnelle, nous avons réalisé un système automatique permettant l'identification de constituants "source, relateur, discours rapporté"[9] et ceci selon des méthodes symboliques ou numériques par apprentissage supervisé.

Sans entrer dans les nombreux méandres d'une analyse du discours rapporté (voir par exemple Jackiewicz, 2006), les citations peuvent prendre différentes formes : citation directe (exemples 1 et 2 ci-dessous) où l'on voit une reprise littérale encadrée par

[9] Le relateur marque le lien entre la source du discours —l'auteur— et le discours lui-même. Il peut être un verbe tel que "*souligner*" dans la phrase "*M. a souligné ...*".

des guillemets, citation indirecte (exemple 3) où le discours est intégré dans le texte et l'opinion reprise ou encore intermédiaire (exemple 4) où nous avons un mixte des deux cas précédents.

> (1) **Le quotidien économique** *souligne :* "Si le rapport ne veut pas associer ces montants à l'idée d'une nouvelle 'cagnotte' budgétaire, ni au débat électoral sur le niveau de prélèvements obligatoires, le montant est équivalent au déficit budgétaire de l'Etat, à savoir 36,5 milliards d'euros l'an dernier."
> (2) "En 2003 , *explique-t-***il**, j'ai fait effectuer douze tests sur des vols en France, dans onze cas sur douze, des armes et explosifs ont pu être introduits dans les avions".
> (3) *D'après* **sa mère**, Julien faisait une sieste dans l'appartement familial lorsqu'il a disparu.
> (4) **Le Figaro** *estime*, **lui**, *que* "les techniciens et les cadres sont en première ligne", notamment ceux de la "Central Entity" de Toulouse.

Deux points de vue sont considérés :

- un élément cité ne doit pas être pris en compte pour le calcul des similarités dans la détection de plagiats. Autrement dit : reprendre une citation n'est pas un plagiat ;
- à l'inverse, les citations directes peuvent être un indice supplémentaire dans la détection de plagiats. En effet, une citation directe, surtout dans un article journalistique, a de forte chance d'être reprise sans modification d'un document à l'autre si celui-ci est plagié (autrement dit : un auteur qui en plagie un autre modifiera légèrement le texte mais reprendra probablement les citations telles quelles). *A contrario*, pour un texte que l'on soupçonne d'avoir été plagié, seuls les textes qui contiendront les mêmes citations directes que lui seront analysés en profondeur.

En coopération avec le LINA de l'Université de Nantes, nous avons constitué un corpus d'entraînement et de test constitué de 108 articles journalistiques en français (environ 70 000 mots) issus de différentes sources : La Tribune, Challenges, Le Soir, Le Figaro, Libération, L'Humanité, Le Monde, AFP et Reuters. Une annotation manuelle a conduit à repérer 846 sources (auteurs) et 938 discours repris (citations) directe, indirecte ou intermédiaire.

Les implémentations et évaluations brièvement décrites ci-dessous sont au cœur du Master Recherche que Thierry Waszak a réalisé sous ma direction (Waszak, 2007) et du Master et de la thèse de Doctorat en cours de Fabien Poulard (LINA) dirigés par Nicolas Hernandez et Béatrice Daille. Deux approches ont été suivies : la première à base d'automates à états finis et la seconde selon des méthodes d'apprentissage supervisé sur des critères numériques. La figure 2.6 réprésente l'automate de niveau 2 permettant d'identifier et de mettre en relation source et relateur. Différents niveaux de FSM sont en effet nécessaires : pour identifier les constituants du schéma SRD au 1^{er} niveau, pour les mettre en relation au niveau 2 et pour ordonner le tout au sein des documents et décider entre citations directes et indirectes (niveau 3).

En parallèle avec l'approche à base d'automates, nous avons testé des méthodes numériques (arbres de décision, SVM, réseaux bayésiens) appliquées aux différents éléments caractéristiques d'un texte permettant le repérage des citations, eux-mêmes inspirés des travaux de Mourad et Desclès (2004) :

- marques lexicales ;

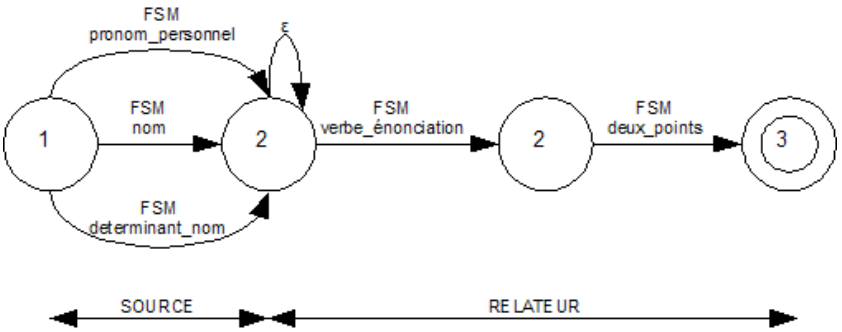


FIG. 2.6: *Exemple de FSM de niveau 2 pour l'identification de citations (couple source / relateur).*

- marques syntaxiques (conjonction "*que*", verbes d'énonciation, étiquettes morpho-syntaxiques en début de discours candidats) ;
- marques typographiques ;
- distance en nombre de mots entre le discours repris candidat, le plus proche verbe d'énonciation caractérisant le discours indirect et le plus proche syntagme prépositionnel introducteur de citation ;
- le temps des verbes (changement ou non) dans le discours par rapport à celui du reste de la phrase ;
- la taille du discours candidat.

Evaluation. Le tableau 2.2 présente les résultats[10] obtenus par validation croisée pour les méthodes numériques et sur l'intégralité du corpus pour les FSM (ceux-ci ont été conçus à partir de l'analyse d'un autre corpus). Les performances sont élevées pour le discours direct avec un léger avantage pour les SVMs en ce qui concerne la précision (0,92) et pour les FSM et les réseaux bayésiens pour la combinaison entre rappel et précision (F-mesure de 0,91 contre 0,89 avec les SVMs). Pour le discours indirect, les différences entre les méthodes sont du même ordre : meilleure précision pour les SVMs et meilleur rappel pour les FSM et les réseaux bayésiens, ce qui, à l'aide d'une précision qui reste correcte, donne au final une meilleure F-mesure. Dans le même temps, nous estimons plus finement la qualité du repérage d'une citation indirecte en évaluant la segmentation avec la mesure WindowDiff (formule 2.5, p.56). Cette dernière vaut 0,23 pour le repérage des citations indirectes avec les FSM, montrant que s'ils sont effectivement identifiés, les discours indirects ne sont pas optimalement délimités.

Les résultats donnés ici ne concernent que le repérage du discours. Pour une évaluation de l'identification automatique des autres constituants des citations (les sources notamment), ainsi que pour une présentation plus détaillée de l'ensemble, se reporter à (Poulard et al., 2008). La figure 2.7 donne un exemple de citations automatiquement détectées avec les *indices* qui les accompagnent.

[10] Dans le cas du discours direct, un segment de texte est considéré correct si ses frontières coïncident exactement avec celles de la référence. Dans le cas du discours indirect, il suffit qu'il y ait chevauchement avec la référence pour que le segment soit considéré correct.

Méthode	Discours direct			Discours indirect		
	Précision	Rappel	F-Mesure	Précision	Rappel	F-Mesure
FSM	0,89	**0,93**	**0,91**	0,58	**0,72**	**0,68**
Réseaux Bayésiens	0,86	**0,93**	**0,91**	0,58	0,68	0,64
SVM	**0,92**	0,88	0,89	**0,70**	0,52	0,57
Arbres de décision	0,91	0,88	0,89	0,66	0,52	0,56

TAB. 2.2: *Évaluation des méthodes de repérage de citations directes et indirectes.*

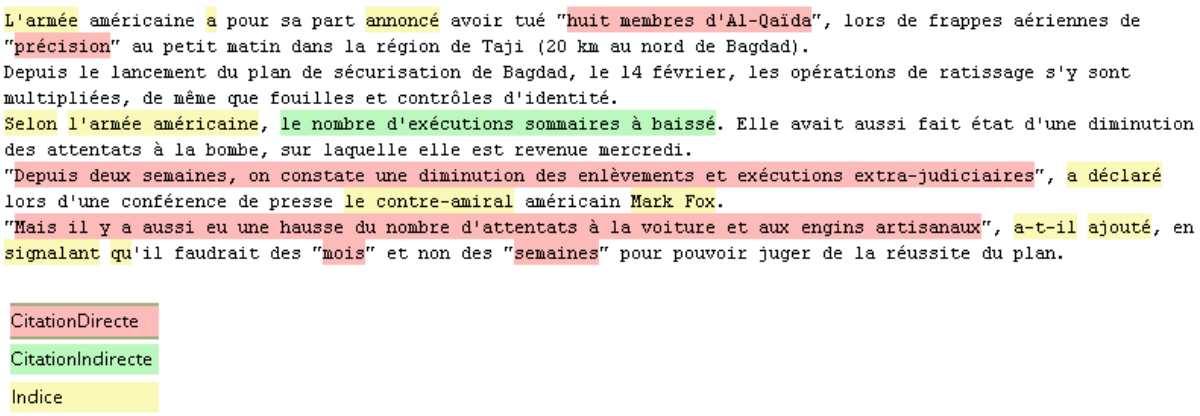


FIG. 2.7: *Exemple de citations automatiquement détectées. Figure reproduite de Waszak, 2007 dans lequel une discussion sur les erreurs produites peut être également trouvée.*

Adaptation vers la langue anglaise. Un sous-ensemble du corpus MPQA en anglais[11] contenant 280 citations a été annoté afin d'estimer la portabilité des méthodes et paramètres mis au point sur le français. Pour les FSM, les mots et expressions qui leur correspondent ont été directement traduites en anglais sans reprendre leur structure. Pour les approches numériques, aucune adaptation n'a été effectuée, les attributs paraissant pouvoir être conservés tels quels.

En ce qui concerne l'identification des citations directes, les résultats sont comparables avec ceux obtenus sur le français (F-mesure de 0,90 au lieu de 0,91 pour les SVMs et de 0,87 au lieu de 0,91 pour les FSM). Pour les citations indirectes par contre, les résultats des FSM chutent fortement (F-mesure de 0,26 au lieu de 0,68) tandis que les méthodes numériques se maintiennent mieux (F-mesure de 0,46 au lieu de 0,57 pour les SVM) voire s'améliorent (F-mesure de 0,63 au lieu de 0,56 pour les arbres de décision). Ces résultats confirment nos attentes : une meilleure portabilité des approches numériques et des performances similaires pour les citations directes qui s'expriment en anglais de façon très proche à celle du français.

[11] http://nrrc.mitre.org/NRRC/02_results/mpqa.html

2.3.4 Segmentation de textes pour la détection de plagiats

La segmentation peut avoir un double rôle dans la détection de plagiats : réduire le champ de recherche, et donc la complexité, en réduisant la taille des textes considérés mais aussi participer à la détection proprement dite en repérant des ruptures « symptomatiques ». Nous proposons de diversifier la segmentation selon plusieurs axes : ruptures thématiques, ruptures stylistiques et ruptures structurelles (au sens de la structure logique du discours). Aussi bien les textes sources (documents de référence), que les textes cibles (documents faisant l'objet d'une analyse pour plagiat potentiel) seront segmentés et leurs segments comparés deux à deux pour aboutir à un score local de ressemblance.

Au moins deux sous-problèmes se présentent : la détection de points d'ancrages pour la segmentation (utilisation des citations directes (section 2.3.3, p. 68) et de zones de plagiat *verbatim* sans indication de la source) et le calcul des similarités entre deux passages de texte (une bonne segmentation peut être celle qui maximise cette similarité mais aussi la longueur des passages : retrouver la plus grande zone de plagiat potentiel).

Au delà de ces aspects, nous proposons une méthode de structuration des documents textuels qui construira de manière automatique une représentation de l'organisation du contenu informatif des textes à partir d'indices linguistiques de forme. Elle s'inspire largement de plusieurs modèles, tels que *Dynamic Syntax* (Kempson et al., 2000) et différentes théories de l'analyse de discours (voir par exemple Régnier, 2007). Elle se présente comme une représentation à mi-chemin entre syntaxe dépendancielle et organisation discursive. Mais ce qui l'en éloigne le plus est qu'elle fait appel à une ressource unique, laquelle est conçue de manière à être la plus succincte et surfacique possible : un lexique procédural qui recense les propriétés structurelles de certains indices de rupture sans nécessiter d'analyse linguistique approfondie.

La détection de copies *verbatim*

La reprise la plus simple, à savoir le plagiat *verbatim* est la première à devoir être considérée. Il s'agit d'un cas trivial mais, à défaut de plagier l'intégralité d'un texte, il est fréquent que seules certaines parties le soient. Le problème est dès lors de repérer ces zones identiques : le nombre et la taille des suites de mots contigus identiques dans un texte et dans un autre sont un indice de plagiat.

Les figures 2.8 et 2.9 présentent respectivement des extraits de deux textes, le premier stipule une source tandis que le second n'en précise pas. Il y a de fortes chances pour que le second soit un plagiat du premier. Sur ces textes, sont coloriées deux citations identiques, deux phrases très proches d'un texte à l'autre qui sont autant d'indices pour calculer une similarité ainsi que deux blocs proches moyennant quelques transformations.

[...] Le patron du groupe européen de défense et d'aéronautique EADS, Louis Gallois, a déclaré mardi qu'il allait proposer la suppression totale du système des stocks-options dans le groupe, dans un entretien au quotidien français Le Monde.
"J'avais déjà dit (...) que je considérais l'attribution de stock-options aux dirigeants comme un système contestable qui s'apparente à une loterie. Je vais proposer au conseil d'administration sa suppression totale", déclare le président exécutif du groupe.
La majorité des hauts dirigeants d'EADS ont dégagé d'importantes plus-values sur des ventes d'actions issues de stock-options à la fin 2005 et en mars 2006, quelques mois avant l'effondrement du titre. Cet effondrement était lié à la révélation d'importants retards du programme A380 le 13 juin 2006, des plus-values qui font l'objet de l'enquÍte actuelle sur de possibles délits d'initiés.
Dans cet entretien, M. Gallois ajoute qu'il **"faut leur substituer un mode de rémunération plus transparent comme l'attribution d'actions gratuites qui sont un complément de salaire"**.
Par ailleurs, [...]

FIG. 2.8: *Exemple de texte source*

Recherche semi-automatique de copies *verbatim*. Une application graphique a été développée par M. Estratat permettant de viualiser des textes sous la forme de rubans colorés. L'objectif est d'accélérer le processus d'identification de zones communes à deux textes (copies *verbatim*). La coloration du texte s'effectue de la manière suivante :

1. le texte est lu une première fois et, à chaque mot nouveau est associée une valeur de couleur spécifique ;
2. si un mot a déjà été rencontré il reprend la couleur affectée précédemment ;
3. l'application affiche le ruban coloré correspondant.

Cette application peut être utilisée avec deux textes en utilisant les mêmes lexiques de correspondance entre couleurs. Ainsi, les textes peuvent être comparés "à l'œil nu". Mais il est également envisageable de faire glisser les fenêtres l'une sur l'autre et de ne retenir que les les différences proches de 0 (à un seuil près). La figure 2.10 présente le résultat de la coloration du texte 2.8 et la figure 2.11 celui du texte 2.9.

Recherche automatique de copies *verbatim*. Les algorithmes de recherche de sous-chaînes communes à deux textes que nous avons pu trouver dans la littérature sont itératifs et partent du premier mot de la source pour le chercher dans le texte cible et ainsi de suite. Si nous voulions utiliser ces techniques, il faudrait effectuer une recherche pour tous les tuples de mots. Ce qui se solderait par une complexité (en posant n la taille de la source S, et m celle de la cible C, avec $n \geq m$) de l'ordre de :

$$\sum_{i=0}^{(m-1)} (n-i)(m-i) \leq \sum_{i=1}^{(n)} i^2 = \frac{n(n+1)(2n+1)}{6} \in O(n^3)$$

– soient deux textes, l'un source (noté S) et l'autre cible (noté C), et un lexique contenant l'intersection des ensembles de mots de S et de C. La figure 2.12 illustre la construction des structures de départ ;

[...] Louis Gallois, patron du groupe européen de défense et d'aéronautique EADS, a déclaré qu'il allait proposer la suppression totale du système des stock-options dans le groupe, lors d'un entretien accordé au (à Le) Monde, mardi 9 octobre.
"J'avais déjà dit (...) que je considérais l'attribution de stock-options aux dirigeants comme un système contestable qui s'apparente à une loterie. Je vais proposer au conseil d'administration sa suppression totale", déclare le président exécutif du groupe.
Attribution d'actions gratuites
Quelques temps avant la chute de l'action EADS, lors de la révélation des importants retards du programme A380 en juin 2006, la plupart des hauts dirigeants d'EADS ont dégagé d'importantes plus-values sur des ventes d'actions issues de stock-options, fin 2005 et en mars 2006. [...] Dans cet entretien, Louis Gallois ajoute qu'il "**faut leur substituer un mode de rémunération plus transparent comme l'attribution d'actions gratuites qui sont un complément de salaire**" [...]

FIG. 2.9: *Exemple de texte cible proche*

FIG. 2.10: *Coloration du texte source*

– à partir des positions des mots dans S et dans C, les chaînes de mots communes à S et C sont recherchées. Coût approximatif dans le pire des cas (textes identiques avec un seul mot répété m fois) : $O(nm)$.

Nous proposons une autre approche illustrée par la figure 2.13 qui représente les liens créés lors de l'analyse.

L'application graphique de cette méthode permet de colorer les segments communs à deux textes. La figure 2.15 présente le résultat de la comparaison entre le texte source et le texte cible.

Une segmentation non linéaire

Partant du principe que de nombreux plagiats essayent d'être maquillés le plus rapidement possible, nous nous intéressons à l'exploitation linguistique de l'information portée par les objets appelés "mots grammaticaux", "mots outils" ou "mots vides" qui sont porteurs de la structure de la phrase. Autrement dit : le maquillage peut consister à remplacer des "mots non outils" par d'autres mots au sens proche mais ne touche pas ou peu à la structure de base des phrases, aux inversions et à l'insertion ou la suppression

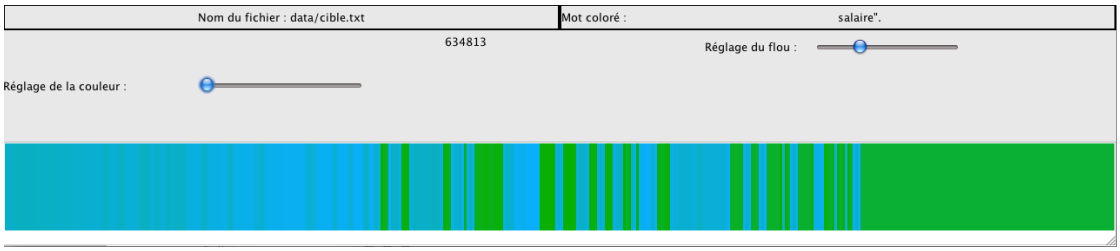


FIG. 2.11: *Coloration du texte cible*

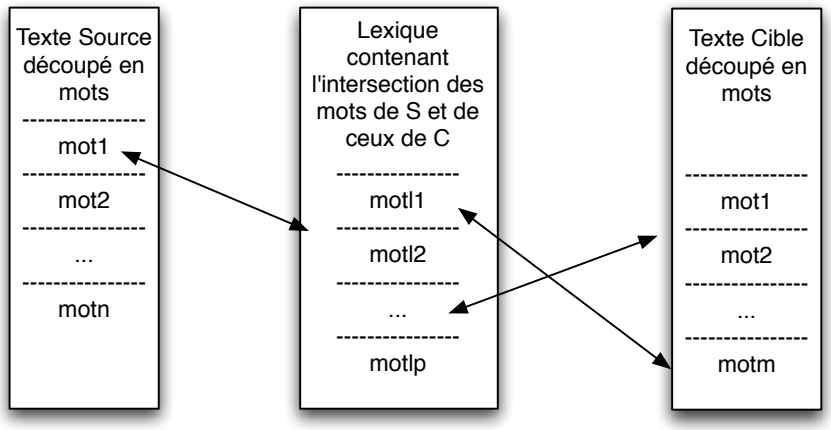


FIG. 2.12: *Construction des structures de base de VSearch*

près de subordonnées ou mises entre parenthèses. Notre proposition se rapproche en cela du Modèle Linguistique de Discours de Polanyi (1987) qui se base sur une opposition de surface entre des constituants portant l'information propositionnelle et d'autres, comme les connecteurs, qui sont des opérateurs de discours. Cela se rapproche des méthodes qui calculent des distances entre arbres de dépendances syntaxiques pour appareiller des réponses candidates à une question avec des réponses type (Kouylekov et al., 2006). L'objectif est de construire dynamiquement une représentation arborescente de la structure logique et du contenu informationnel des documents puis de comparer les structures des arbres source et cible.

La recherche de sous-arbres isomorphes conduit à une segmentation des textes (voir figure 2.14) par le simple fait d'en isoler certaines parties communes. Ce problème peut être résolu en temps polynomial (Garay et Johnson, 1979). Nous nous intéressons donc à limiter au plus possible la création d'arcs rendant le graphe non planaire pour limiter les coûts de calcul : le problème est de complexité non polynomiale sinon (Kobler et al., 1993; Jenner et al., 2003).

Ces structures pourront ensuite être comparées afin d'en dégager une mesure de proximité (similarité). Au-delà des questionnements que nous serons amenés à nous poser concernant la définition de la grammaire permettant d'aboutir à une arbores-

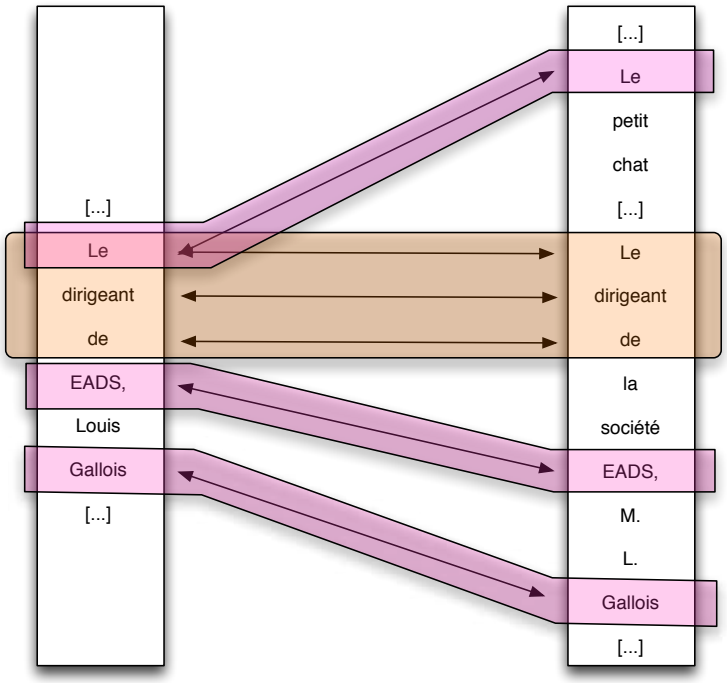


FIG. 2.13: *Mise en correspondance des occurrences des mots pendant la recherche de sous-chaînes communes à deux textes pour la recherche automatique de copies verbatim.*

cence utile, un enjeu scientifique réside dans la capacité à rechercher des sous-arbres isomorphes *à "quelque chose" près*. Par exemple, pour un nœud racine (du sous-arbre) donné, il pourra y avoir une différence de nombre de noeuds- fils, de même que lesdits noeuds descendants pourront se situer à un intervalle différent d'un sous-arbre à l'autre. Puisque l'ordre des arcs n'est pas significatif (coordination), c'est l'ensemble des nœuds-fils d'un nœud donné qui sera étudié sans tenir compte de leur ordre d'apparition linéaire dans le texte. Cela permet de gérer la possibilité d'inversion de constituants au sein du texte cible par rapport à la source.

La question de la définition d'une mesure de similarité rendant compte d'isomorphismes partiels semble être toujours un problème ouvert, et ceci indépendamment de l'*objet* associé aux nœuds (en l'occurrence les nœuds représentent des mots ou des syntagmes, ils sont typés selon leur fonction de coordination ou de subordination, et, le cas échéant, avec leur étiquette morpho-syntaxique : cela rend la recherche d'isomorphismes partiels d'autant plus complexe). Une approche structurelle de la recherche d'isomorphismes partiels conduit à envisager différentes réductions possibles puis à déterminer d'éventuelles bijections entre les squelettes réduits. Par opposition, une recherche *par le contenu* consiste à identifier des isomorphismes (stricts) puis à quantifier la structure et la nature des arcs qui relient deux sous-arbres de arbre du texte cible, sachant que ces deux sous-arbres sont isomorphes à deux autres sous-arbres du texte source. Cela s'apparente à l'approche mise en œuvre par Baziz et al. (2007) pour l'in-

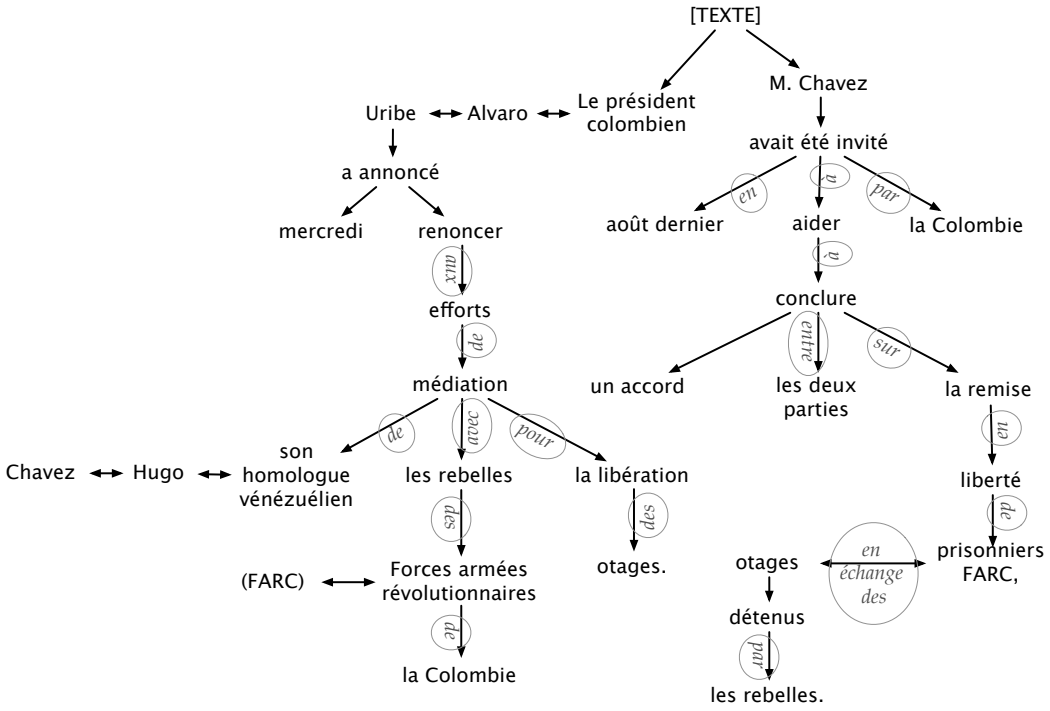


FIG. 2.14: *Exemple de structure propositionnelle pour la détection de phrases similaires. Le squelette de cette structure est un indice qui servira au calcul de similarité pour la détection de plagiats.*

dexation conceptuelle. Il a été montré que le calcul d'une distance d'édition entre des arbres étiquetés et non ordonnés (l'ordre des fils n'importe pas) est un problème NP-difficile (Zhang et al., 1992) qui a malgré tout fait l'objet d'implémentations (Zhang, 1996; Wang et al., 1994). Le cas où les arbres sont ordonnés (l'ordre entre les fils d'un nœud doit être conservé) a été étudié pour la comparaison de structures secondaires d'ARN (Dulucq et Tichit, 2001) où, comme en ce qui nous concerne, la séquence et la structure sont deux indications importantes de la similarité entre les objets étudiés.

La suite de ce travail donnera lieu à deux séries de travaux : l'une sur la définition de grammaires appropriées à la tâche de détection de plagiat et, éventuellement, en la réduisant à certains types de plagiats, et l'autre à l'étude de mesures de similarité continues entre sous-arbres.

FIG. 2.15: *Coloration des segments communs de taille ≥ 5 pour la recherche de copies verbatim.*

Deuxième partie

Recherche personnalisée : navigation et lisibilité

Chapitre 3

Recherche d'informations et résumé automatique audio

Sommaire

3.1 Recherche d'informations au sein de documents audio 81
3.2 Navigation dans des bases audio par résumé automatique 83
3.3 Une approche de résumé automatique audio orienté requête 84
3.4 Expériences durant DUC 2006 87

NB. Ce chapitre correspond aux travaux réalisés durant la thèse de Doctorat de Benoît Favre, en convention CIFRE avec la société Thalès et co-encadrée par Jean-François Bonastre (LIA) et moi-même entre septembre 2003 et début 2007.

3.1 Recherche d'informations au sein de documents audio

L'indexation de la parole et la recherche audio sont des thèmes qui ont été largement étudiés à la fin des années 1990, notamment lors des campagnes d'évaluations TREC, piste *Spoken Document Retrieval* SDR (Garofolo et al., 2000). La recherche de documents audio à partir d'une requête, elle-même audio ou écrite, peut correspondre à un calcul de similarités au niveau des caractéristiques acoustiques du signal — fréquence, rythme, silences, mélodie, locuteur — (Jolion, 2001; Cai et al., 2003; Bakker et Lew, 2002), des unités phonétiques de base (Jones et al., 1996), ou encore à une recherche en fonction des mots prononcés et du contenu informationnel. Ce dernier type de recherche n'est possible qu'au prix de coûteuses transcriptions humaines des enregistrements sonores ou lorsqu'un logiciel de transcription automatique est disponible (voir, entre autres, Nocéra et al. (2004) ou bien, pour une application aux *webcasts*, Munteanu et al. (2006)). Les différentes évaluations TREC ont montré que le taux d'erreur dans les transcriptions n'affectait pas trop les performances en recherche documentaire : une chute de 10 % de la précision pour un *Word Error Rate* (WER) de 50 % (Allan, 2001), à

mettre en relation avec les performances obtenues par les meilleurs systèmes actuels de reconnaissance de la parole qui obtiennent un WER de l'ordre de 10 % sur des données non bruitées. Pour une évaluation récente de l'effet des erreurs de transcriptions ou des techniques de compression du signal sur la qualité de l'indexation et de la recherche, on pourra se reporter à (Ranjan et al., 2006). Enfin, pour un aperçu général des questions posées par l'indexation multimédia (caractéristiques, formats dont MPEG-7), on lira (Bachimont, 2003).

Dans le cadre de son Master Recherche, co-encadré par Jean-François Bonastre et moi-même en 2003, Benoît Favre a étudié le comportement des modèles vectoriels et probabilistes de recherche d'informations lorsque le corpus cible contient à la fois des documents textuels classiques et des transcriptions issues de documents audio (Favre, 2003a,b; Favre et al., 2004a,b). Le corpus utilisé correspond à la réunion des quelques 500 000 documents (2,5 Go) de la piste *ad-hoc* de TREC-8 (Voorhees Ellen et Harman, 1999) avec la transcription des 500 heures de documents audio de la piste SDR (Garofolo et al., 2000). Il s'agit d'un corpus déséquilibré en volume mais cela reflète assez bien la réalité de la plupart des collections, web compris, où les transcriptions sont beaucoup moins nombreuses que les documents écrits. L'objectif était de mesurer les conséquences de ce déséquilibre lorsque l'utilisateur souhaite, à partir d'une requête (en l'occurrence les champs `DESC` des *topics ad-hoc* et SDR), obtenir des réponses mélangeant les deux modalités texte et parole.

Les statistiques que nous avons relevées sur ce corpus montrent qu'à partir d'un nombre de documents 25 fois plus grand et, en moyenne, d'une longueur 3 fois plus élevée pour la modalité "texte", la taille du vocabulaire est 18 fois plus importante. Dans le modèle vectoriel, ces différences ont des conséquences directes sur les valeurs de la composante tf. Cependant, des écarts majeurs sont trouvés sur les valeurs idf comme le montrent les figures 3.1 et 3.2. Ils soulignent une différence bien connue entre langage oral et langage écrit même si ce phénomène est réduit ici par le fait que les documents oraux ne sont pas de la parole spontanée. Dans ces conditions, le moteur de recherche a toutes les chances de sélectionner majoritairement des documents textuels par rapport aux documents audio.

Nous avons évalué la précision des réponses retournées par le système SMART (Buckley et al., 1996) après y avoir implémenté un équivalent de la fonction de score probabiliste BM-25 (p. 151). Parmi les 30 premiers résultats des requêtes *ad-hoc*, il n'y a que 2 % de documents issus de la modalité "parole". Cela pourrait s'expliquer par le fait qu'il s'agit de requêtes adaptées au corpus de la modalité "texte" mais il y en a à peine plus, 17 %, lorsque ce sont les requêtes de la piste SDR qui sont employées.

Nous avons alors proposé (Favre et al., 2004a) de réduire le déséquilibre entre les deux modalités en exploitant des techniques proches de celles de l'expansion de requêtes dans le modèle probabiliste. Ce type d'approche avait alors déjà été utilisé pour essayer de compenser les erreurs de transcription (Johnson et al., 2000) mais pas, à notre connaissance, le déséquilibre entre les modalités. Malheureusement, une évaluation reste toujours à faire faute de référentiel satisfaisant qui mélange les deux modalités. Bien sûr, des techniques employées pour la méta-recherche pourraient également

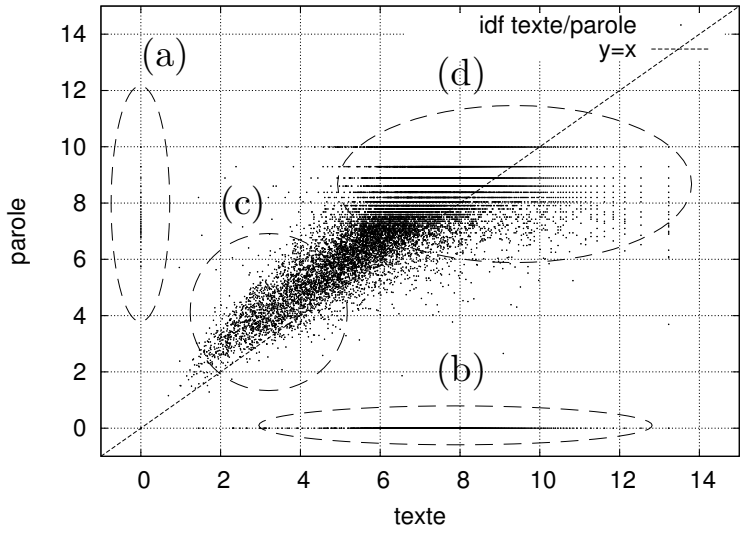
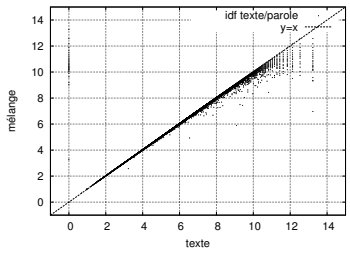
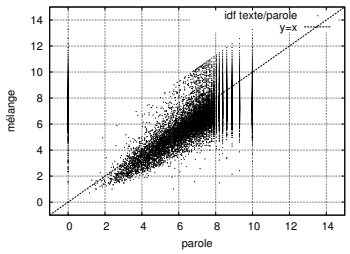


FIG. 3.1: *Différences entre les valeurs d'idf pour les modalités parole et texte. Le nuage central (c+d) représente les mots communs aux deux modalités. Plus un point est éloigné de l'axe $y = x$, plus les différences d'idf entre les deux modalités sont grandes. Figure tirée de (Favre et al., 2004a).*



texte comparé au mélange



parole comparé au mélange

FIG. 3.2: *Valeurs d'idf pour les modalités "texte" et "parole" au sein du corpus mélangé. Les valeurs idf de la modalité "texte" sont bien plus proches de celles calculées sur le corpus mélangé que les valeurs de la modalité "parole". En conséquence, le moteur de recherche a toute les chances de sélectionner majoritairement des documents textuels par rapport aux documents audio. Figure tirée de (Favre et al., 2004a).*

être utilisées afin de réduire le déséquilibre inhérent à tout mélange de corpus de tailles différentes.

3.2 Navigation dans des bases audio par résumé automatique

Une grande différence entre les données audio et le texte, est qu'il n'est pas possible pour un humain d'avoir un aperçu rapide du contenu audio en le survolant : une

écoute, coûteuse en temps, est nécessaire. Dans le cadre de sa thèse de Doctorat, Benoît Favre a proposé de réduire la durée d'écoute en supprimant la redondance informationnelle présente au sein des documents audio. Cette réduction est réalisée par l'intermédiaire d'un moteur de recherche et d'une interface utilisateur spécifique, à partir d'approches issues du résumé automatique et de la segmentation thématique présentée précédemment. Les différentes approches et composants qui constituent ce système sont brièvement présentés dans ce qui suit. Pour de plus amples informations, nous renvoyons le lecteur aux articles que nous avons publiés sur ce thème (Favre et al., 2006, 2007) ainsi qu'à la thèse de Doctorat de B.Favre (2007).

La question du développement d'interfaces utilisateurs *ad-hoc* pour manipuler aisément des documents audio est déjà ancien. Il est possible de trouver un état de l'art dans (Foote, 1999; Whittaker et al., 1999; Ranjan et al., 2006) ainsi que des exemples — figure 3.3 — d'application sur une messagerie ou des compte-rendus de réunions (Whittaker et al., 2002) ou encore des *webcasts*, figure 3.4 (Dufour et al., 2005). Pour une étude des usages, se reporter à (Lin et Smucker, 2008).

La figure 3.5, p.87, donne une idée de l'interface développée par B. Favre. On y distingue un champ de saisie de requête et une frise chronologique qui permet de visualiser les dates de parution des documents audio trouvés ainsi que les zones temporelles les plus denses (Favre, 2007). Elle permet la navigation selon plusieurs échelles temporelles (heures, jours, mois, années) en déplaçant les curseurs temporels à la souris. L'utilisateur a alors la possibilité d'entendre le document audio ou d'examiner un résumé automatique créé selon la procédure décrite dans la section suivante.

3.3 Une approche de résumé automatique audio orienté requête

Nous avons développé un système de recherche d'informations que nous avons testé sur les données audio issues de journaux radiophoniques de la campagne ESTER (Galliano et al., 2005). Ces documents ont tout d'abord été segmentées en classes acoustiques (Fredouille et al., 2004) puis en locuteurs (Istrate et al., 2005). Elles ont ensuite été transcrites (Nocéra et al., 2004), annotées selon une reconnaissance d'entités nommées (Favre et al., 2005) puis segmentées thématiquement en fonction d'indices de cohésion lexicale (voir chapitre 2) avant d'être sélectionnées en fonction de la requête (modèle probabiliste, type Okapi) et résumées par extraction de phrases.

Le résumé par extraction met en jeu une phase de sélection de segments. Il y a deux grands types d'approches : l'apprentissage des caractéristiques de phrases candidates à l'extraction sur des résumés manuels (Kupiec et al., 1995a) et la résolution d'un problème d'optimisation sur des critères informationnels, linguistiques et prosodiques (Kupiec et al., 1995b; Goldstein et al., 2000; Gong et Liu, 2001), voir figure 3.6. Nous nous sommes intéressé plus particulièrement à l'approche de Goldstein, connue sous le nom de *Maximal Marginal Relevance* (MMR), voir figure 3.7. C'est une approche qui détermine une sélection de segments de façon itérative : à chaque itération, le segment le plus similaire (au sens du modèle vectoriel de la recherche d'informations) au

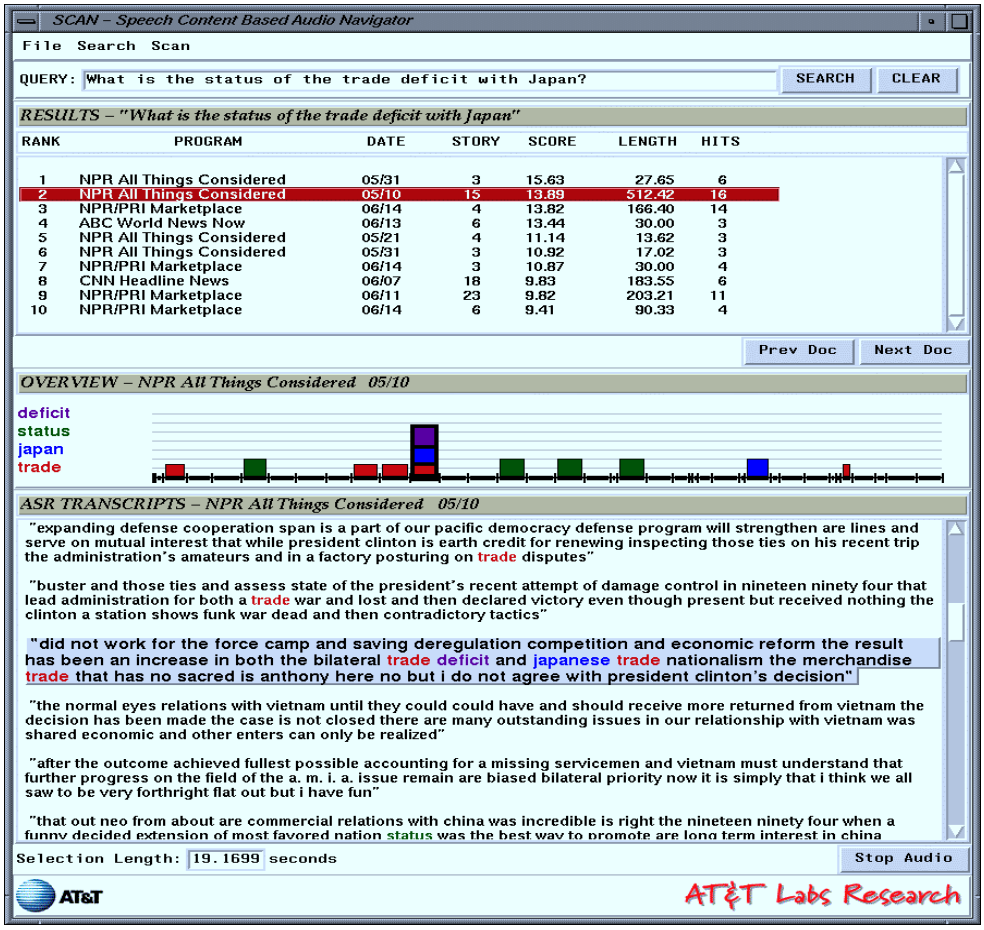


FIG. 3.3: *Interface graphique du moteur de recherche de documents audio SCAN (Whittaker et al., 1999). Les documents sont transcrits, les mots de la requête surlignés et leur présence est schématisée par un histogramme qui indique leur nombre d'occurrences paragraphe par paragraphe.*

centroïde et le moins similaire aux segments déjà sélectionnés, est ajouté à la sélection. Comme proposé par Murray et al. (2005), les scores de similarité destinés à appliquer l'approche MMR ont été calculés en utilisant une indexation sémantique latente (LSA) à l'aide de l'environnement logiciel *infomap-nlp*[1]. Afin de réduire le risque d'isoler dans le résumé une phrase qui deviendrait incompréhensible sans celle qui la précède dans le document intégral, le score d'une phrase en position i est pondéré par le score de la phase $i-1$ qui la précède.

Il existe deux types de phénomènes potentiellement générateurs d'erreurs : d'une part, les phénomènes intrinsèquement liés à la parole qui affectent le système même si la structuration est réalisée manuellement, et d'autre part, les phénomènes liés à l'incertitude de la structuration automatique.

Parmi les phénomènes intrinsèques, on trouve :

[1] http://infomap-nlp.sourceforge.net

FIG. 3.4: *Interface de navigation au sein de "webcasts" de présentations diapositives/vidéo comme proposée par Dufour et al. (2005) : on y voit une ligne temporelle représentant la chronologie entre les différentes diapositives et une table des matières construite dynamiquement.*

- les problèmes d'élocution (bégaiement, coupures, reprises, pauses) qui perturbent le contenu et nuisent à la compréhension ;
- le manque de grammaticalité à l'oral et notamment la notion toute relative de phrase ;
- la cohérence des dialogues (les tours de parole des locuteurs sont considérés comme une bonne base de segmentation mais il ne faut pas sélectionner une question posée par un journaliste et écarter la réponse qui lui a été donnée) ;
- les références (pronoms, titres, ...) à des entités citées qui perdent leur sens une fois isolées de leur contexte ;
- l'identité des locuteurs joue un rôle important lorsque l'on cherche à résumer les différents points de vue débattus ;

Chaque phase peut apporter son lot d'erreurs :

- les erreurs de transcription se retrouvent immédiatement dans l'analyse lexicale, les mots hors vocabulaire (inconnus du système de transcription) n'ont aucune chance d'apparaître dans les transcriptions ;
- l'impact des erreurs des différentes segmentations (classes acoustiques, locuteurs, thèmes) n'est pas mesurable directement, par contre, la segmentation en phrases a une grande importance dans le sens où elle peut générer des coupures, impliquant une forte réduction de la cohérence du contenu ;
- les concepts utilisés, les entités nommées, sont très adaptés au domaine d'application, notamment pour la génération de hiérarchies thématiques afin d'affiner les résultats ; il est évident que la qualité de leur extraction est essentielle à l'exploitation des données.

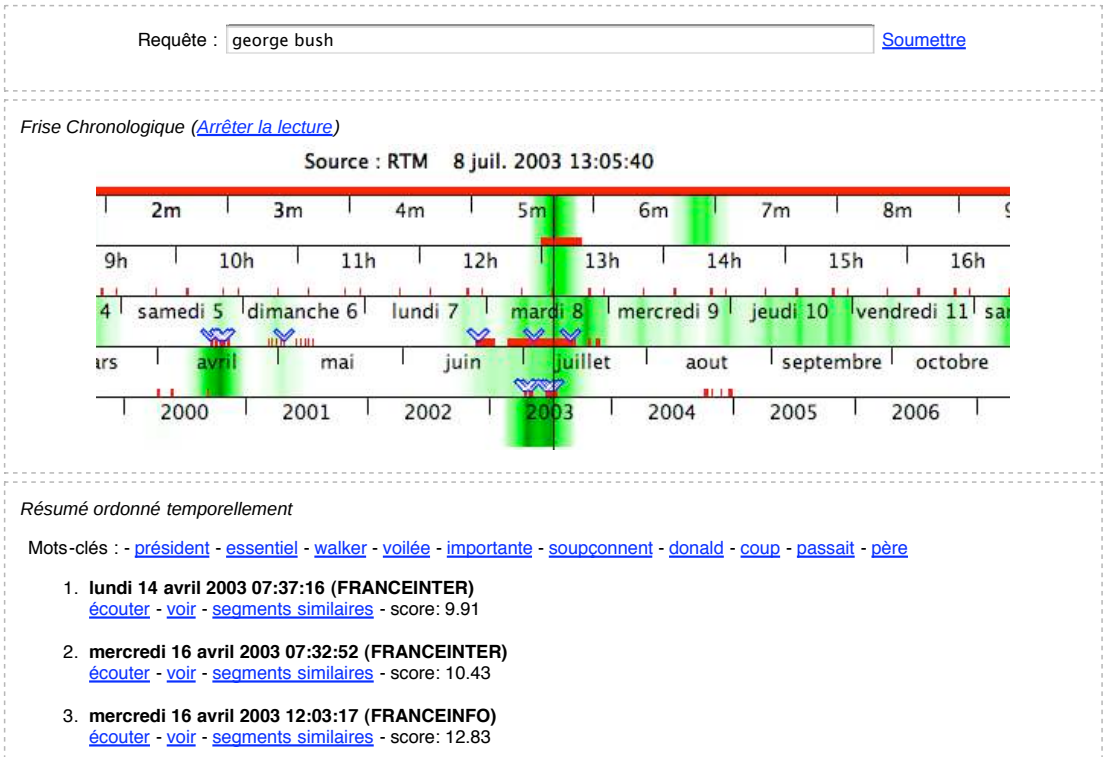


FIG. 3.5: *Interface de navigation du système de navigation, de résumé automatique et de recherche de documents audio créé par B. Favre durant sa thèse de Doctorat. Une frise chronologique permet de visualiser les dates de parution des documents trouvés ainsi que les zones temporelles les plus denses. Elle permet la navigation selon plusieurs échelles temporelles (heures, jours, mois, années) (Favre, 2007).*

3.4 Expériences durant DUC 2006

Une approche identique à celle qui vient d'être décrite a été testée avec succès durant les campagnes DUC 2006. Elle a été confrontée et exploitée conjointement à quatre autres systèmes du LIA : le système de résumé automatique Cortex (S2) (Torres-Moreno et al., 2005), une approche à base de modèles de langue (S3) *4-gram*, *4-lemmas*, *4-stems*, le module de recherche de passages du moteur questions-réponses SQuaLIA (S4) — section 1.3.2, p. 36 —, le module d'extraction de réponses de SQuaLIA (S5) — section 1.4, p. 40 —. Pour une description détaillée des approches suivies, voir (Favre et al., 2006).

La figure 3.8 donne un exemple de thème de recherche proposé durant la campagne DUC-2006. Les figures 3.9 et 3.10 donnent les résultats obtenus par nos différents systèmes, indépendamment les uns des autres et après fusion. Celle-ci est réalisée selon une stratégie qui consiste à identifier toutes les phrases retenues par au moins un système en les représentant, ordonnées, par autant de transducteurs à états finis (FST) reliés les uns aux autres et pondérés par une fonction de coût. Cette dernière est estimée

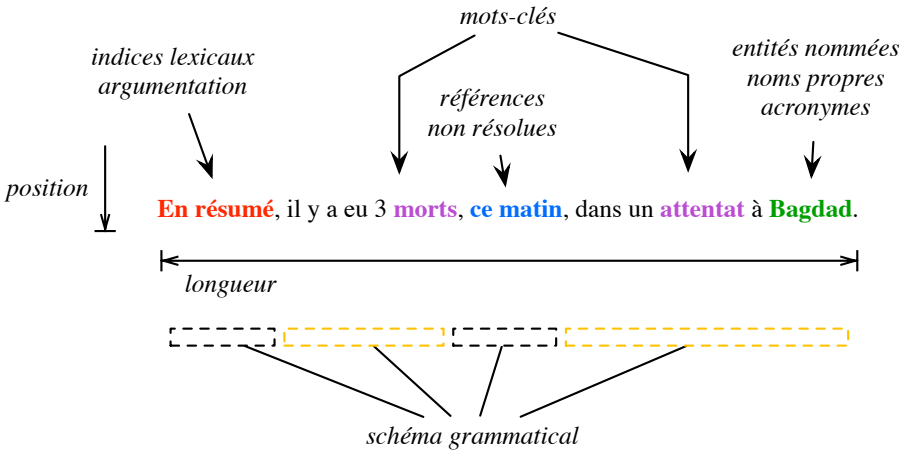


FIG. 3.6: *Indices linguistiques utilisés pour la sélection de phrases dans le résumé automatique (position dans le document, longueur, indices lexicaux, références anaphoriques, mots-clés porteurs du contenu, entités spécifiques, schéma grammatical). Figure tirée de (Favre, 2007).*

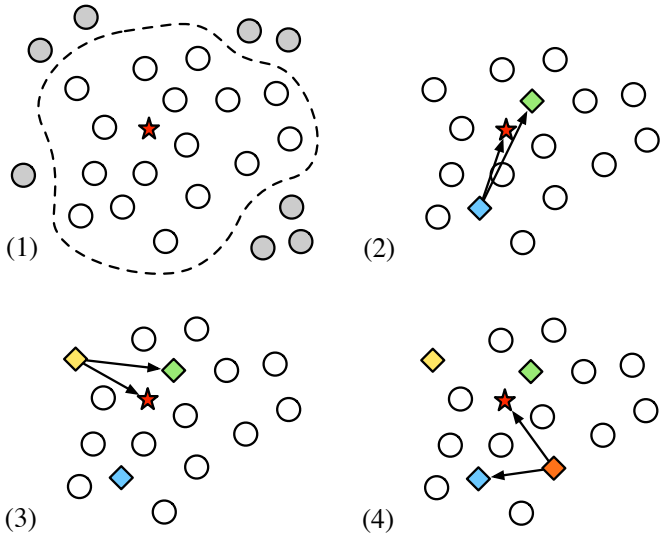


FIG. 3.7: *Illustration du fonctionnement de Maximal Marginal Relevance (MMR). La projection du besoin utilisateur est représentée par une étoile, les phrases candidates par des cercles et les phrases sélectionnées par des losanges. La première étape est d'écarter les phrases non pertinentes à l'aide, par exemple, d'une approche issue de la recherche documentaire (1). La première phrase sélectionnée est celle qui est la plus proche de la requête. Puis, les phrases sont sélectionnées itérativement en fonction de leur distance vis à vis de la requête, contrebalancée par leur redondance, elle-même estimée selon la distance avec la phrase déjà sélectionnée la plus proche (2,3 et 4). Figure tirée de (Favre, 2007).*

selon deux paramètres : le nombre de systèmes ayant sélectionné la phrase parmi ses premiers choix et le meilleur rang obtenu par la phrase pour l'ensemble des systèmes-. Il est à noter que les systèmes (S4) et (S5) sont entièrement non supervisés et qu'aucun paramètre n'a été adapté aux données de la campagne DUC. Enfin, quelques post-traitements ont été rajoutés de manière à résoudre certains acronymes, à réunir différentes écritures d'un même nom propre, à formater clairement les dates et à supprimer les phrases en doublon.

```
Num: d313e
Title: Development of Magnetic Levitation (MAGLEV) Rail Systems
Narrative:
In what countries are MAGLEV rail systems being proposed? Are the proposals for
short or long haul? Is government financing required for construction?
Granularity: specific
```

FIG. 3.8: *Exemple de thème de recherche issu de la campagne DUC 2006. Le titre définit la thématique tandis que la partie Narrative est constitué de questions précises.*

Les résultats selon les mesures d'évaluation ROUGE-2 et ROUGE-SU4 montrent que le système brièvement décrit dans ce chapitre est celui qui obtient les meilleurs scores sur les données de 2006 parmi les systèmes du LIA. Les résultats sur les données 2005 sont quant à eux biaisés car elles ont servi à paramétrer les systèmes S1 à S3. Remarquons que la fusion des 5 systèmes (F2) est celle qui obtient les meilleurs résultats, mieux que (S1) seul et mieux que (F1) qui réunit les 3 meilleurs systèmes sur les données 2005. Les résultats obtenus à partir d'une évaluation humaine restent bons puisque la fusion se classe 8^e sur l'ensemble des 34 soumissions (5^e et 6^e sur les tests automatiques).

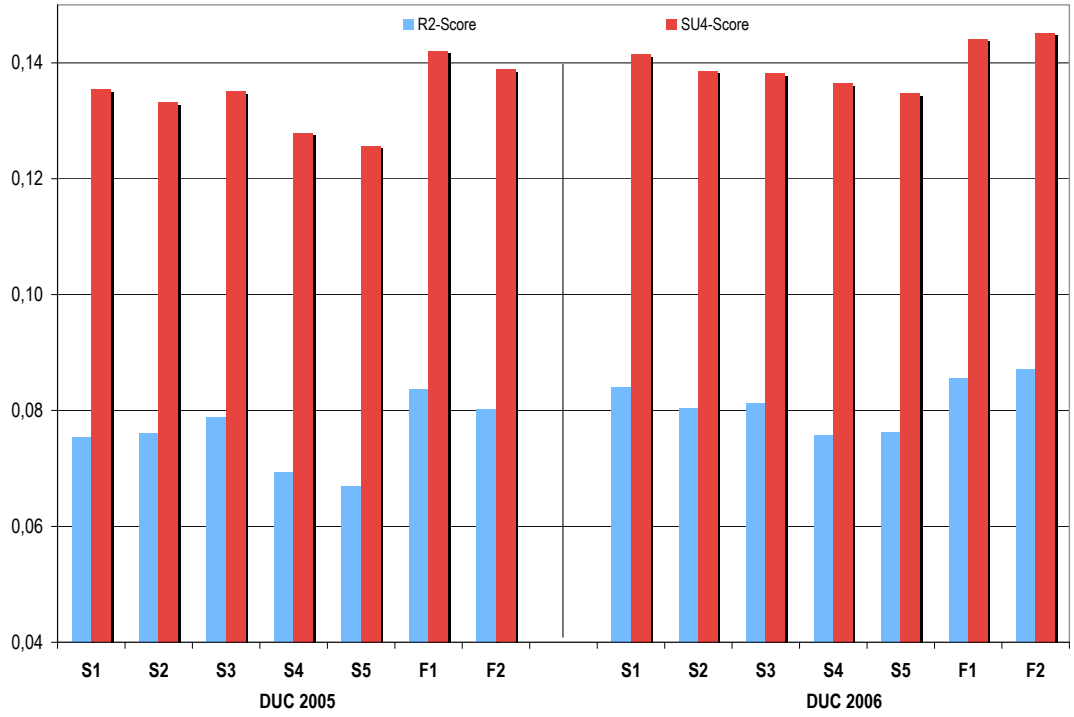


FIG. 3.9: *Scores Rouge 2 et Rouge SU4 sur les données de DUC 2005 et DUC 2006 pour les 5 systèmes LIA-Thales (S1 à S5), la fusion des 3 meilleurs (F1) et la fusion des 5 systèmes (F2). L'apprentissage sur le corpus 2005 est bénéfique car les systèmes non optimisés restent significativement moins performants que les autres sur DUC 2006. La fusion limite le sur-apprentissage lorsqu'elle est appliquée sur les 5 systèmes (F2 est meilleure sur DUC 2006). Figure tirée de (Favre et al., 2006).*

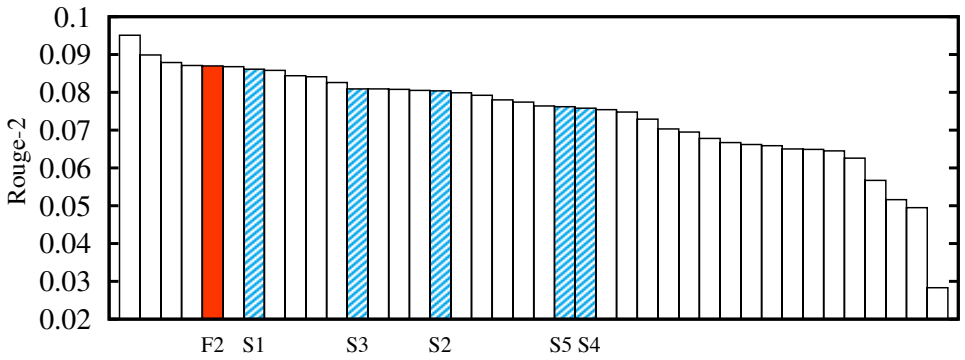


FIG. 3.10: *Classement des sous-systèmes S1 à S5 et de leur fusion F2 par rapport aux soumissions des autres participants à DUC 2006. Les scores sont exprimés selon la mesure Rouge-2. Le système S1 est à la hauteur des systèmes état de l'art en résumé automatique de texte. Les systèmes S4 et S5 sont issus de notre moteur de questions-réponses SQuaLIA et n'ont pas été spécifiquement adaptés à la tâche de résumé orienté requête (systèmes entièrement non supervisés). Figure tirée de (Favre, 2007).*

Chapitre 4

Recherche d'informations, lisibilité et dyslexie

Sommaire

4.1 Introduction 92
4.2 Modélisation cognitive de la lecture 94
4.3 Critères pour estimer la difficulté de lecture d'un texte 99
4.4 La dyslexie comme trouble du langage 104
4.5 Robustesse de SQuaLIA face à des requêtes bruitées 108
4.6 Hypothèses de réécriture de questions dysorthographiées 110
4.6.1 Données recueillies 111
4.6.2 Correction par phonétisation et retranscription 112
4.7 Lisibilité 116
4.7.1 Les mesures de lisibilité de Flesch 117
4.7.2 Réordonnancement des documents trouvés selon la mesure de Flesch 117
4.7.3 Proposition d'une mesure de lisibilité adaptée à la dyslexie . . . 118
4.8 Perspectives 121
4.8.1 Fonction de score d'un document combinant pertinence et lisibilité 121
4.8.2 Un processus spécifique d'expansion par retour de pertinence... et de lisibilité 123

NB. Ce chapitre correspond pour l'essentiel aux travaux réalisés durant la thèse de Doctorat de Laurianne Sitbon, titulaire d'une bourse BDI du CNRS (sections STIC et SHS), co-encadrée par Philippe Blache (LPL/CNRS-Université de Provence) et moi-même entre septembre 2004 et l'été 2007.

4.1 Introduction

Les travaux concernant les handicaps langagiers, la personnalisation et l'accessibilité doivent se pencher sur des problématiques concernant le langage, l'apprentissage et la cognition. Le cerveau n'est pas un ordinateur au sens où ce dernier est avant tout *logique*, c'est à dire basé sur des circuits électroniques où toute *action fondamentale* (primitive) est exprimable par des opérateurs booléens. Si la Logique permet de simuler l'activité du cerveau dans certains cas, voire des processus stochastiques, elle est moins adaptée pour rendre compte de nos interactions avec le monde extérieur, de nos émotions et sensations, qui sont l'une des sources essentielles de nos comportements et autant de stimuli continus. Les avancées les plus récentes de la technique et de l'imagerie médicale nous offrent des modélisations plausibles de nos fonctionnements cognitifs dont nous pouvons nous inspirer afin de simuler l'humain dans des domaines tels que le langage et la pensée. Et ceci à l'inverse des approches qui ont tenté de modéliser l'intelligence humaine à partir de la structure et l'architecture physique interne des machines. Ces avancées permettent de remettre au goût du jour les modélisations computationnelles neuronales même si les finalités poursuivies par la communauté des neurosciences ne sont pas identiques aux nôtres : nous essayons de faire en sorte de développer des logiciels *utiles* et non de modéliser le comportement humain. Les modèles neuronaux sont apparus dans les années 1960 avant de disparaître suite à la preuve que les perceptrons de l'époque ne pouvaient accomplir des tâches de classification pourtant élémentaires (Minsky et Papert, 1969). Ils ont resurgi dans les années 1980 à la suite des travaux de Paul Werbos qui a proposé dans sa thèse de doctorat en 1974 un algorithme de rétro-propagation pour l'apprentissage de réseaux comportant plusieurs couches (Werbos, 1994). Cet algorithme a ensuite été popularisé par Rumelhart et al. (1985) et il est à la base de la majorité, si ce n'est de la totalité, des systèmes de recherche d'informations connexionnistes (Kwok, 1989; Boughanem et Soulé-Dupuy, 1997) ainsi que des modèles d'apprentissage de la lecture qui ont fait l'objet d'implémentations informatiques (voir section suivante).

Nous allons nous intéresser dans ce chapitre plus particulièrement aux modèles cognitifs de la lecture et à la tentative de les exploiter afin de définir des systèmes de recherche d'informations capables d'estimer l'effort nécessaire à la compréhension d'un document et d'être suffisamment robustes pour accepter des requêtes incorrectement orthographiées.

S'il existe de nombreux travaux autour des moteurs de recherche d'informations interactifs, de grandes lacunes concernent leur adaptation contextuelle à des utilisateurs aux capacités en écriture ou en lecture limitées. Il peut s'agir de personnes atteintes de pathologies (dyslexie, mauvaise vision, ...) mais aussi, plus simplement, de personnes ne maîtrisant pas suffisamment la langue d'un document en consultation. De manière générale, la prise en compte du *contexte* et l'adaptation aux utilisateurs en recherche d'informations fait l'objet de nombreuses conférences : *Information Retrieval in Context*[1] durant SIGIR 2004 (Ingwersen et Belkin, 2004), *Adaptive Information Retrieval*[2] durant

[1] http://ir.dcs.gla.ac.uk/context/
[2] http://www.dcs.gla.ac.uk/workshops/air2008/

la conférence IIiX 2008, *NLP for Reading and Writing*[3] durant la conférence SLTC 2008 à Stockholm... Par ailleurs, des groupes d'études ont été formés afin de permettre l'accès au Web par des personnes handicapées. C'est le cas de la *Web Accessibility Initiative* (W3C, 2001) qui dresse une liste d'utilisations potentielles du Web et préconise certaines solutions techniques. Par exemple, une personne atteinte de dyslexie sera aidée si l'on ajoute aux documents des représentations graphiques et si l'on rend immobiles les animations tandis qu'une personne daltonienne souhaitera gérer elle-même les couleurs d'affichage. Une personne ayant des problèmes d'acuité visuelle appréciera l'interfaçage d'un module de synthèse de la parole tandis qu'une autre qui ne peut se servir d'un clavier standard emploiera un outil de reconnaissance de la parole pour saisir ses requêtes (Scott et Galan, 1998; Fairweather et al., 2002).

Dans ce cadre, la personnalisation de la recherche d'informations et la prise en compte des caractéristiques cognitives individuelles des utilisateurs est l'une des problématiques majeures. Les modèles de recherche d'informations usuels permettent d'ordonner des documents en fonction de la quantité d'informations qu'ils véhiculent vis à vis de ce que l'utilisateur a exprimé dans sa requête tout en tenant compte, dans le meilleur des cas, du taux de nouveautés apportées par rapport à d'autres documents déjà connus (Allan, 2002). Il s'agit d'une vision purement informationnelle de la pertinence posant l'hypothèse que plus le *nombre* d'informations nouvelles est grand, plus le document est susceptible d'intéresser l'utilisateur. Cela s'avère exact dans une certaine mesure mais ne tient pas compte du fait que les besoins sont différents suivant le niveau d'expertise de l'utilisateur : une personne novice dans un domaine sera certainement plus intéressée par un document de vulgarisation que par une étude approfondie au vocabulaire et à la structure complexes. Ainsi, de nombreuses études se sont penchées très tôt sur la notion de *pertinence* en tentant de la définir en fonction de paramètres le plus souvent extra-linguistiques et contextuels, non explicites dans une requête (Mizzaro, 1997). Cela est vrai à plus forte raison, pour des personnes ayant des difficultés élevées de lecture. Il s'agit alors de définir de nouvelles mesures prenant en compte cet aspect tout en offrant la possibilité de présenter d'abord les documents les plus "simples", les plus "lisibles". Notons que cette fonctionnalité peut aussi être profitable pour des adultes ayant des capacités en lecture et écriture normales et pour des enfants en phase d'apprentissage.

Pour ce faire, nous devons dans un premier temps clairement définir ce que nous entendons par *lisibilité*. Cette notion est étroitement liée à la caractérisation d'un profil utilisateur, lui-même fonction de son niveau de connaissance du domaine et de la langue du document ; autrement dit, de ses capacités de lecture. S'il existe un *continuum* évident depuis la personne analphabète ou illettrée jusqu'au lecteur expert qui peut être reflété par les nombreux tests de lecture disponibles, nous avons choisi de travailler sur un handicap courant, la dyslexie. Dans un deuxième temps, les caractérisations des handicaps entraînant des déficits en lecture et écriture (voir par exemple Rossignol, 2001; Rey et al., 2001) doivent être exploitées en étudiant comment ils peuvent se traduire au niveau d'implémentations informatiques. Celles-ci peuvent être destinées à l'aide à la détection ou à la remédiation des handicaps étudiés mais aussi, plus modes-

[3] http://spraakbanken.gu.se/personal/sofie/SLTC_2008/SLTC_2008.html

tement, à l'adaptation de logiciels basés sur des interactions textuelles, orales ou écrites. Certaines estimations font état qu'entre 3 et 9 % de la population adulte ou en âge d'être scolarisée connaît des difficultés importantes dans l'apprentissage de la lecture (Ducrot et Nguyen, 2003) leur rendant d'autant plus complexe la manipulation d'outils informatisés. Un effort particulier doit être entrepris afin de faciliter l'accès à "l'information" pour ces personnes et, *a fortiori*, pour celles présentant un handicap plus important.

À titre d'exemple, l'étude exposée dans Bruza et al. (2000) qui mesure l'effort cognitif correspondant à différents modes de recherche d'information ainsi que le logiciel de traitement de textes pour dyslexiques décrit par Dickinson et al. (2002) peuvent servir de points de départ, notamment pour l'assistance dans la formulation de requêtes. Pour des handicaps plus extrêmes, ces dernières peuvent ne pas être constituées uniquement de mots mais aussi de symboles ou d'images tel que cela est réalisé dans les plateformes de communication alternative — voir par exemple les logiciels de la société AEGYS d'aide à la communication verbale et non verbale[4] (Bellengier et al., 2004; Blache et Rauzy, 2007, 2008) ou encore VITIPI (Boissière et Dours, 2000) développé à l'IRIT.

Il est enfin particulièrement important d'établir un mode évaluatoire spécifique pour chaque module informatique développé : définitions de métriques autres que les classiques rappel et précision mais aussi constitution de données de test dans la voie des travaux de Jansen et Kroner (2003). Des évaluations dans des conditions réelles doivent être réalisées, par exemple pour des personnes atteintes de dysphonésie (dyslexie entraînant la reconnaissance visuelle d'un faible nombre de mots seulement) et/ou de dyseidésie (dyslexie visuelle rendant la lecture très lente). Il s'agit par exemple de vérifier que les méthodes proposées permettent aux utilisateurs de trouver plus rapidement l'information qu'ils recherchent, et, pour la recherche documentaire, d'accéder à un plus grand nombre de documents pertinents qu'en employant un moteur de recherche non dédié.

4.2 Modélisation cognitive de la lecture

La simple présentation des faits ne suffit pas pour pouvoir acquérir le langage... Il nous faut découvrir ce qui est indispensable pour faire fonctionner le système.

(N. Chomsky)

Nous allons tenter dans cette section et la suivante de définir les critères objectifs, et éventuellement subjectifs, qui peuvent permettre d'estimer la lisibilité d'un texte (les capacités de lecture nécessaires) en exploitant les modélisations psychocognitives et neurocognitives les plus récentes.

De nombreux modèles de la lecture ont été proposés depuis une quarantaine d'années. Ferrand (2007) en dresse une liste exhaustive depuis le modèle Logogène de Morton (1969) dans lequel un *détecteur* cognitif spécifique est associé à chaque mot dans un

[4] http://aegys.fr

lexique mental jusqu'aux récents modèles à double voies en cascade ou connexionnistes incorporant un codage phonologique (Seidenberg et McClelland, 1989; Coltheart et al., 2001; Perry et al., 2007) et permettant des simulations informatiques performantes. Historiquement, une des questions soulevées par la compréhension des processus en jeu durant la lecture (silencieuse ou non), concerne le rôle de l'information phonologique et la manière dont celle-ci est utilisée pour accéder à la compréhension du mot et, le cas échéant, à leur prononciation. En ce qui nous concerne, ces modèles sont intéressants car ils permettent d'envisager des moyens de simuler par ordinateur les processus de la lecture humaine, et de tenter de distinguer ainsi un texte facile d'un texte difficile. Remarquons qu'il est *a priori* possible de lire (prononcer) sans comprendre et, à l'inverse, de comprendre sans lire mot à mot. La quasi totalité des chercheurs considère aujourd'hui qu'un codage phonologique est obligatoire et automatique dans la lecture silencieuse (Ferrand, 2007), seule son importance varie suivant la langue ou selon le mot considéré.

Parmi les modèles qui simulent le mieux la lecture experte, figure le modèle à double voies (Coltheart, 1978) dont une évolution est schématisée en figure 4.1. Selon ce modèle, le *décodage* d'un mot peut se faire selon :

- *une voie lexicale* — le mot est *reconnu instantanément* sans découpage ni identification explicite de ses constituants (morphèmes, syllabes...) par concordance avec une entrée d'un lexique visuel puis mis en correspondance avec un lexique phonologique[5] (route B sur la figure) et le système sémantique (route A) de compréhension — ; ,
- *une voie non lexicale* — un ensemble de règles, *apprises* de conversion graphèmes-phonèmes permet de phonétiser le mot dans le système phonémique (route C) qui est lui-même connecté au système sémantique afin de permettre la compréhension du mot.

La route lexicale est considérée comme la voie privilégiée de la lecture experte, la plus rapide, tandis que la route non lexicale est essentielle pour la lecture de mots irréguliers ou de non-mots (déchiffrage). Parmi d'autres, Southwood et Chatterjee (2000) ont posé l'hypothèse d'activation simultanée des différentes voies (lexicales et non lexicale). La prise de décision lexicale (système phonémique) se fait alors en fonction des contraintes phonologiques et de celles imposées par la tâche et le contexte : contrainte temporelle dans le cas d'une lecture à haute voix (le débit doit être maîtrisé), contrainte sémantique pour la compréhension.

Toujours selon ce modèle, une rétroaction d'activation peut avoir lieu entre le niveau phonologique et le niveau orthographique pour le traitement des homophones (flèches en pointillés sur la figure 4.1 entre le système phonémique, le lexique phonologique et le lexique visuel et l'unité sémantique d'autre part). Par exemple, la distinction entre les mots *cygne* et *signe* pourrait se produire uniquement après identification des phonèmes correspondants puis détection d'une ambiguïté sémantique possible et retour vers l'orthographe pour une analyse plus précise et un accès au lexique visuel (route

[5] Une connexion directe entre le lexique visuel et le lexique phonologique peut s'illustrer par l'exemple qui consiste à considérer un mot dont on connaît la prononciation (on s'en *souvient*) mais dont on a oublié le sens.

lexicale). D'autre part, l'accès au système sémantique se fait vraisemblablement à la fois par l'usage du code orthographique (voie lexicale) et du code phonologique (voie non lexicale) comme cela est montré par de nombreuses études citées dans Ferrand (2007) : (Coltheart et Coltheart, 1997; Harm et Seidenberg, 2004). Une bonne illustration de l'accès systématique au système sémantique a été proposée par Stroop (1935) qui relevait la difficulté que l'on a à énoncer la couleur avec laquelle est écrit un mot lorsqu'il désigne lui-même le nom d'une autre couleur (par exemple, dire que le mot *jaune*, écrit en bleu, est bleu). Les interactions entre les voies lexicales et non lexicales sont visibles lorsque l'on est amené à lire un mot tel que *Monsieur* qui doit être *reconnu* avant de pouvoir être prononcé ou encore *fille/ville* et *chœur/chou* qui comportent des graphèmes identiques aux prononciations différentes.

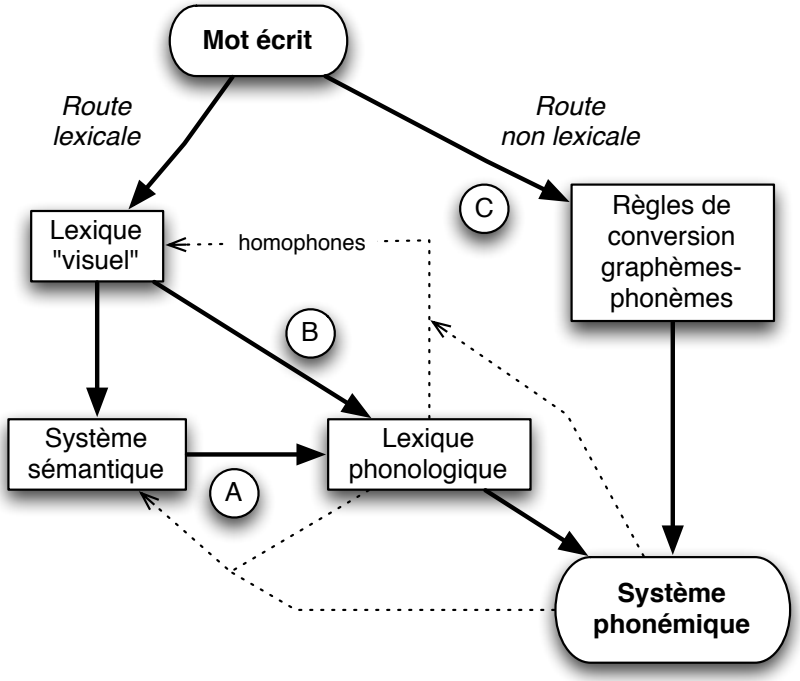


FIG. 4.1: *Les différentes routes de la lecture experte permettant de passer du mot écrit à une séquence de phonèmes et à sa compréhension. Le modèle présenté ici est le modèle "dual route" (A) et (C) étendu avec une troisième route (B) par Southwood et Chatterjee (2000). Il permet par exemple d'expliquer pourquoi certaines personnes ayant une voie directe (C) déficiente parviennent à lire sans difficulté les mots fonctionnels du langage (mots outils) malgré leur représentation sémantique réduite.*

Une extension du modèle à double voies a été proposée par Coltheart et al. (2001) dans laquelle les lexiques visuel et phonologique ainsi que les systèmes sémantique et phonémique sont reliés à double sens par des connexions activatrices ou inhibitrices continues (par opposition à l'aspect strictement séquentiel et en tout-ou-rien du modèle originel). Selon ce modèle, les unités lexicales sont pondérées selon leur fréquence d'occurrence dans le lexique visuel et les différentes routes sont activées de manière simultanée et parallèle. Il suppose de nouveau l'existence de règles de conversion expli-

cites et symboliques. Selon les simulations informatiques réalisées par Coltheart et al. (2001), ce modèle permet de définir les difficultés dans l'apprentissage de la lecture comme liées à une capacité plus ou moins grande à acquérir l'un des composants de cette architecture (Sprenger-Charolles et Colé, 2003) ou bien à suivre l'une des routes les connectant.

Par opposition aux modèles à double voies qui supposent l'existence de *règles* de conversion explicites, le modèle connexionniste *à traitement parallèle distribué (PDP)* de Seidenberg et McClelland (1989) propose une mise en correspondance entre les mots écrits, le système phonémique et le système sémantique, progressive et établie selon des régularités statistiques. Selon ce modèle (figure 4.2), il n'y a plus de référence à plusieurs voies distinctes consacrées aux mots réguliers pour l'une, et aux irréguliers pour l'autre, l'interconnexion forte entre les nœuds du réseau neuronal permettant de représenter parallèlement le traitement de l'ensemble des mots et des non-mots. L'ensemble du système est activé simultanément, les différences entre les niveaux d'activation conduisant, par propagation et selon les potentiels de chaque nœud, à la prise de décision lexicale.

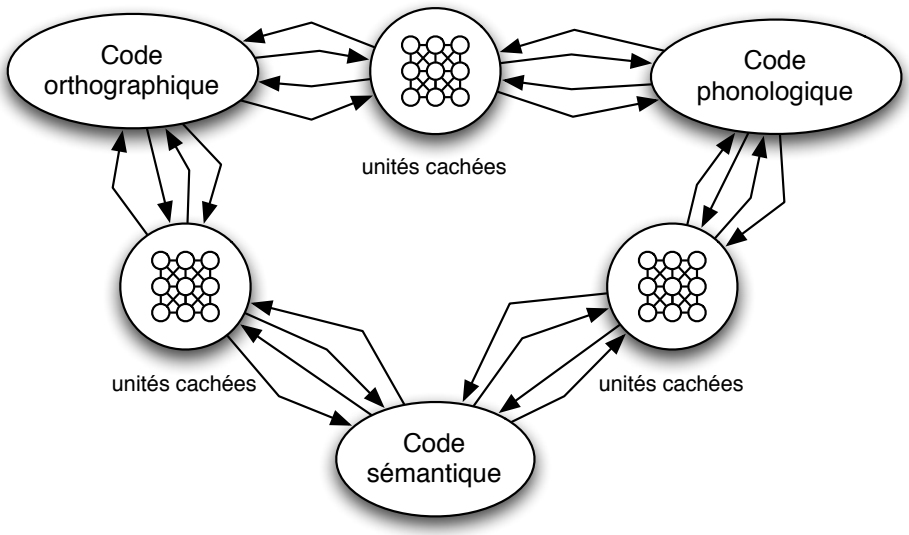


FIG. 4.2: *Le modèle connexionniste parallèle distribué (PDP) proposé par Seidenberg et McClelland (1989) puis étendu par Harm et Seidenberg (2004) suppose un apprentissage progressif des mises en correspondance orthographiques, phonétiques et sémantiques.*

Si le modèle connexionniste de Seidenberg et McClelland (1989) est séduisant par sa simplicité apparente, il ne permet pas de rendre compte des difficultés plus ou moins grandes éprouvées par certaines personnes à lire des mots inconnus ou irréguliers mais familiers. L'absence de séparation entre deux voies de la lecture ne permet pas de représenter, au sein du modèle, ces familles de mots comme étant de nature différente des mots réguliers. La figure 4.3 reproduit le modèle implémenté par Brown (1997) qui met en correspondance des triplets de lettres et des triplets de phonèmes selon un réseau en trois couches et en faisant varier le nombre d'unités cachées dans la couche

intermédiaire. Selon les valeurs de ce nombre, ils parviennent à simuler quelques unes des conséquences de la dyslexie : plus le nombre d'unités cachées est faible, plus la lecture de non-mots réguliers est difficile. Ils proposent ainsi de mettre en relation directe, complexité structurelle du réseau neuronal et capacité de lecture. À partir de là, il serait possible d'estimer le niveau de lisibilité d'un texte en fonction de la complexité nécessaire (nombre d'unités cachées) du réseau pour obtenir une lecture correcte. Nous n'avons trouvé ce potentiel exprimé dans aucun article mais pensons qu'elle pourrait s'avérer efficace dans le cadre qui nous occupe.

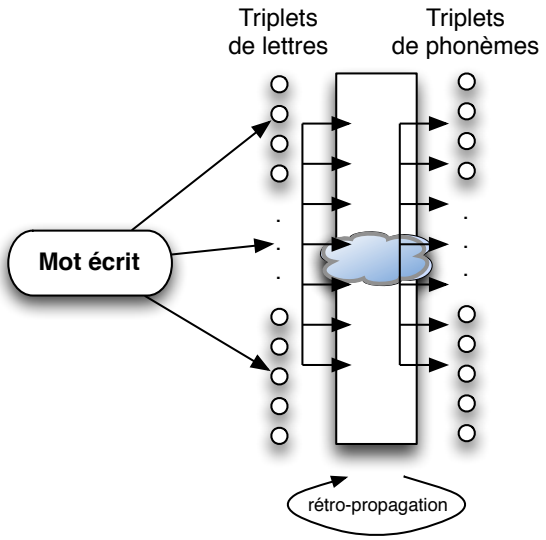


FIG. 4.3: *Un modèle connexionniste simplifié de la lecture permettant la mise en correspondance entre un mot écrit isolé de son contexte et une séquence de phonèmes (Brown, 1997). La variation du nombre d'unités cachées dans la couche intermédiaire permet de simuler certaines difficultés de lecture éprouvées par des personnes dyslexiques. Nous proposons qu'il soit un critère afin de déterminer le niveau de lisibilité d'un texte.*

Ce modèle a été largement étendu par Harm et Seidenberg (1999) puis par Perry et al. (2007) en y intégrant distinctement les deux routes du modèle de Coltheart et al. (2001) mais en conservant les propriétés structurelles des réseaux connexionnistes (apprentissage statistique par rétro-propagation, niveaux d'activation continus, règles non explicites). Les simulations informatiques ont montré des performances élevées de ce modèle en tâche de prononciation puisqu'il est capable de prononcer correctement près de 99 % des mots qui lui sont présentés et près de 94 % des non-mots (Perry et al., 2007). Ferrand (2007) ajoute que ce modèle permet en outre de simuler de nombreux effets selon des critères qui seront énumérés dans la section suivante : effet de voisinage orthographique, effet de fréquence, de longueur, d'amorce *etc.*

Il faut toutefois pondérer ces performances en relevant que ces modèles ne permettent de simuler, au moins actuellement, que des mots monosyllabiques et indépendamment des contextes gauche (historique) et droit (empan lexical)[6].

[6] Les systèmes de synthèse de la parole sont nettement plus complexes et performants sur ces différents

4.3 Critères pour estimer la difficulté de lecture d'un texte

La majorité des critères ci-dessous sont tirés de Sprenger-Charolles et Colé (2003); Ferrand (2007); Dehaene (2007) en considérant que la *difficulté* d'un mot peut être mise en correspondance avec le temps nécessaire à son identification. Ils nous serviront ultérieurement à définir une mesure de lisibilité que l'on exploitera en recherche d'informations.

Les graphèmes. Ils sont définis comme étant la représentation écrite d'un phonème (voir page 139). Ils peuvent être constitués d'une ou plusieurs lettres (ex. "*ch*" en anglais ou en français). Remarquons que pour certaines langues, telles que le russe à l'exception de certains signes d'accentuation, tous les graphèmes sont constitués d'une seule lettre. Pour un nombre de lettres égal, les mots ayant un faible nombre de graphèmes sont plus longs à être lus que les autres (Rastle et Coltheart, 1998; Rey et al., 1998, 2000). Cela peut s'expliquer par le fait que les syllabes sont plus complexes à isoler et que la correspondance entre les lettres et les sons n'est pas directe. Cela semble être surtout le cas pour les mots non fréquents pour lesquels la voie lexicale directe est moins significative.

La correspondance lettres-graphèmes-phonèmes. Suivant les langues considérées, la correspondance entre les lettres, les graphèmes et les phonèmes est plus moins transparente, un corollaire de la transparence étant que deux mots proches orthographiquement ont tendance à être prononcés de façon similaire. Parmi les langues alphabétiques, le français — environ 10 % des mots sont irréguliers (Ziegler et al., 1996) — et surtout l'anglais — environ 30 % de mots irréguliers (Ziegler et al., 1996) — présentent de nombreuses irrégularités, l'anglais essentiellement sur les voyelles et le français sur les consonnes finales[7]. L'impact de la régularité sur le processus de décision lexicale n'est pas clairement identifié même s'il semble qu'il soit réel pour les mots peu fréquents pour lesquels le décodage phonétique est plus important par rapport à un accès lexical direct (Seidenberg et al., 1984).

points puisqu'ils tiennent compte du contexte et simulent l'intonation (voir par exemple (Véronis et al., 1998) et Iida et al., 2003). Néanmoins les approches employées (voir Dutoit (1997); Yvon et al. (1998); Donovan (2003); Bellegarda (2005) pour un aperçu des principales méthodes et de leurs évaluations) ne permettent pas de, et ne sont pas conçus pour, modéliser la phase d'apprentissage de la lecture ni de souligner les effets d'activation ou d'inhibition dans la prise de décision lexicale tels que l'on peut les observer chez des humains face à des mots plus ou moins complexes. Ils ne permettent pas non plus de schématiser les impacts des atteintes physiologiques ou des handicaps tels que la dyslexie. Pour la seule conversion graphèmes-phonèmes (indépendamment de toute prédiction prosodique) dans des langues alphabétiques telles que le français, l'anglais ou l'italien, un lexique phonétique de formes fléchies couplé à un analyseur morpho-syntaxique et à un ensemble de règles de conversion peut s'avérer suffisant. Cela n'empêche pas que des systèmes à base de réseaux de neurones ont été implémentés pour le module de conversion graphèmes-phonèmes (Sejnowski et Rosenberg, 1987) mais ce problème n'est désormais plus vraiment au cœur des travaux en synthèse de la parole.

[7] On parle de *mot régulier* lorsque la prononciation est prédictible à partir de l'orthographe et de *mot irrégulier* dans le cas contraire.

L'attaque et la rime. Elles se situent à un niveau intermédiaire entre la syllabe et les lettres. L'*attaque* est le regroupement de consonnes, ou la consonne, en début de syllabe et la *rime* ce qui suit, décomposé en noyau vocalique (voyelles) et en *coda* (consonnes finales). S'il n'a pas été montré d'effet de latence spécifique suivant la nature de l'attaque ou de la rime, elles jouent néanmoins un rôle significatif dans la lecture comme unités clairement identifiées : la lecture est difficile lorsque les mots sont coupés (technique de masquage) à des positions inhabituelles (Grainger et Ferrand, 1996).

Le voisinage orthographique. Selon la définition donnée par Coltheart et al. (1977), il correspond, pour un mot donné, à l'ensemble des autres mots de la langue qui lui sont identiques à une lettre près et à la même position (*vendre, rendre*). La taille du voisinage et la fréquence d'usage des mots du voisinage sur le processus de décision lexicale sont des critères qui ont souvent été étudiés mais les conclusions sur leur impact potentiel divergent et sont souvent contradictoires (Ferrand, 2007). Il semblerait qu'un mot ayant un voisinage orthographique avec des mots de haute fréquence soit plus long à être reconnu (décision lexicale). La taille du voisinage exercerait également un rôle inhibiteur, potentiellement plus fort en français qu'en anglais. Relevons que la définition du voisinage orthographique peut être étendue à un voisinage syllabique et non plus graphémique. Enfin, il pourrait se définir de manière discrète en considérant des voisins plus ou moins *proches* suivant le nombre de lettres, ou de syllabes, qu'ils ont en commun.

Le voisinage phonologique. Le voisinage orthographique entraîne le plus souvent une proximité phonologique (ceci est d'autant plus vrai que la langue considérée est *régulière*) et l'on peut se demander si les effets inhibiteurs ou facilitateurs du voisinage sont d'ordre orthographique et/ou phonologique voire phono-graphique (voir figure 4.4). Lorsque le nombre de voisins orthographiques est maintenu constant, Yates et al. (2004) ont obtenu un effet de facilitation en décision lexicale lorsque le nombre de voisins phonologiques est élevé. Toutefois, là encore, les conclusions de ces études sont sujet à caution tant il est difficile de s'affranchir de biais expérimentaux et d'isoler une variable non corrélée avec d'autres. Dans cet ordre d'idée, le nombre d'homophones est un critère à considérer.

La longueur des mots. Certaines études indiquent que les mots longs sont plus difficiles à reconnaître que les mots courts tandis que d'autres concluent à une absence d'effet du nombre de lettres arguant que les mots longs sont aussi des mots rares pour lesquels l'accès lexical direct est rendu difficile et qui favorise un décodage phonologique plus lent. Notons que si l'on considère que les deux voies sont systématiquement utilisées par des normo-lecteurs, une différence dans la durée de décision lexicale n'est alors possible que si la décision est prise alors que les deux voies n'ont pas encore été entièrement suivies (autrement dit qu'il n'y a pas de synchronisation entre elles). Cela n'est pas incompatible avec le modèle connexionniste où les connexions entre les

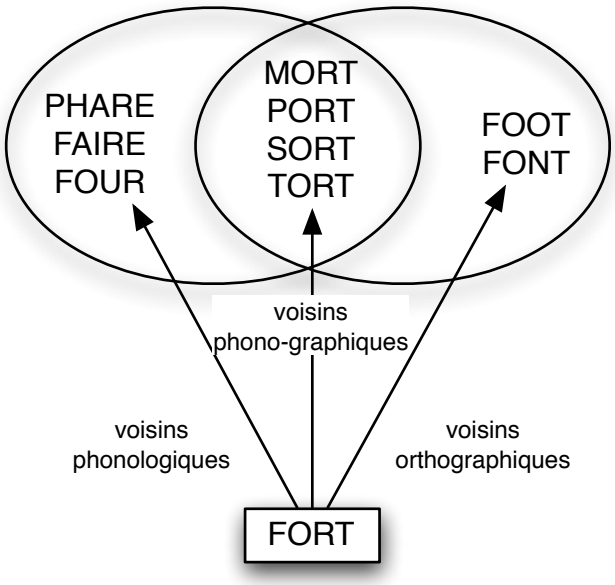


FIG. 4.4: *Exemple de voisinages (orthographique, phono-graphique et phonologique) du mot FORT en français (figure reprise de Ferrand (2007), p. 149)*

formes visuelles et les mots du lexique sont permanentes et associées à un poids. Stopper le processus de décision lexicale en cours de route correspond à considérer les probabilités lexicales en cet instant même si elles sont susceptibles d'évoluer encore durant le processus. Le nombre de lettres étant fortement corrélé avec le nombre de phonèmes et de syllabes (voir ci-dessous), il est d'autant plus difficile d'en mesurer l'effet indépendamment des autres critères. Notons que New et al. (2006) ont relevé un effet inhibiteur pour les mots de plus de 9 lettres ou de moins de 5 lettres sur les données du lexique *The English Lexicon Project*[8] contenant près de 3 millions de temps de réactions collectés à partir de plus de 40 000 mots et de 40 000 non-mots. L'effet de la longueur des mots semble moins important au fur et à mesure de l'augmentation de l'âge et du niveau scolaire. Il peut par contre être très important chez les personnes dyslexiques qui *épellent* (Zoccolotti et al., 2005).

Le nombre de syllabes. Pour des mots de basse fréquence, Ferrand et New (2003) ont montré que plus le mot contenait de syllabes, plus la décision lexicale était longue à prendre. Aussi bien pour le français que pour l'anglais, certaines études soulignent que cet effet est indépendant du nombre de lettres dans le mot, du nombre de voisins orthographiques voire même de la fréquence lexicale. Cet effet dénote un mode visuel de lecture (Ashby et Rayner (2004) cité par Ferrand (2007)) où la syllabe est reconnue immédiatement sans découpage en lettres puis regroupement soulignant un traitement parallèle des lettres conduisant à l'indentification syllabique. Ce rôle est toutefois régulièrement remis en question par d'autres études.

[8] http://elexicon.wustl.edu

La fréquence d'occurrence. Elle correspond au nombre de fois que le lecteur a rencontré le mot considéré (on parle de *fréquence de surface* lorsqu'une seule forme du mot est considérée dans le décompte, par opposition à la *fréquence cumulée* où l'ensemble de ses flexions ou dérivations est dénombrée). Si cette valeur est difficile à calculer autrement que de manière subjective en demandant au lecteur de classer les mots en classes d'équivalence suivant son degré de familiarité avec le mot, une approximation peut être donnée à partir de l'analyse d'ouvrages littéraires ou de documents journalistiques. C'est ce type d'informations que l'on peut trouver dans la base LEXIQUE[9] (New et al., 2001). De très nombreuses études (voir Ferrand (2007), p. 91, pour une liste particulièrement riche) ont montré que l'identification lexicale, la prononciation (de façon moins nette, les mots pouvant être prononcés sans indentification lexicale en exploitant uniquement les correspondances phonologiques) ou encore la catégorisation sémantique étaient plus rapide pour les mots fréquents que pour les mots rares aussi bien pour le français que pour l'anglais, l'hébreu, le serbo-croate ou le chinois (Frost et al., 1987). Remarquons à cette occasion que l'anglais est une langue où les nombreuses irrégularités dans la correspondance graphèmes-phonèmes rendent fréquent l'usage des deux voies de lecture, lexicale et non lexicale, tandis que l'hébreu, aux correspondances encore bien plus complexes, utilise obligatoirement la voie lexicale contrairement au serbo-croate, à l'italien et à l'espagnol, langues très régulières.

La familiarité. La fréquence d'occurrence d'un mot dans la langue est une valeur complexe à estimer car dépendante du corpus utilisé (journaux, ouvrages, web, langue orale / écrite *etc.*). En outre, il paraît naturel que le corpus qui serait idéalement à prendre en compte devrait être lié aux habitudes de lecture de la personne pour laquelle on souhaite estimer la lisibilité d'un texte. Certaines études ont permis d'observer l'impact de la *familiarité* d'un mot sur la lecture en demandant aux lecteurs de l'estimer de manière subjective. Il est alors observé que cette dernière constitue un critère prédictif plus pertinent pour le temps de décision lexicale que la seule fréquence d'occurrence objective (Gernsbacher, 1984) et ce même s'il s'agit de valeurs corrélées. Par contre, il ne s'agit pas d'un facteur décisif pour la prononciation contrairement à l'âge d'acquisition du mot (voir ci-dessous) comme cela a été relevé par Brown et Watson (1987).

La répétition à court ou long terme. Si l'effet de répétition exerce à long terme une action proche de celle de la fréquence d'occurrence ou de la familiarité, il semble qu'un effet à court terme existe également. Ceci semble être vérifié aussi bien pour des mots réels que pour des non-mots (par exemple *freper*) qui sont lus une deuxième fois plus vite lorsqu'ils ont été rencontrés dans un passé très proche et ce d'autant plus que la tâche de décision lexicale est rendue difficile (alternance de polices de caractères par exemple). Cette hypothèse de *trace épisodique*, sorte de mémoire-cache, a été formulée par Bodner et Masson (1997). Ces effets de répétition sont à mettre en relation avec les différents effets d'amorce, morphologique ou sémantique, trouvés par ailleurs.

[9] http://www.lexique.org

L'âge d'acquisition du mot. Il est défini comme étant l'âge où le mot a été *appris* pour la première fois. Morrison et Ellis (1995) ont montré que plus les mots ont été acquis tardivement, plus il est difficile de les prononcer et ceci indépendamment de l'effet de fréquence (Bonin, 2007). Il s'agit d'une mesure difficile à estimer. Parmi les méthodes proposées figure celle consistant à analyser les livres pour enfants selon une classification qui suit l'âge ciblé. À ce titre, le lexique Manulex (Lété et al., 2004) constitue une ressource précieuse qui permet en outre d'établir une *trajectoire fréquentielle* des mots au cours du temps. Lorsque la trajectoire est descendante, c'est à dire lorsqu'il s'agit de mots moins souvent rencontrés avec l'âge, la tâche de décision lexicale pourrait être plus ardue avec le temps que dans le cas contraire où elle serait facilitée. Cela tendrait à souligner un effet adaptatif même si l'état initial du système de reconnaissance reste primordial et que la *plasticité* du réseau connexionniste diminue avec l'âge. Remarquons qu'un tel effet a été vérifié expérimentalement pour la tâche de décision lexicale ou la dénomination d'objets mais pas pour la prononciation de mots écrits (Bonin et al., 2004).

La morphologie et la fréquence cumulée. Le morphème est défini comme étant la plus petite unité de sens à l'intérieur du mot et correspond aussi bien à l'étude des *racines* qu'aux *dérivations* et *flexions*. L'influence de la taille et de la fréquence de la famille morphologique du mot sur le temps de décision lexicale fait débat mais ils peuvent également être des critères importants et il est probable qu'un mot fréquemment au pluriel permette de reconnaître plus rapidement la forme au singulier par rapport à un autre mot dont les formes pluriel et singulier sont rares. Il va de soi que la corrélation de ces paramètres avec la fréquence cumulée est forte et que les expériences sur le temps de lecture ne permettent pas de décider si une aire spécifique du cerveau est dédiée à l'analyse morphologique.

La concrétude et l'imageabilité. Les mots concrets (ceux qui correspondent à une personne, un objet, un lieu *etc.*) sont lus plus rapidement que les mots abstraits (Bleasdale, 1987) ce qui tend à prouver l'existence d'un lexique imagé.

La polysémie et l'homonymie. Aussi bien pour des tâches de décision lexicale que de prononciation, il a été montré que les mots ayant plusieurs sens (par exemple *table* ou *clé*) étaient identifiés plus rapidement que les mots non ambigus, voir par exemple (Pexman et al., 2004). Il est probable que cet effet puisse s'expliquer à la fois par une quantité d'informations sémantiques associées plus importante rendant leur *attraction* supérieure et par des fréquences cumulées élevées. Par contre, la *compréhension* est ralentie lorsque les mots lus ont de nombreux sens candidats. Cependant, il faut distinguer les mots polysémiques des homonymes qui, eux aussi, ont plusieurs sens mais qui ne sont pas reliés sémantiquement entre eux (par exemple *avocat*). Les temps de décision lexicales sont significativement plus importants pour les homonymes que pour les autres mots (Beretta et al., 2005).

Par contre, la forme graphique globale des mots n'est pas un critère déterminant pour le temps de décision lexicale (Perea et Rosa, 2002) même si certaines paires de lettres sont plus souvent confondues que d'autres ("O" et "Q" contrairement à "S" et "W") et que les lettres minuscules sont lues 5 % plus rapidement que les lettres majuscules (Ferrand, 2007).

On l'aura compris, une des difficultés dans la sélection des meilleurs critères de lisibilité réside dans la mise en œuvre d'expérimentations fiables et généralisables ainsi que dans la forte corrélation pouvant exister entre eux (par exemple entre la familiarité, la fréquence d'occurrence et l'âge d'acquisition). Enfin, des impacts différents sont observés suivant que la tâche consiste en une prise de décision lexicale, étroitement liée à la *compréhension*, ou à la prononciation du mot lu. Ils sont encore différents chez des personnes atteintes de trouble du langage tels que la dyslexie pour lesquels l'altération des voies de lecture est plus ou moins prononcée. Dans l'objectif d'établir une mesure de la lisibilité, nous devons déterminer de quelle manière les différents critères que l'on vient d'énoncer interagissent et concourent à un score commun. Malheureusement, les expériences conduites depuis de nombreuses années confirment que le temps de lecture ne peut être relié linéairement à ces différents critères. En section 4.7, nous proposerons toutefois une mesure de lisibilité adaptée à des enfants dyslexiques et nous l'évaluerons dans le cadre de la recherche d'informations.

4.4 La dyslexie comme trouble du langage

Les troubles du langage sont très nombreux et peuvent toucher la communication en général, la production de parole spontanée, la lecture à haute voix ou silencieuse, la compréhension de l'écrit ou encore l'écriture. Pour une description de ces multiples affections, le lecteur pourra se référer à (Danon-Boileau, 2004; Campolini, 2004) ; concernant spécifiquement les troubles de la lecture et la dyslexie : (Messerschmitt et Flohic, 2002; Dehaene, 2007) ou encore (Habib, 1997) pour un point de vue plus neurologique et neurophysiologique ; sur les rapports entre cerveau et connaissance et les dernières avancées des sciences du cerveau : (Edelman, 2007; Changeux, 2002) ; sur l'apprentissage du langage : (Piaget et Chomsky, 1979; Bruner, 1983; Chomsky, 2000).

Citons, à titre d'illustration de la diversité des troubles possibles, une étude conduite en 1887. Cette étude, considérée par certains comme étant à l'origine de la neurologie, a été conduite auprès d'un homme ayant souffert d'une incapacité soudaine à lire (Déjerine, 1892). Il pouvait malgré cet handicap brutal recopier un texte, très lentement, mais était incapable de le relire, à tel point qu'il ne fallait l'interrompre durant l'écriture, sous peine de ne pouvoir reprendre, ne sachant où il en était. Il était par ailleurs capable de reconnaître son journal grâce à sa mise en page, de reconnaître que telle forme référait à des lettres mais sans avoir aucune idée des phonèmes auxquelles elles pouvaient correspondre et, enfin, pouvait toujours lire des nombres et faire des calculs. Dehaene (2007) rapporte, d'après Déjerine (1892), qu'une autopsie pratiquée à la mort du patient après une seconde attaque cérébrale, a montré des lésions de la partie postérieure de l'hémisphère gauche du cerveau alors que la partie droite était intacte.

Parmi les nombreux enseignements récents des neurosciences, relevons le fait que l'alphabétisation modifie l'anatomie du cerveau, que les dyslexies correspondent à des déficits physiologiques identifiés (Ziegler, 2006; Monaghan et Shillcock, 2008) — voir figure 4.5 — (on écarte ici le cas des enfants qui ont des difficultés à acquérir la lecture indépendamment d'une quelconque affection physiologique et que l'on nomme à tort, d'un point de vue neurocognitif, *dyslexiques*, mais à raison d'un point de vue social) et qu'une rééducation adaptée permet, au moins en partie, de réactiver des zones du cerveau *endormies*— voir figure 4.6 — (Dehaene, 2007). Rappelons au passage que de nombreuses difficultés d'apprentissage de la lecture peuvent être liées à une scolarisation irrégulière, une mauvaise maîtrise de la langue orale, un environnement peu stimulant, des troubles psychologiques ou encore une déficience auditive. Les définitions les plus récentes de la dyslexie se positionnent sur des critères précis et indépendants du niveau intellectuel *i.e.* de la capacité à comprendre un texte dans sa globalité :

> *[...] la recherche est maintenant en mesure de fournir des indicateurs comportementaux fiables des processus en œuvre dans la lecture permettant de diagnostiquer un dyslexique. La manifestation la plus probante d'une dyslexie réside dans l'échec à développer des capacités de reconnaissance des mots écrits en dehors de tout contexte.*
>
> *[...] selon l'Observation National de la Lecture, est considéré comme dyslexique un mauvais lecteur "chez qui le déficit résulte, en partie en tout cas, d'une anomalie de la capacité d'identification des mots écrits. L'origine de cette anomalie se trouve dans les structures cérébrales et cognitives qui sous-tendent cette capacité.*
>
> *[...] Quelques rares enquêtes épidémiologiques permettent de penser qu'environ 5 % des enfants sont dyslexiques, c'est-à-dire, approximativement un par classe.*
>
> Sprenger-Charolles et Colé (2003), p. 132.

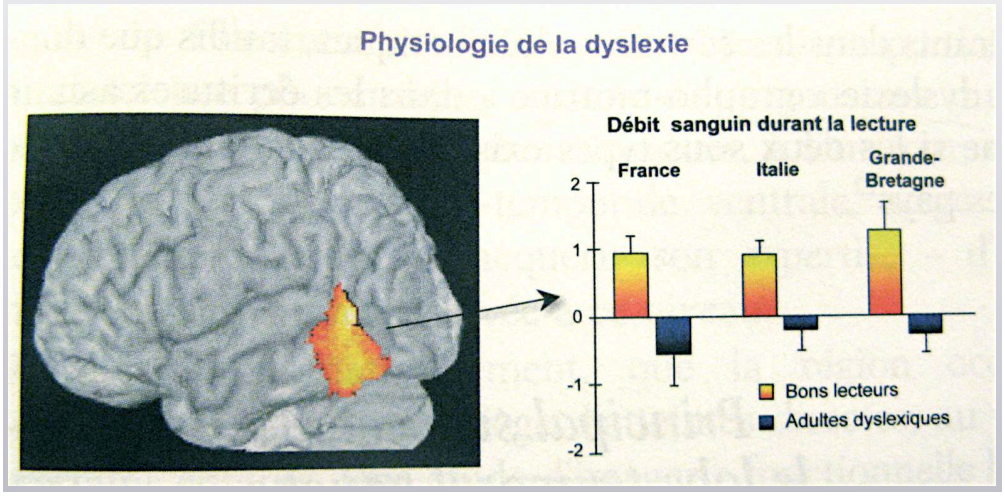


FIG. 4.5: *Le débit sanguin durant la lecture est différent chez des normo-lecteurs et chez des adultes dyslexiques. Illustration reprise de (Dehaene, 2007), elle-même issue de (Paulesu et al., 2001)*

De nombreux types de dyslexies sont ainsi recensés (et parfois contestés, ou du

moins, discutés) qui touchent plus ou moins gravement certaines aptitudes de lecture[10]. Parmi les dyslexies plus courantes, citons, en se référant aux différentes routes de la lectur (figure 4.1, p. 96) :

- *la dyslexie dyseidétique* (ou *dyslexie de surface*) : lecture lente sans altération de la qualité de la compréhension au moins pour les mots réguliers (la faculté à lire des mots inconnus est plus ou moins atteinte et en tout cas très ralentie). Elle correspond à une perte de l'accès direct au sens (route A). La route (C) est privilégiée. Par exemple *femme* est lu *fèm* et le mot peut être alors non reconnu ;
- *la dyslexie dysphonique* : lecture à vitesse normale mais certains mots sont substitués par d'autres sans altération profonde du sens (par exemple *viande* est lu lorsque *jambon* est écrit). La route (A) est efficace contrairement à la route (B) ;
- *la dyslexie phonologique* : incapacité à prononcer de nouveaux mots ou des non-mots. La route (C) est déficiente. Cette forme de dyslexie est parfois associée à une aphasie voire à une difficulté à nommer les objets vus (système sémantique lexical dégradé) ;
- *la dyslexie profonde* : il s'agit d'une forme aggravée de dyslexie phonologique qui touche les trois routes (A), (B) et (C) et engendre une incapacité à lire des mots nouveaux (perte de la capacité à convertir des graphèmes en phonèmes), des erreurs sémantiques et phonétiques.

En lien avec les critères de lisibilité mentionnés dans la section précédente, énonçons maintenant quelques remarques autour de la dyslexie :

- si l'effet de fréquence est constaté chez une personne dyslexique comme chez un normo-lecteur, c'est à dire, si les mots fréquents sont plus rapidement lus que les mots rares et ceci indépendamment des autres caractéristiques du mot telles que la longueur, cela témoigne de la possibilité d'utiliser au moins partiellement la voie lexicale (routes B et C) ;
- les effets de régularité graphèmes/phonèmes sont constatés aussi bien chez des personnes atteinte de dyslexie développementale que chez des normo-lecteurs. Autrement dit, un mot régulier est toujours lu plus rapidement qu'un mot irrégulier ou encore qu'un non-mot (Brown, 1997). Cela signifie que si déficit phonologique il y a, la lecture peut tout de même se faire en empruntant des voies

[10] Etant donné que les effets extérieurs de la dyslexie peuvent ressembler, du point de vue de l'observation, à ceux des personnes apprenant à lire, nous sommes amenés à nous intéresser au mode d'apprentissage de la lecture tel que pratiqué en milieu scolaire. Jusqu'au milieu des années quatre-vingt dix, l'apprentissage de la lecture était souvent considéré comme un apprentissage réalisé en trois étapes (Frith, 1985) plus ou moins communément admises :

1. reconnaissance globale des lettres (stade logographique) en tant que symboles abstraits spécifiques ;
2. reconnaissance de la structure interne des mots (syllabes, attaque et rime, graphèmes et morphèmes) et de la position des lettres mais aussi acquisition des premières correspondances phonologiques (stade alphabétique) : cette reconnaissance peut s'avérer suffisante pour la lecture de mots pour lesquels la correspondance entre les graphèmes et les phonèmes (voir les figures 6.3 et 6.4 en pages 139 et 140) est régulière pour la langue considérée ;
3. reconnaissance *instantanée* des mots sans conversion phonologique explicite (stade orthographique).

Depuis, il s'est avéré que ce découpage séquentiel n'était pas aussi net dans la réalité et que les différents stades pouvaient être atteints en parallèle (Sprenger-Charolles et Colé, 2003) et les processus cognitifs sous-jacents définitivement conservés et utilisés même chez un lecteur expert.

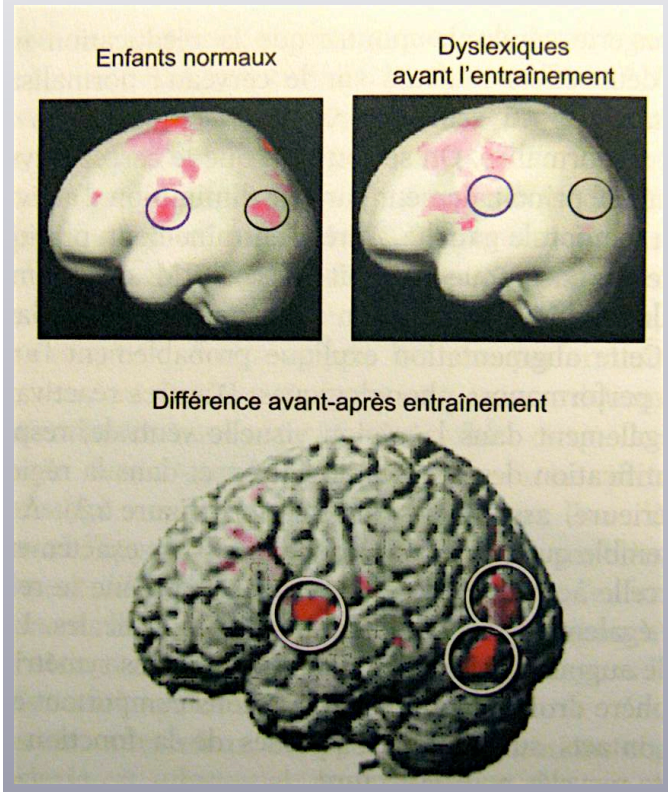


FIG. 4.6: *Une rééducation adaptée permet de restaurer une activité cérébrale proche de la normale. Mais les méthodes sont à adapter à chaque personne. Illustration reprise de (Dehaene, 2007), elle-même issue de (Temple et al., 2003).*

différentes parmi les routes (A), (B) et (C) ;
- la longueur des mots est un facteur plus important pour la personne dyslexique que pour un normo-lecteur (Zoccolotti et al., 2005; Fiset et al., 2006) ;
- les dyslexiques profonds ont de grandes difficultés à identifier les mots abstraits alors que les mots concrets sont lus correctement (Marchand et Friedman, 2005) ;
- le modèle à double voie de Coltheart et al. (2001) (voir figure 4.1) simule efficacement deux types de dyslexie, la dyslexie phonologique (route non lexicale déficiente) et le dyslexie de surface (trouble de la route lexicale).

Dysorthographie. Les dyslexies se traduisent fréquemment en dysorthographies, troubles du langage *écrit*. Selon Sprenger-Charolles et Colé (2003), certains enfants peuvent n'écrire correctement que les mots qu'ils connaissent par cœur ou bien éprouver des problèmes de mémorisation qui provoquent des confusions phonologiques. D'autres auteurs cités par Jumel (2005) remarquent des confusions auditives, omissions ou ajouts de lettres, ou encore inversions de lettres ou de syllabes. Enfin, des problèmes de segmentation de mots sont constatés (phénomène d'agglutination ou, au contraire, de subdivision) ainsi

qu'un nombre supérieur à la normale d'erreurs grammaticales[11].

Dans les sections suivantes, nous nous intéressons aux conséquences que peuvent avoir des requêtes mal orthographiées pour un moteur de questions-réponses (dont le cas extrême est la dysorthographie) puis à la définition d'une mesure de lisibilité pour réordonner les documents trouvés par un système de recherche d'informations. Ces deux problèmes, écriture et orthographe d'une part, et lisibilité d'autre part, peuvent concerner tout utilisateur qui ne pourra accéder à ce qu'il recherche. Mais les personnes touchées par une dyslexie ou une dysorthographie seront bien plus fortement handicapées. Nous essayerons d'adapter au mieux les systèmes à ces besoins qui ne sont pas spécifiques mais plus fortement exprimés.

4.5 Robustesse de SQuaLIA face à des requêtes bruitées

Les ateliers *Analytics for Noisy Unstructured Text Data* qui ont eu lieu durant les conférences IJCAI 2007[12] et SIGIR 2008[13] ont été l'occasion de réfléchir aux problèmes rencontrés lorsqu'un système automatique est confronté à des documents bruités tels que ceux issus de processus de reconnaissance automatique de caractères (le taux d'erreurs mots peut aller de 2 % à plus de 50 % en fonction des polices de caractères ou de la qualité d'impression), de SMS, de courriels ou de blogs, ou encore de transcriptions automatiques de requêtes orales (taux d'erreurs mots souvent aux alentours de 30 à 40 %). Pour chacun de ces types de données, des corpus de référence en anglais peuvent être exploités[14].

Dans le cadre de sa thèse (Sitbon, 2007), Laurianne Sitbon s'est intéressé à évaluer les performances de notre moteur de questions-réponses SQuaLIA lorsque des questions réelles lui sont posées (Sitbon et al., 2008a). Nous avons ensuite proposé une stratégie conduisant à formuler des hypothèses de correction orthographique des questions dans le cadre général (Sitbon et al., 2007a) puis dans le cas spécifique d'utilisateurs dyslexiques et dysorthographiques (Sitbon et Bellot, 2008a). Nous présentons dans cette section et dans la suivante les principaux résultats obtenus concernant l'évaluation de SQuaLIA et des propositions de réécritures.

Evaluation de SQuaLIA à partir de questions réelles

À partir d'une interface web développée pour les besoins de l'expérimentation, 17 personnes ont saisi des questions à partir de 20 besoins en informations définis depuis une sélection de questions de la campagne EQUER pour lesquels notre système de

[11] Ces dernières peuvent être tout simplement liées à un problème de mémorisation provenant d'efforts trop importants consacrés à l'écriture, générateurs d'étourderies.

[12] http://research.ihost.com/and2007/index.html

[13] http://and2008workshop.googlepages.com/

[14] http://research.ihost.com/and2007/data.html

questions-réponses SQuaLIA fournit une réponse correcte et supportée. Parmi les personnes ayant accepté de participer à cette évaluation, 11 ont la langue française comme langue maternelle (dont 2 sont dyslexiques) et 6 vivent en France mais ont pour langue maternelle le chinois, l'allemand ou l'espagnol. L'objectif de cette étude était d'évaluer globalement SQuaLIA sur des questions réelles mais également d'estimer les baisses de performance de chacun des modules qui le composent (catégorisation de la question, recherche de documents, recherche de passages, extraction de la réponse). Les 20 questions sélectionnées concernent la recherche de noms de personnes, de nombres, de dates, de lieux ou encore d'âge ou de distance et contiennent des noms propres courants ou rares (figure 4.7). Ces questions ont conduit à définir des besoins en information. Par exemple, la question *"Comment s'appelle le Président Tchétchène ?"* est transformée en une instruction donnée oralement aux participants de la forme : *"Demandez le nom du Président Tchétchène."*. La figure 4.8 donne des exemples de questions écrites par les 17 participants. Globalement, chaque utilisateur a fait au moins une fois une erreur d'écriture sur un nom propre. Ce taux va jusqu'à 39 % des noms propres pour les personnes pour lesquelles le Français n'est pas la langue maternelle soulignant par là une difficulté particulière bien connue dans la traduction. Les erreurs syntaxiques concernent surtout des fautes d'accords et, rarement, l'ordre des mots.

Référence	Libellé de la question
GF222	Qui est le maire de Bastia ?
GF30	Combien de personnes souffrent d'acné en Suisse ?
GF266	Quelle est la monnaie nationale en *Hongrie* ?
GF219	A combien de kilomètres de Paris se trouve la gare de Tours ?
GF178	Comment s'appelle le président Tchétchène ?
GF6	Quel âge a l'abbé Pierre ?
GF232	Combien y a t il de chômeurs en Europe ?
GF245	Qui est le frère de la princesse Leia ?
GF17	Combien y a t il d'habitants en *Lettonie* ?
GF29	Quel grade occupe *Juan Carlos Rolon* dans la marine ?
GF273	Quand est mort *Kurt Cobain* ?
GF105	Quel est le nom du roi du *Maroc* ?
GF298	Quelle est la capitale de Terre Neuve ?
GF147	En quelle année *Hitler* est arrivé au pouvoir ?
GF176	Qui est le président d'*Aérospatiale* ?
GF99	Combien de personnes sont mortes dans des accidents de la route en 1997 ?
GF132	Où se situe *San Cristobal* de Las Casas ?
GF206	Quand a été votée la loi *Evin* ?
GF84	En combien de langues a été publié le Petit Prince ?
GF78	Quel journal publie chaque année le top 50 des personnalités ?

FIG. 4.7: *Les 20 questions d'EQUER utilisées comme base dans le cadre d'une expérimentation de la robustesse de SQuaLIA. Ces 20 questions ont permis de définir oralement des besoins en information à partir desquels des utilisateurs ont écrit de nouvelles requêtes*

Si les 20 questions originelles étaient toutes correctement étiquetées (le système n'aurait pas pu répondre correctement sinon), seules 64 % des *questions réelles* l'étaient. Parmi les 36 % restant, 20 % se sont vu attribuer un type incorrect et 16 % pas de type du tout ou bien un type générique rendant l'extraction d'une réponse correcte plus difficile. Pour 18 % des questions mal étiquetées, aucune erreur orthographique n'était

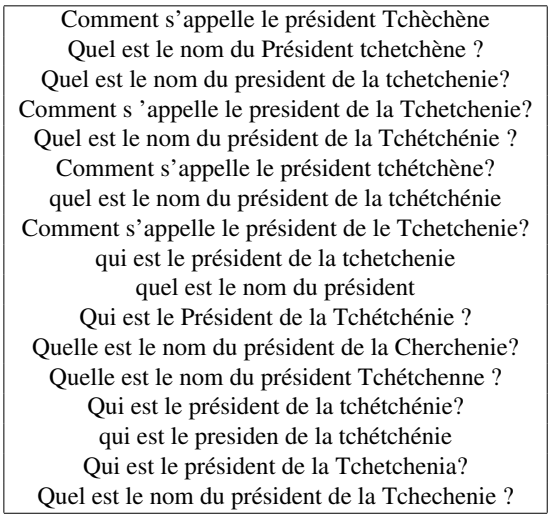


FIG. 4.8: *Exemple de questions écrites par des utilisateurs à partir du besoin en information "Demandez le nom du président tchétchène.*

présente. Ce cas de figure correspond à des formulations absentes des modèles utilisés pour l'étiquetage. Près de la moitié des questions mal étiquetées contiennent une ou plusieurs erreurs d'accent, de syntaxe ou d'orthographe des noms propres. Après l'étape de catégorisation, ce ne sont plus que 80 % des questions qui peuvent espérer recevoir une réponse correcte. L'évaluation des étapes de recherche documentaire ou de recherche de passage est par la suite plus complexe à effectuer étant donné que le module d'extraction de réponse ne se contente pas de fonctionner sur la première solution (document ou passage) fournie. Si l'on examine l'ensemble des passages retenus à partir des questions réelles, seules 67 % d'entre elles peuvent encore recevoir une réponse correcte. Le module final d'extraction de la réponse fait chuter ce score à 60 %. Grâce à la redondance et à la variabilité des manières d'écrire les noms propres, y compris dans le corpus de la campagne EQUER constitué d'articles du quotidien Le Monde, 56 % des questions avec un nom propre mal orthographié obtiennent tout de même une bonne réponse, ce qui n'est le cas que de 31 % des questions ayant au moins une faute d'orthographe portant sur un autre mot qu'un nom propre. Au final (figure 4.9), SQuaLIA ne répond correctement en moyenne qu'à 12 questions par utilisateur, 10 pour les dyslexiques et 14 pour les francophones natifs.

4.6 Hypothèses de réécriture de questions dysorthographiées

Selon une approche expérimentale similaire à celle présentée dans la section précédente, cinq questions de la campagne EQUER ont servi à définir des besoins en information qui ont été soumis à 19 enfants dyslexiques. Ces cinq questions ont la propriété de ne contenir que des mots appartenant au lexique Manulex (Lété et al., 2004), c'est à dire censés être connus par des enfants d'âge scolaire élémentaire. Les propositions

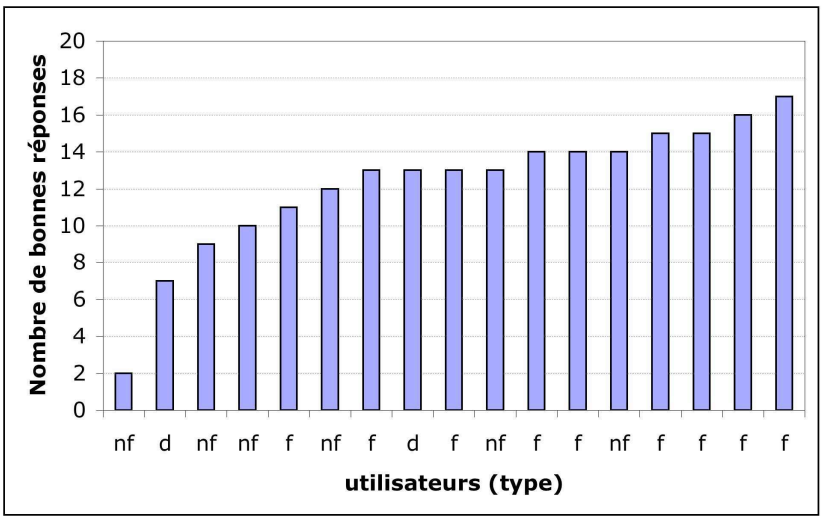


FIG. 4.9: *Nombre de bonnes réponses obtenues avec SQuaLIA pour chaque utilisateur : "f" correspond aux personnes dont le Français est la langue maternelle, "nf" aux personnes chinoises, allemandes ou espagnoles et "d" aux personnes francophones dyslexiques.*

et résultats présentés dans cette section ont été publiés dans un article de la revue TAL (Sitbon et al., 2008c), durant la conférence *Interspeech 2007* (Sitbon et al., 2007a) puis dans un atelier conjoint à la conférence SIGIR 2008 (Sitbon et Bellot, 2008a) ainsi que dans (Sitbon et al., 2007b).

4.6.1 Données recueillies

Le corpus que nous avons recueilli est constitué de questions tapées par des enfants dyslexiques qui sont également dysorthographiques. Il a été réalisé lors de séances d'orthophonie de huit enfants entre 9 ans et demi (classe de CE2) et 13 ans (classe de 4^e). Notons au passage que du fait de leur manque de familiarité avec les moteurs de recherche, les enfants formulent plus facilement que des adultes des questions en langue naturelle plutôt que des mots-clés isolés.

Le corpus ainsi obtenu, bien que de taille réduite (37 questions) permet beaucoup d'observations communes aux huit participants. En premier lieu, il apparaît que la plupart des observations faites par ailleurs sur les manuscrits d'enfants dyslexiques ne sont pas validées sur les écrits dactylographiés. Cela est dû non seulement à une organisation motrice différente pour la production écrite (il ne s'agit pas de former les lettres mais de les repérer sur le clavier où elles apparaissent en majuscules), mais également à une plus grande motivation impliquant une plus grande attention au niveau de la production comme de la relecture. Ainsi, on ne rencontre pas de substitutions de lettres dites *miroirs* (*p*, *b*, *d*, *q* ou *m* et *w*, *n* et *u*). De même, on n'observe que deux cas d'inversion de lettres, et aucun cas d'inversion de syllabes.

Les erreurs que l'on rencontre sont à la fois des erreurs de conversion graphèmes-

phonèmes et des erreurs de segmentation des mots. Cela correspond à une écriture partiellement phonétique qui n'est pas dénuée de *complexité orthographique*. Ainsi, *monnaie* s'écrit *monné, monais, moner, monnaie, moner, monaie* ou *monai*. Un problème central concerne la segmentation en mots : *s'appelle* peut s'écrire *ca ple* ou bien *sapel*, et *l'abbé Pierre* s'écrit *labe pierre, la Bepierre, labepier, labée pierre, l abepier, l'abée pierre* ou *labpier*. Apparaissent également des omissions ou substitutions de lettres dans des cas où les phonèmes ne sont pas assez distincts (comme pour *Bastia* ou *monnaie*). Une autre conséquence de l'écriture phonétique est la substitution de certains mots par des homophones (*mer* remplace *maire*).

Par ailleurs, on remarque des motifs d'erreurs constants pour chaque individu et propres à chacun. Par exemple, pour un enfant, les pronoms interrogatifs souffrent systématiquement d'un remplacement du *u* par une apostrophe (*q'elle* au lieu de *quel*) ou, pour un autre enfant, d'un ajout d'apostrophe (*qu'el* au lieu de *quel*). Cependant la définition de modèles individuels s'avère difficile car, en plus de nécessiter beaucoup d'exemples, elle devrait nécessairement être dynamique puisque les utilisateurs sont généralement en cours d'apprentissage et les erreurs type peuvent évoluer.

Un second corpus, composé de 46 questions plus variées (portant sur le nombre d'habitants au Liban ou en France, le lieu des jeux Olympiques une année donnée, le premier homme ayant marché sur la lune, la date de la chute du mur de Berlin, l'âge de décès du plus gros homme au monde...), a été recueilli afin de valider nos propositions. Il a été réalisé auprès de onze enfants d'une classe d'enseignement spécialisé pour des dyslexiques (CLasse d'Intégration Scolaire des Grands Cyprès à Avignon), entre 8 et 11 ans. Chaque enfant a tapé entre 3 et 6 phrases.

4.6.2 Correction par phonétisation et retranscription

Le problème de la correction orthographique adaptée à des personnes dysorthographiques dyslexiques a fait l'objet de travaux que nous mentionnons brièvement. L'hypothèse la plus communément admise est qu'indépendamment de l'origine des erreurs commises, ces dernières n'apparaissent pas de nature différente de celles générées par des normo-lecteurs même si elles sont plus fréquentes. Loosemore (1991); Deorowicz et Ciura (2005) ont ainsi proposé une modélisation à base de réseaux de neurones ou d'automates des erreurs rencontrées. Cette modélisation peut être spécifique à un utilisateur, comme proposé par Spooner (1998), mais les performances qu'il obtient semblent être d'un niveau seulement équivalent à celles d'un correcteur orthographique générique. De manière générale, il a été montré que ces performances sont faibles et les correcteurs standards non adaptés aux personnes dysorthographiques (James et Draffan, 2004). Des modèles phonétiques ont été introduits par Toutanova et Moore (2002), en se basant sur les approches probabilistes de canal bruité introduites par Brill et Moore (2000). La grande majorité des systèmes à l'exception par exemple de Pedler (2001), n'ont pas de composant sémantique qui permettrait de gérer une correction couplée à une désambiguisation et une prise de décision entre des homophones tels que *cygne, signe*.

Méthode

Dans le cadre d'un système de recherche d'informations devant *accepter* une requête dysorthographiée et partant du constat qu'une fois oralisées les questions peuvent être comprises, nous avons proposé une approche basée sur leur phonétisation conduisant non pas à *une* correction mais à *plusieurs* hypothèses de correction. Le système mis en œuvre parcours un graphe d'homophones en utilisant des composants logiciels développés au LIA et exploités habituellement pour la synthèse et la reconnaissance vocales. La figure 4.10 illustre la séquence de processus impliqués. Les systèmes de synthèse vocale permettent de transformer une séquence de lettres en une séquence de phonèmes. Pour cela, ils s'appuient sur un lexique de correspondances (le lexique phonétique) et sur un étiquetage morpho-syntaxique permettant de résoudre certains ambiguïtés ainsi que sur un ensemble de règles de conversion des graphèmes vers les phonèmes (voir par exemple Dutoit (1997); Yvon et al. (1998); Bellegarda (2005)). Les systèmes de transcription automatique utilisent quant à eux des modèles de langage associés aux lexiques phonétiques. À partir d'une séquence sonore, ils extraient les phonèmes correspondants sous forme de treillis de phonèmes réunissant toutes les hypothèses de reconnaissance pour chaque partie du signal. Ensuite, l'ensemble des sous-séquences de ce treillis ayant une correspondance dans le lexique phonétique permet de se ramener à un treillis d'hypothèses de mots. Enfin, des modèles de langage sont appliqués afin d'extraire la séquence de mots la plus probable.

Même si le cas de l'inversion de graphèmes s'avère moins fréquent qu'envisagé au départ, la prise en compte de ce type d'erreurs nécessiterait d'envisager tous les cas d'inversions possibles, chacun conduisant à une hypothèse, plus ou moins vraisemblable, de question. Afin de réduire le nombre de possibilités à explorer et de s'affranchir, au moins provisoirement, de ce problème, nous avons choisi d'exploiter le correcteur orthographique libre GNU ASPELL[15] qui utilise à la fois des distances d'édition et des distances phonologiques pour proposer des alternatives aux mots rencontrés hors de son lexique. Ce correcteur montre de bonnes performances par rapport aux autres outils commerciaux et libres grand public[16]. La figure 4.11 montre un exemple de graphe de mots à partir duquel sont produites des représentations phonétiques des questions. La phonétisation est effectuée à l'aide de l'outil LIA_phon (Béchet, 2001), qui dispose à la fois d'un lexique phonétique de 80.000 mots et d'un système de 1.996 règles de conversions ordonnées des plus générales aux plus spécifiques. La combinaison de ces deux ressources rend la phonétisation robuste, ce qui est essentiel compte tenu des dégradations orthographiques qui peuvent être rencontrées. La matrice de confusion pour obtenir le graphe de phonèmes étendu contient uniquement les confusions entre les voyelles ouvertes et fermées. La figure 4.12 donne un exemple d'hypothèses phonétiques obtenues à partir du graphe de mots de la figure 4.11. Ce graphe est enfin réduit aux chemins les moins coûteux selon un modèle de langage générique et un transducteur du lexique phonétique. Le tableau 4.1 indique les trois premières réécritures retenues pour la question "*kel aje a la Bepierre*" après phonétisation et re-transcription.

[15] http://aspell.sourceforge.net
[16] http://aspell.net/test/

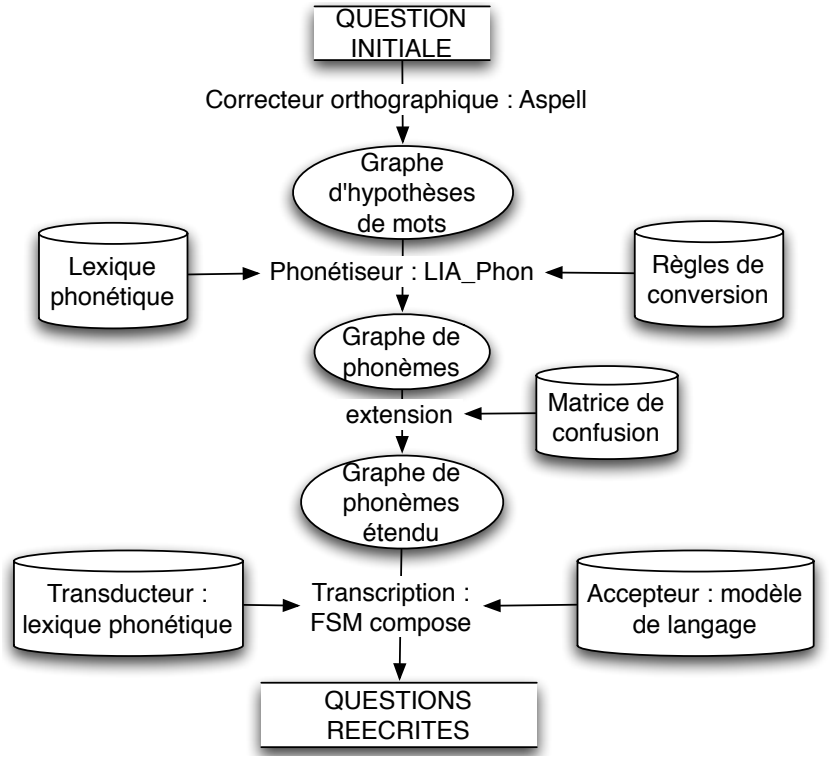


FIG. 4.10: *La procédure de correction orthographique conduisant à la génération d'hypothèses de questions utilise un correcteur orthographique standard, un phonétiseur (LIA_Phon) et un système de transcription automatique (figure reprise de Sitbon et al. (2008c)).*

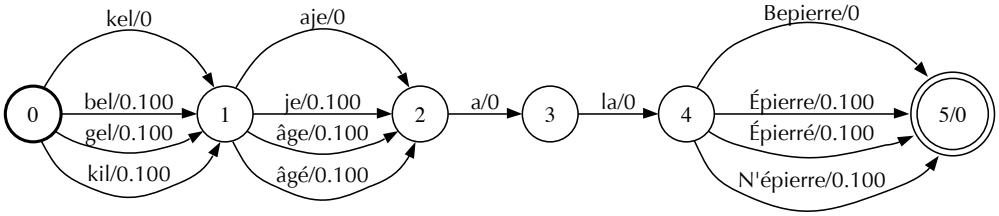


FIG. 4.11: *Exemple de graphe d'hypothèses de mots généré par le correcteur orthographique GNU Aspell à partir de la question : kel aje a la Bepierre (Quel âge a l'abbé Pierre ?) —figure reprise de Sitbon et al. (2008c).*

Evaluation

Nous avons effectué une transcription manuelle des phrases tapées par les enfants de manière à s'approcher au mieux de leur intention (il n'y a pas d'ambiguïtés dans les choix de transcription puisque l'on connaît par avance l'objet des questions). L'évaluation des hypothèses de réécritures obtenues a ensuite été réalisée grâce à la plate-

Phrase	Coût du chemin
quel âge a l' abbé pierre	45,79
quel âge à l' abbé pierres	46,5
quel âge alla et pierre	48,44

TAB. 4.1: *Les trois premières réécritures retenues pour la question "kel aje a la Bepierre" après phonétisation et re-transcription.*

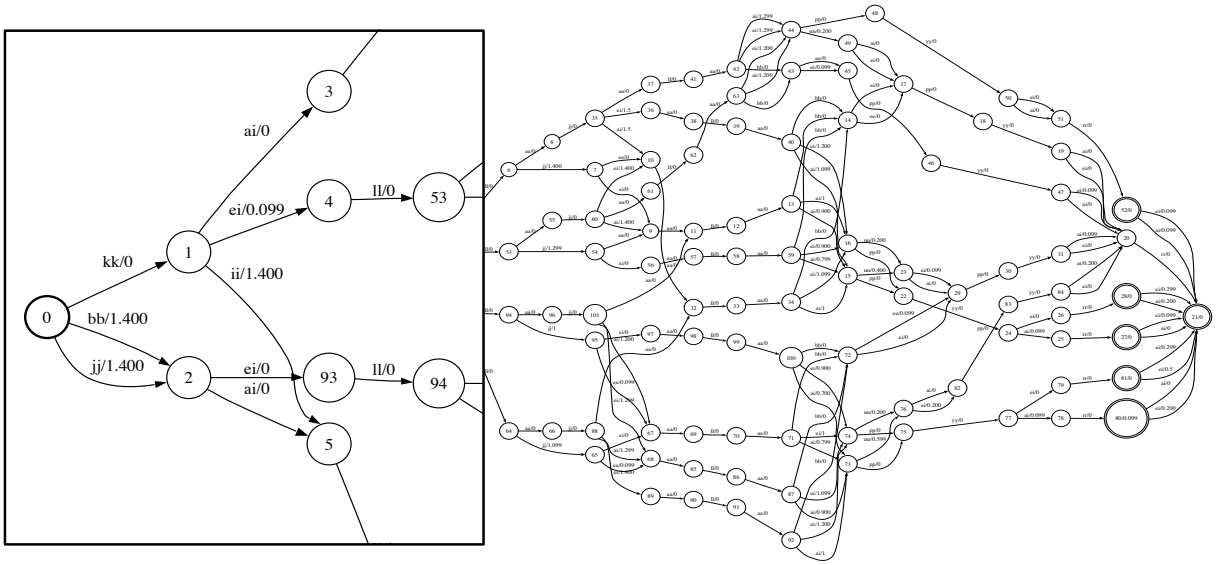


FIG. 4.12: *Exemple d'hypothèses phonétiques (phonème/poids, numéro de nœud) produites par LIA-Phon à partir des hypothèses graphémiques du correcteur Aspell et de la question : kel aje a la Bepierre (Quel âge a l'abbé Pierre ?) —figure reprise de Sitbon et al. (2008c).*

forme d'évaluation des outils de reconnaissance de la parole SCTK[17]. SCTK inclut l'outil SCLITE qui implémente un algorithme de programmation dynamique pour calculer des taux d'erreurs mots dans le meilleur des cas entre une phrase de référence et des hypothèses (représentées par un graphe de mots), en prenant en compte les insertions, omissions et les substitutions.

Le tableau 4.2 contient les résultats de l'évaluation de la qualité orthographique des questions d'origine (colonne *Initial*), de la première hypothèse graphémique issue du système Aspell (*Asp 1*) et de celles fournies selon notre méthode (*PhonTrans 1*), ainsi que de la réunion des trois premières hypothèses (*Asp 3* et *PhonTrans 3*). L'évaluation est effectuée selon deux critères : le taux de mots exacts et de questions correctes (pourcentage de réécritures totalement correctes). Les résultats pour les trois premières hypothèses montrent que si l'amélioration orthographique est déjà importante dans l'absolu (de 30 % rapport à l'initial), elle l'est aussi par rapport à un correcteur orthographique performant tel qu'Aspell (dont le taux est de 11 % plus bas).

[17] http://www.nist.gov/speech/tools

Mesure	Initial	Asp 1	Asp 3	PhonTrans 1	PhonTrans 3
Taux de lemmes corrects	51,6	67,1	70	78,9	**81,2**
Taux de phrases correctes	5,4	13,5	18,9	43,2	**45,9**

TAB. 4.2: *Taux de lemmes corrects et pourcentage de phrases identiques à la référence après lemmatisation et filtrage, sur les questions initiales ou réécrites à l'aide de Aspell (Asp) ou de notre système (PhonTrans), pour une ou trois hypothèses.*

Nous avons appliqué la même méthode sur le corpus de questions recueillies en école primaire. Le taux d'erreur sur ce second corpus sont supérieurs mais ils montrent toujours que les meilleures performances de correction sont obtenues avec notre système plutôt qu'avec Aspell. Cette hausse du taux d'erreur peut s'expliquer par la différence de niveau entre les individus ayant participé aux tests dans les deux cas : parmi les enfants de l'école primaire, figuraient deux enfants dysphasiques ayant du mal à s'exprimer oralement. Comme il fallait s'y attendre, cela suggère que les résultats varient selon les individus ; voir Sitbon et al. (2008c) pour une analyse détaillée et individualisée.

Au final, les performances obtenues permettent de proposer des réécritures qui multiplient très significativement le nombre de questions correctement orthographiées. L'évaluation des phrases filtrées et lemmatisées montre que l'on peut faire descendre le taux d'erreur de 51 % à 23 % en considérant uniquement la première hypothèse fournie, alors qu'un correcteur orthographique tel qu'Aspell ne permet de descendre qu'à 35,7 %. Une observation détaillée des résultats a montré que la cause des erreurs qui restent réside dans la surabondance d'insertion de lettres. De plus certaines confusions de phonèmes n'ont pas été prises en compte (l'utilisation de *promiet* pour *premier* apparaît plusieurs fois). Le point fort du formalisme des automates est de permettre des traitements supplémentaires en amont du traitement phonétique, afin d'augmenter le nombre d'hypothèses de réécritures. Cependant, cette multiplication risque de générer du bruit. Notons que les différentes hypothèses de réécriture peuvent constituer autant de candidats pour l'expansion de la requête posée au moteur de recherche. Cette potentialité est brièvement décrite dans (Sitbon et Bellot, 2008a) où les poids des mots de la requête étendue dérivent de leur impact dans le calcul du coût de chaque chemin issu du graphe d'hypothèses phonétiques. Une expérimentation sur l'effet de l'enrichissement sur la précision de la recherche d'informations reste à être conduite.

4.7 Lisibilité

Cette section est plus particulièrement consacrée à une manière de tenir compte des capacités de lecture en recherche d'informations. Pour un aperçu d'autres propositions dans l'adaptation des systèmes d'informations, se référer par exemple aux actes de l'atelier PeCUSI qui s'est tenu durant la conférence Inforsid 2007 et plus particulièrement de (Chevalier et al., 2007) qui discutent de la notion de profil utilisateur pour la Recherche d'Informations.

4.7.1 Les mesures de lisibilité de Flesch

Pour un utilisateur dyslexique, le temps nécessaire à l'obtention de l'information recherchée — le *coût* associé — est *a priori* plus important que pour un normo-lecteur rendant dès lors cruciale la prise en compte de notions non uniquement informationnelles dans le calcul de la pertinence d'un document. Comme cela a déjà été dit plus haut, la prise en compte de la *lisibilité* dans le processus de recherche s'adresse aussi à toute personne lisant dans une langue qui n'est pas sa langue maternelle ou qui n'est pas experte du domaine exploré.

Pour la langue anglaise, la formule générique la plus communément employée, y compris par un logiciel tel que Microsoft Word dans les outils statistiques de l'onglet *Grammaire et orthographe*, est la mesure définie par Flesch (1948) qui définit la lisibilité d'un document en fonction de la longueur moyenne de ses phrases ASL et du nombre moyen de syllabes par mot ASW (L_{Flesch}, formule 4.1). Pour la langue française, Kandel et Moles (1958) ont déterminé des valeurs *ad-hoc* des coefficients (L_{French}, formule 4.2) qui demeurent très proches de celles utilisées pour l'anglais. Les valeurs prises par ces deux mesures de lisibilité L_{Flesch} et L_{French} varient entre 0 et 100 : 30 correspondant à du texte très difficile à lire pour un normo-lecteur et 70 un texte de difficulté normale. Dans les expériences qui suivent, ce sont ces formules qui ont servi de mesure de référence. Les critères plus complets définis en début de chapitre devront y être intégrés. Notons que de nombreux autres facteurs, particulièrement sensibles pour les personnes mal voyantes, pourraient être pris en compte dans l'estimation de la lisibilité : formes, espacements et tailles des caractères, contrastes, alignements des paragraphes, présence d'images, structure physique des documents... Ces critères ont été définis dans les *Web Accessibility Initiative guidelines*[18] mais nous les avons écartés de notre étude, au moins provisoirement, pour nous concentrer sur des aspects plus linguistiques.

$$L_{Flesch}(d) = 206,835 - 1,015 \times ASL - 84,6 \times ASW \tag{4.1}$$

$$L_{French}(d) = 207 - 1,015 \times ASL - 73,6 \times ASW \tag{4.2}$$

4.7.2 Réordonnancement des documents trouvés selon la mesure de Flesch

Les mesures de Flesch (formules 4.1 et 4.2) produisent un score pour chaque document trouvé. Pour prendre en compte la lisibilité dans le processus de recherche, ces scores de lisibilité doivent être combinés avec les scores de *pertinence thématique* initiaux ou, mieux (Lee, 1997), en fonction des rangs initiaux des documents (voir discussion en section 4.8.1). À cet effet, une fonction de combinaison linéaire est définie ci-dessous où q est une requête, $Rank(d)$ le rang initial du document d, N le nombre de documents

[18] http://www.w3.org/WAI/

trouvés, L_F une des deux mesures de Flesch et enfin λ un coefficient tel que $0 \leqslant \lambda \leqslant 1$:

$$Sri(d) = (1 - \lambda) \cdot \left(1 - \frac{Rank(d)}{N}\right) + \lambda \cdot \frac{L_F(d)}{100} \tag{4.3}$$

Nous avons testé cette combinaison[19] sur les données de la tâche *ad-hoc* de la campagne TREC-8 qui comprennent 50 requêtes (*topics*) et un corpus de 530 000 documents. Les documents ont été initialement ordonnés en utilisant le moteur de recherche Lucene à partir des champs *title* des *topics*. Faisant l'hypothèse que la lisibilité d'un document est indépendante de la requête, il est naturel que sa prise en compte ne puisse pas, sauf dans des cas de pur hasard, entraîner une hausse de la précision : il n'y a pas de raison pour que les documents les plus lisibles soient les plus proches thématiquement de la requête. L'inverse étant également vrai (les documents les moins lisibles ne sont pas nécessairement les plus pertinents), nous nous attendions à ce que le réordonnancement des documents en fonction de la lisibilité reste neutre vis à vis de la mesure de précision. Lorsque la lisibilité est pondérée par un coefficient λ faible, le réordonnancement agit, vis à vis de la mesure de la précision, comme une redistribution locale et forcément limitée où l'espérance du gain en précision est quasi nul. Si la valeur λ est trop forte (à l'extrême, $\lambda = 1$, seule la lisibilité est prise en compte), le réordonnancement devient global et aboutit à redistribuer les documents pertinents dans un ensemble où ils sont minoritaires et donc à faire fortement chuter la précision. L'expérience confirme cette hypothèse, mais au delà de ces résultats, elle souligne une fois de plus le manque de pertinence... de la mesure de pertinence usuelle dans des conditions réelles : la prise en compte de la lisibilité entraînant de fait une hausse de la lisibilité des premiers documents trouvés (à condition bien sûr d'en accepter son bien fondé), il est probable que l'utilisateur puisse trouver plus rapidement l'information recherchée, et donc, en ce sens, que les premiers documents soient plus pertinents qu'auparavant. Cela ne peut être pris en compte par la mesure de précision seule et devra être validé par des expériences interactives en conditions réelles. Pour cela, se reporter à la méthodologie expérimentée par les évaluations de la piste HARD durant les campagnes TREC ainsi qu'au chapitre écrit par S. Chaudiron (2004b).

La figure 4.13 indique les taux de précision et de lisibilité moyenne obtenus sur les 10 et 20 premiers documents réordonnés en fonction de différentes valeurs du coefficient de pondération λ. Jusqu'à la valeur $\lambda = 0,2$, la précision ne diminue pas tandis que la lisibilité augmente. Au-delà de cette valeur, la chute de la précision est significative.

4.7.3 Proposition d'une mesure de lisibilité adaptée à la dyslexie

La dyslexie entraînant une capacité réduite à identifier les mots écrits, le nombre de mots *difficiles* est augmenté par rapport à celui d'un normo-lecteur. D'autre part, la

[19] De façon générale, l'utilisation d'une addition plutôt que d'une multiplication entre les deux scores normalisés permet de donner une plus grande importance au fait que les documents sont trouvés par plusieurs approches par rapport au score brut. N'ayant que deux scores utilisés ici, l'effet de ce choix devrait être insignifiant.

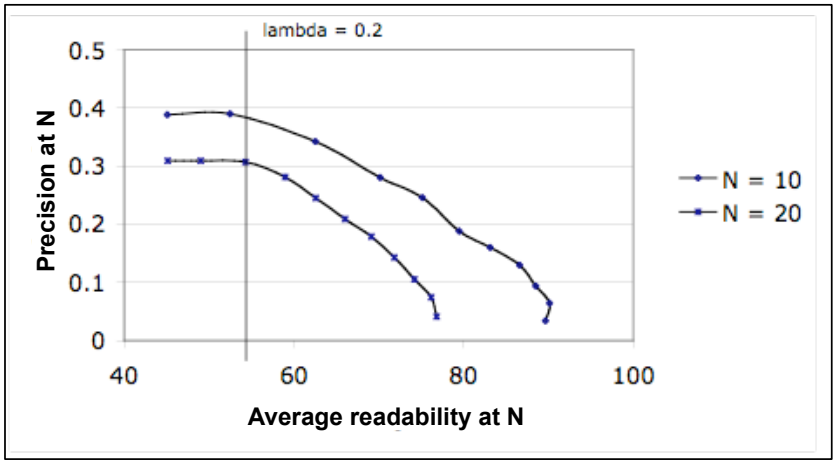


FIG. 4.13: *Précision et lisibilité moyenne pour les $N = 10$ ou $N = 20$ premiers documents trouvés après réordonnancement en fonction de différentes valeurs du coefficient de pondération λ (à gauche, $\lambda = 0$, et à droite, $\lambda = 1$). Sur les données de TREC-8, le meilleur compromis entre lisibilité et précision est obtenu pour la valeur $\lambda = 0,2$.*

variabilité et la complexité des types de dyslexie est telle qu'il est inadapté de prédire la lisibilité d'un texte à partir des seuls critères de longueur moyenne des phrases ou des mots comme cela a été défini plus haut. Cela nous a incité à utiliser d'autres critères tels que le nombre de lettres qui composent un mot, sa rareté dans le langage courant, sa catégorie morpho-syntaxique et sa cohésion graphème-phonème. Nous avons estimé cette dernière par le rapport entre le nombre de phonèmes et le nombre de lettres dans le mot[20]. Elle permet de tenir compte du fait qu'un mot contenant des lettres muettes ou bien des phonèmes de plusieurs lettres (*ph* vis à vis de *f* seul) est plus complexe à lire qu'un mot pour lequel la correspondance graphème-phonème est bijective[21]. À partir de la définition de la complexité d'un mot, celle d'une phrase peut être estimée en fonction de la moyenne des complexités des mots qu'elle contient.

Afin de déterminer les coefficients à appliquer à chacun des critères retenus pour estimer la difficulté d'un mot, nous avons choisi d'entraîner un classifieur à partir des temps de lecture d'un ensemble de phrases lues par des enfants. Ces données ont été recueillies par une équipe de psycholinguistes conduite par S. Ducrot, du Laboratoire Parole & Langage (LPL) du CNRS et de l'Université de Provence, dans le cadre d'expérimentations sur le diagnostic de la dyslexie par l'empan perceptif (Lété et Ducrot, 2007). Neuf enfants pour lesquels le français est la langue maternelle ont dû lire vingt phrases d'une longueur de douze mots[22]. L'expérimentation a été conduite par l'intermédiaire d'un logiciel réalisé par des étudiants en Master Informatique sous la supervi-

[20] Un niveau de consistance graphème-phonème est accessible pour les mots de la base de données lexicales Manulex-Infra constituée de mots issus de manuels scolaires en français (Peereman et al., 2007).

[21] Dans le même ordre d'idées, il serait judicieux de considérer le fait qu'une lettre seule, par exemple *c*, peut correspondre à différents phonèmes. Cela n'a pas été fait dans les expériences décrites ici, où la cohésion n'est donc qu'une première approximation.

[22] *Le chien de ma grand-mère aime beaucoup jouer avec mes chaussons* est un exemple de phrase à lire.

sion de L. Sitbon et moi-même. Les phrases ont été lues mot à mot (le passage d'un mot au suivant se faisant par activation d'une touche au clavier), ce qui a permis de mesurer des temps de lecture globaux et mot à mot. La lecture effective de chaque phrase a été validée par une épreuve visuelle de compréhension (l'enfant, après avoir lu chaque phrase, devait choisir l'image qui la représentait le mieux parmi deux).

Après avoir testé plusieurs classifieurs (voir Sitbon et al. (2008b)), selon une régression linéaire ou des machines à vecteurs supports, nous avons déterminé, pour chacun des critères retenus, les coefficients permettant d'estimer au mieux le temps de lecture et donc, par hypothèse, la difficulté de lecture. Après validation croisée, la meilleure estimation peut être trouvée selon la formule 4.4 où, comme l'on pouvait s'y attendre, le facteur principal est la cohésion graphème-phonème (ADV désigne le nombre d'adverbes, CON le nombre de conjonctions et COH la cohésion graphème-phonème).

$$Time(d) = 1.12 \times ADV - 0.69 \times CON + 6.48 \times COH + 15.58 \tag{4.4}$$

Cette définition permet de définir une nouvelle mesure de lisibilité L' considérant à la fois les difficultés spécifiques aux personnes dyslexiques (formule 4.4) et la lisibilité générique (formule 4.2) :

$$L'(d) = \frac{Time(d) + (100 - L_{French}(d))}{2} \tag{4.5}$$

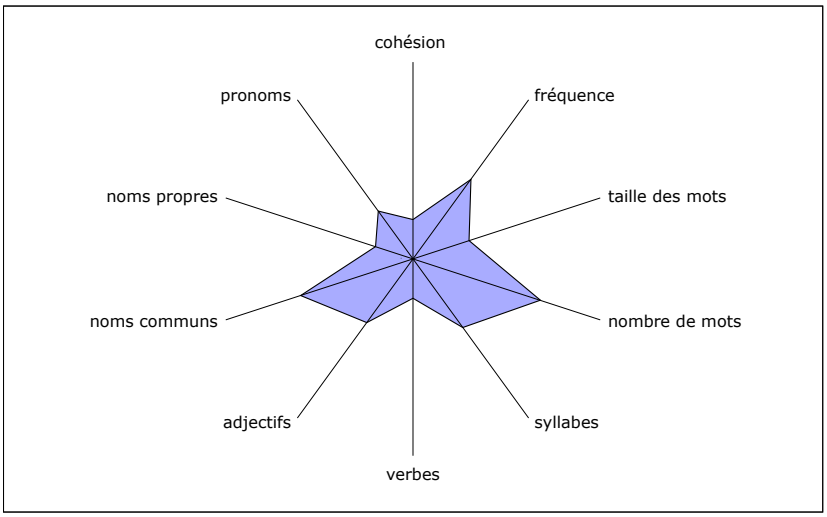


FIG. 4.14: *Coefficients d'impact des critères envisagés permettant d'estimer la difficulté de lecture d'un mot apprise. À partir d'expérimentations conduites avec 9 enfants dyslexiques, tous ces critères n'ont pas été retenus.*

Sur les données en français de la piste *ad-hoc* de la campagne CLEF, nous avons pu observer une gain de lisibilité de 20 % pour une perte relative de précision inférieure à 10 % sur les 10 ou 20 premiers documents (en prenant $\lambda = 0,3$). Si ces résultats sont encourageants, ils restent à être validés, et comparés avec une utilisation de la mesure

de Flesch seule, en conditions réelles d'utilisation. En l'état, ces résultats ont été présentés au *Second international workshop on Adaptive Information Retrieval* qui s'est tenu en conjonction avec la conférence IIiX 2008 (Sitbon et Bellot, 2008b).

4.8 Perspectives

4.8.1 Fonction de score d'un document combinant pertinence et lisibilité

Souhaiter combiner une mesure de lisibilité avec une mesure de similarité du type BM25 est une problématique qui s'apparente à celle de la fusion de résultats et à la méta-recherche où doivent être pris en compte différents scores pour un même document (Chidlovskii, 2003; Fox et Shaw, 1994; Lee, 1997; Montague et Aslam, 2001)[23]. À cet effet, la plupart des auteurs ont proposé d'établir de nouvelles fonctions de score, comme par exemple (Aslam et Montague, 2000, 2001; Ogilvie et Callan, 2003) à partir d'approches bayésiennes, (Lillis et al., 2006) selon une approche probabiliste entraînée sur les niveaux de performance pré-établis de plusieurs systèmes différents et établissant un score à partir des rangs obtenus. Manmatha et al. (2001) montrent que les scores obtenus par des moteurs de recherche durant les campagnes TREC suivent une distribution normale pour les documents pertinents et exponentielle pour les autres. À partir de cette observation, il est possible de déduire une probabilité de pertinence uniquement à partir de la valeur du score et de permettre alors la sélection automatique d'un système ou d'une méthode par rapport à une autre.

Une autre façon de voir est de faire le parallèle avec la recherche de documents structurés où le score retenu est une combinaison des scores des parties qui composent les documents (Wilkinson, 1994; Carmel et al., 2003; Piwowarski et Gallinari, 2003; Lalmas et Tombros, 2007). Pour un document d structuré en K parties d_k, La forme courante de la fonction de score est :

$$s(d,q) = \sum_k v_k \cdot s'(d_k,q) \tag{4.6}$$

avec s la fonction de score sur le document, s' la fonction de score sur une partie de d, v_k le poids de la k^e partie de d tel que $\sum v_k = 1$ et q la requête.

Selon une démarche opposée à celle qui consiste à modifier la fonction de score, Robertson et al. (2004) ont proposé d'agir directement sur les valeurs des indices permettant d'estimer la pondération BM25. Cela permet non seulement une reformulation simple du problème de la prise en compte d'indices multiples mais également de conserver toutes les propriétés mathématiques du modèle probabiliste de type *Okapi* : non linéarité vis à vis de la composante tf[24], et, pour les documents structurés, conser-

[23] La sélection des seuls documents lisibles ou d'un niveau d'expertise adéquat pourrait aussi s'apparenter à un problème de filtrage (Berrut et Denos, 2003; Boughanem et al., 2004b).

[24] La désormais classique pondération BM25 (en Annexes 8.0.6, p. 151) est basée sur un modèle de distribution des occurrences des mots de type 2-poisson qui part du principe que, pour un document donné, les mots peuvent être séparés en deux classes : la classe des mots *élites* qui sont caractéristiques d'un document et la classe des autres mots. De manière toujours aussi classique, les mots sont pondérés

vation des valeurs idf et de l'impact de la longueur du document sur le score final.

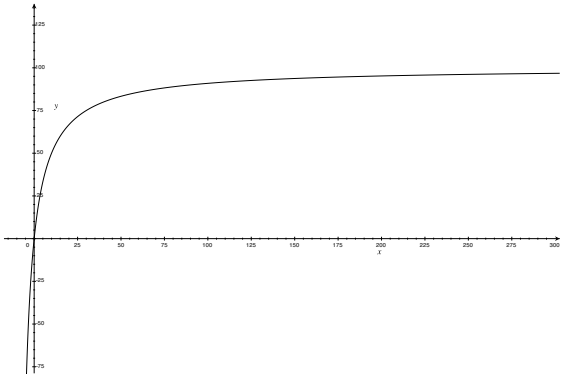


FIG. 4.15: *Pondération BM25 : évolution non linéaire du poids d'un mot dans un document (axe des ordonnées) en fonction de la valeur de la composante $tf \geqslant 0$) (axe des abscisses).*

Dans la section précédente, nous avons défini une combinaison linéaire des scores mais cela peut ne pas être la meilleure solution. D'abord car il a été montré que les différentes stratégies de classement internes peuvent interférer avec un tel choix et que d'autre solutions sont souvent préférables (Savoy et al., 1997). Ensuite, parce que la redondance des indices pris en compte dans chacun des scores peut entraîner une sélection déséquilibrée. Les expériences que nous avons faites sont basées sur une fonction de score définie selon la formule 4.5, elle-même liée aux formules 4.4 et 4.2 (p. 117 et 120). Ces définitions tiennent compte d'indices qui ne sont pas liés à la fréquence d'apparition des mots dans les documents et la combinaison linéaire du score de lisibilité avec une fonction de score BM25 ne fait pas perdre la non linéarité vis à vis de la composante tf.

Cependant, si l'on considère l'ensemble des critères de lisibilité envisagés en début de chapitre (et non uniquement ceux que nous avons utilisés dans nos expériences), et donc la fréquence d'apparition des mots (rappelons qu'un mot est plus facilement lu s'il est rencontré souvent), cette propriété de non linéarité serait mise à mal. En effet, un mot de la requête qui apparaîtrait plusieurs fois dans un document verrait cette quantité prise en compte deux fois dans le calcul du score, une fois pour la lisibilité et une fois pour la similarité BM25. Ainsi, il est possible qu'un document contenant plusieurs fois un même mot de la requête, et uniquement celui-ci, soit préféré à un document qui contient deux mots de la requête pour des valeurs idf identiques. Pour répondre à ce problème précis, Robertson et al. (2004) ont choisi de modifier la valeur de la composante tf plutôt que de combiner des scores qui, indépendamment les uns des autres, utilisent la valeur tf initiale. Si l'on estime par exemple qu'un mot qui apparaît 3 fois dans un document est 2 fois plus lisible qu'un mot qui n'apparaît qu'une seule

en tenant compte de leur nombre d'occurrences au sein du document considéré (tf) et dans la collection entière (idf). Il a été montré que l'évolution du poids en fonction du nombre d'occurrences devait être croissant mais non linéaire pour une meilleure efficacité. C'est le cas des pondérations *Okapi*/BM25 qui ont une composante tf de la forme $\frac{k \cdot tf}{\beta \cdot k + tf}$ (figure 4.15).

fois alors on modifie son tf en le multipliant par deux. Pour un document structuré d, cela correspond à définir :

$$tf(w,d) = \sum_k v_k \cdot tf(w,d_k) \tag{4.7}$$

avec w un mot, d_k une partie de d et v_k le poids de la k^e partie de d tel que $\sum v_k = 1$.

Relevons que les références données en début de section concernent toutes la combinaison de score de nature identique : tous les moteurs de recherche calculent leurs scores selon un modèle de pertinence. Dans notre cas, la combinaison doit agir sur des scores *a priori* orthogonaux, rendant la nature du problème quelque peu différente. Cela n'est d'ailleurs pas sans rappeler la question du mélange des modalités texte/audio que nous avons vue au chapitre précédent concernant la recherche multimédia.

Une perspective intéressante réside donc dans la définition de fonctions de score qui préservent l'efficacité des mesures de type *okapi* tout en tenant compte de la lisibilité. Les arguments que nous venons d'avancer en ce sens sont à pondérer par le fait que les mesures de similarité de type okapi/BM25 ne font intervenir que les mots communs à la requête et aux documents, soit un nombre de mots très réduit face à la diversité du vocabulaire des documents trouvés. Même dans le cas d'un utilisateur dyslexique, on peut estimer que les mots de la requête seront facilement lus dans les documents, ne serait-ce que parce que l'utilisateur les a justement utilisés dans sa requête. En ce sens, ils sont autant de mots que l'on pourrait peut être écarter de l'estimation de la lisibilité des documents trouvés. Toutefois, dans le cadre d'un processus d'expansion de requête, la stratégie évoquée ci-dessus retrouve toute sa pertinence puisque des mots non présents dans la requête sont alors utilisés pour ordonner les documents. Il y aurait alors intérêt à essayer d'étendre les requêtes avec des mots *lisibles*.

4.8.2 Un processus spécifique d'expansion par retour de pertinence... et de lisibilité

À la suite des expériences relatées dans les sections précédentes, nous devons procéder à une évaluation en conditions réelles d'un moteur de recherche prenant en compte la lisibilité des documents trouvés dans sa fonction d'ordonnancement. Cette évaluation permettrait non seulement d'estimer l'apport de la méthode non supervisée que nous avons proposée mais aussi de procéder à l'apprentissage de nouvelles fonctions d'ordonnancement en fonction des retours des utilisateurs. Les critères de jugement combinerait pertinence thématique, facilité de lecture et rapidité pour obtenir l'information recherchée.

Nous pourrons alors passer d'un processus qui enchaîne :

1. pondération des mots des documents et de la requête selon une approche type *okapi*/BM25 ;
2. calcul des scores de pertinence vis à vis de la requête des documents de la collection ;

3. calcul des scores de lisibilité des documents extraits de la collection à l'étape précédente ;
4. réordonnancement des documents en fonction des scores de pertinence et de lisibilité.

vers un nouveau processus incluant une rétro-action (ou une pseudo-rétroaction) de pertinence améliorant simultanément la lisibilité et la pertinence des documents trouvés.

La lisibilité est une notion qui dépend de l'utilisateur, de ses capacités de lecture et de son niveau d'expertise vis à vis des thématiques des documents lus. Comme nous l'avons vu dans le cadre de la dyslexie ou d'un enfant apprenant à lire, la difficulté de lecture d'un texte peut être estimée en fonction de caractéristiques directement quantifiables à partir du texte lui-même. Il serait malgré tout plus efficace de disposer de modèles adaptés à chaque utilisateur. Dans l'idéal, ces profils seraient appris automatiquement et représenteraient les capacités de lecture de l'utilisateur dans un cadre générique mais aussi en fonction des thématiques rencontrées.

Selon un processus interactif, il serait possible d'associer à des requêtes, et par suite à des thématiques, des listes de documents que l'utilisateur aura trouvés non seulement pertinents mais également *utilisables* (lisibles). De manière classique, un modèle de langue pourrait convenir à une telle définition.

De façon alternative, nous proposons d'exploiter les acquis des sciences cognitives en modélisation de la lecture pour associer à chaque lecteur un modèle connexionniste représentatif de ses capacités en lecture. Il a été montré que de tels modèles pouvaient simuler un grand nombre des difficultés rencontrées chez des normo-lecteurs ou chez des dyslexiques, notamment en faisant varier le nombre d'unités cachées entre le système lexical et le système phonologique d'un réseau connexionniste computationnel. Autrement dit, il est envisageable d'estimer un paramètre explicite directement lié à la capacité de lecture d'un utilisateur. Après un apprentissage supervisé, ce modèle pourrait être appliqué aux documents trouvés par un moteur de recherche, et les sorties phonologiques comparées dynamiquement aux prononciations correctes (elles-mêmes calculées selon un modèle fiable de lecteur expert). Une autre possibilité, plus complexe à réaliser mais certainement plus efficace, consisterait à mettre en place un processus entièrement interactif de l'évaluation de la lisibilité d'un document en fonction des temps et de la qualité de lecture des documents trouvés observés dynamiquement. Une lecture à voix haute permettrait d'estimer à chaque instant t non seulement la position dans le texte mais aussi l'exactitude de ce qui est lu. On en déduirait une fonction de mise à jour des modèles selon deux paramètres : la vitesse et le taux d'erreur. À un instant $t = 0$, l'erreur serait maximale mais elle serait faiblement prise en compte dans la mise à jour. Au fur et mesure que le temps passe, les erreurs diminuent (les hésitations de lecture conduisant finalement à une reconnaissance correcte des mots ne sont pas pénalisantes) mais leur niveau de prise en compte augmente. Les méthodes d'apprentissage par renforcement constituent un cadre théorique adapté à ce type de problème.

L'apprentissage par renforcement (par l'expérience) permet de décider de la sélec-

tion de documents par un processus adaptatif séquentiel permettant une mise à jour du modèle en fonction des erreurs (en l'occurrence la pertinence des documents et leur lisibilité). Contrairement à un apprentissage supervisé, on ne cherche pas dans ce cas à apprendre une fonction de correspondance requête / documents à partir d'exemples (estimation d'une mesure de similarité), mais on adapte le comportement du système suivant les retours de l'utilisateur de façon séquentielle. Une fonction d'évaluation (estimation d'une probabilité de pertinence/lisibilité) est mise à jour de manière à ce que le système puisse préjuger de la qualité de sa réponse[25]. Le temps de lecture de l'utilisateur agirait comme un signal de retour sur la qualité de la sortie du système. Un temps de lecture court indique soit une grande facilité de lecture soit une non pertinence du document qui n'aurait pas été lu en entier. Il est possible d'estimer ce temps de lecture en fonction de la quantité de texte lu en pistant les mouvements de la souris et des ascenseurs de la fenêtre où le texte est visualisé.

Un procédé expérimental plus évolué encore est utilisé depuis une trentaine d'années pour mesurer l'attention d'un lecteur et sa capacité de lecture. Il enregistre les mouvements oculaires d'un sujet en train de lire et *"a permis, par exemple, de mettre en lumière une stratégie globale de déplacement du regard vers une position optimale, en fonction d'indices uniquement visuels, et des tactiques locales de réajustement en fonction d'interprétations plus élaborées de données visuelles et linguistiques. C'est la théorie dite "stratégie et tactiques" proposée par le groupe Regard de l'Université Paris V (O'Regan, 1990)"* (Caelen et al., 2003) *in* (Gaussier et Stéfanini, 2003). Les mouvements oculaires permettent alors de détecter les zones que l'utilisateur trouve particulièrement attractives et pertinentes (Eglin et Emptoz, 1997). Pour un utilisateur ayant des difficultés de lecture, un ralentissement local peut signifier qu'il rencontre une zone ayant un score de lisibilité faible. Caelen et al. (2003) proposent d'exploiter cette technique expérimentale afin de déterminer automatiquement la pertinence d'un document vis à vis d'une requête. À notre connaissance, cela n'a cependant pas été évalué.

[25] Dans un séminaire donné en 2007 au Collège de France, Rémi Munos illustrait la différence entre l'apprentissage d'une fonction décrivant les données (entrées/sorties) et celui d'une fonction de correspondance en faisant le parallèle avec un enfant qui apprend à faire du vélo : il ne connaît pas les équations qui décrivent les mouvements et ne les connaîtra jamais, mais il apprend les mouvements à faire et à ne pas faire pour réussir dans sa tâche.

Chapitre 5

Perspectives générales

[...] la pensée précède le langage.

(G. Edelman, 2007)

Dans un article récent sur les enjeux de la recherche d'informations, Belkin (2008) rappelle que la question de la personnalisation est annoncée comme majeure depuis une vingtaine d'années mais que la plupart des tentatives vers une étude systématique ont échoué. S'il n'est pas question de prétendre que les évaluations *TREC Interactive* (Hersh et Over, 2001) puis *TREC Hard* (Allan, 2005) et *TREC ciQA* (Kelly et Lin, 2007), pour ne mentionner qu'elles, n'ont abouti à rien, force est de constater qu'aussi bien des paradigmes d'évaluation que des stratégies efficaces restent encore à trouver et que l'utilisateur n'intervient que très peu dans les systèmes de recherche de l'Internet. Belkin (2008) cite Karen Sparck-Jones durant la conférence SIGIR 1988 (Sparck-Jones, 1988) :

> *As it is, it is impossible not to feel that continuing research on probabilistic weighting in the style in which it has been conducted, however good in itself in aims and conduct, is just bombinating in the void...*
>
> *The current interest [...] is in integrated, personalisable information management systems.*

Dans les différents chapitres qui constituent ce mémoire, nous avons présenté trois directions vers une personnalisation de la recherche d'informations. La première (chapitre 1) correspond à une analyse du besoin en information d'un utilisateur qui permet de distinguer recherche documentaire et recherche de réponses précises sachant que celles-ci peuvent être des informations factuelles, des définitions ou des explications. Cette distinction correspond à celle entre Recherche d'informations (RI) et questions-réponses (QR) mais, pour être utilisables, les interfaces des systèmes correspondants devront être fusionnées (il est peu probable qu'un utilisateur soit satisfait d'un système questions-réponses qui ne fournit pas de document support à ses réponses). En outre, l'utilisateur devra changer ses habitudes pour abandonner les requêtes mots-clés pour (re)venir à des requêtes en langue naturelle.

La seconde direction que nous avons suivie (chapitre 3) est celle de l'aide à la navi-

gation dans de grandes bases documentaires audio et textuelles. Elle consiste à définir une interface graphique permettant un survol chronologique, par l'exploitation de méthodes d'indexation sémantique, de segmentation thématique (chapitre 2) et de résumé automatique, des documents de la collection.

La troisième direction (chapitre 4) est, quant à elle, véritablement une personnalisation de la recherche d'informations puisqu'elle correspond à prendre en compte la capacité de lecture d'un utilisateur dans le calcul du score de pertinence d'un document vis à vis d'une requête.

La problématique de la personnalisation et de la prise en compte de l'utilisateur en recherche d'informations renvoie naturellement à celle, bien plus large, des fondements du traitement automatique des langues, au croisement de la linguistique et de l'informatique, toutes deux rejointes par la psychologie pour l'étude des comportements individuels, les neurosciences pour l'étude du cerveau et des racines physiologiques du langage mais aussi par la sociologie et la sémiologie pour des analyses globales des besoins, des attitudes et des significations. Ce croisement pluridisciplinaire est un enjeu majeur des années à venir si l'on veut aller au-delà, pour reprendre K. Sparck-Jones, de la seule étude permettant d'espérer (et encore ne s'agit-il que d'un espoir sans même être convaincu de la significativité des gains) grappiller quelques points de précision en recherche *ad-hoc*. La place de l'informaticien dans ce puzzle au dessein encore flou mérite d'être discutée même si, pour moi, la réponse est à peu près claire.

L'intelligence artificielle, du moins à ses débuts, est partie de l'hypothèse que le traitement humain de l'information était avant tout symbolique et organisé selon des structures logiques fonctionnant à la manière d'un automate. Cela a conduit à des réalisations qui ont pu faire croire que l'ordinateur allait très vite être capable de parler, de dialoguer, de traduire, voire de *comprendre*, comme un humain. Malheureusement les prototypes n'ont jamais pu passer le cap de la mise à l'échelle et n'ont su faire preuve d'une robustesse suffisante pour pouvoir être pleinement exploités. Le Prix Nobel de médecine G. Edelman condamne cette approche logique censée reproduire le fonctionnement du cerveau (Edelman, 2007) :

> [...] *les cerveaux humains fonctionnent fondamentalement en terme de reconnaissance de structures plutôt que de logique. Ils sont hautement constructifs pour établir des structures et en même temps constamment portés à l'erreur. On le voit dans les illusions de perception ainsi que dans les croyances de niveau supérieur. Mais comme le montre l'analyse de l'apprentissage, ils peuvent en général procéder à des corrections d'erreurs en réponse à des récompenses et des punitions adaptées.* [...] *Le langage lui-même reflète l'aspect constructif et pourtant foncièrement ambigu et indéterminé de ce mode de pensée.*

Parallèlement aux approches symboliques, les méthodes numériques, fondées sur une étude statistique étudiée en corpus, ont prouvé leur grande capacité à s'adapter rapidement à diverses thématiques ou langues. Cela s'est fait au prix d'une certaine *approximation* dans les résultats, toujours exacts dans les approches symboliques pour peu que les *règles* soient *justes* et *complètes*.

Ces deux orientations, symbolique et numérique, ont longtemps été opposées, chacune des communautés scientifiques se distinguant par des finalités divergentes opposant prototypes limités mais exacts pour l'une et systèmes fonctionnels mais approxi-

matifs pour l'autre et par des attitudes plus ou moins pragmatiques. Cette opposition rejoint en quelque sorte celle qui a opposé, et oppose toujours, certains linguistes et philosophes sur la nature même du langage et son acquisition à savoir, en simplifiant à l'extrême, la pré-existence ou non d'un système (cognitif) de règles génératrices des phrases possibles au sein d'une langue (ou tout du moins du *degré* de pré-existence). C'est ainsi que F. de Saussure parlait au début du 20^{e} siècle de *faculté de langage comme instinct naturel* inhérent à tout être humain ou que, dans un autre style, Nietzsche écrivait : "[...] *La « raison » dans le langage : ah ! quelle vieille femme trompeuse ! Je crains bien que nous ne nous débarrassions jamais de Dieu, puisque nous croyons encore à la grammaire*" (Nietzsche, 1888)[1]. L'existence de *lois premières* est évoquée par M. Foucault dans l'ouvrage "*Les mots et les choses*" dont nous reprenons ici un court extrait (Foucault, 1966) :

> *Si le mot peut figurer dans un discours où il veut dire quelque chose, ce ne sera pas par la vertu d'une discursivité immédiate qu'il détiendrait en propre et par droit de naissance, mais parce que dans sa forme même, dans les sonorités qui le composent, dans les changements qu'il subit selon la fonction grammaticale qu'il occupe, dans les modifications enfin auxquelles il se trouve soumis à travers le temps, il obéit à un certain nombre de lois strictes qui régissent de façon semblable tous les autres éléments de la même langue ; si bien que le mot n'est plus attaché à une représentation que dans la mesure où il fait partie d'abord de l'organisation grammaticale par laquelle la langue définit et assure sa cohérence propre. Pour que le mot puisse dire ce qu'il dit, il faut qu'il appartienne à une totalité grammaticale qui, par rapport à lui, est première, fondamentale et déterminante.*

Ce débat sur la nature et le rôle de la grammaire, a ses origines au 17^{e} siècle entre tenants de l'empirisme (l'être humain est né vierge et est entièrement façonné par l'expérience) et du rationnalisme (l'homme ne peut être réduit à son expérience). Dans les années 1950, le courant behavioriste, empiriste, a tenté de définir l'acquisition du langage comme un apprentissage sous forme de réactions en chaîne par rapport à des renforcements positifs ou négatifs. En opposition, N. Chomsky a proposé la pré-existence de structures cognitives spécifiques au langage et propres à l'être humain, suggérant ainsi que le langage est quelque chose de réellement biologique (Grodzinsky, 2007).

Une conséquence directe de la pré-existence de structures grammaticales pour l'acquisition du langage a d'abord consisté à définir à la fois une *grammaire universelle* exprimant les universaux linguistiques et *des grammaires particulières* pour les spécificités des multiples langues (Chomsky, 1964) : "*Dans cette conception, l'acquisition de la langue était vue comme un processus d'induction de règles : l'enfant, pourvu de la structure générale de la grammaire universelle définissant une certaine classe de grammaires particulières possibles, doit découvrir les règles particulières qui engendrent la langue particulière à laquelle il est exposé*" (Rizzi, 2007). Ultérieurement, les grammaires particulières ont été réduites à des valeurs spécifiques de paramètres de la grammaire universelle, l'acquisition d'une langue correspondant ainsi en la fixation de ces paramètres (Chomsky, 1981). Il n'y a alors plus besoin d'induire de quelconques règles (Rizzi, 2007), l'induction étant remplacée par un processus de détermination et de sélection parmi toutes les productions linguistes *a priori* possibles (Mehler et Dupoux, 1992).

Selon cette description de l'acquisition du langage, la notion d'*apprentissage* est bien sûr de première importance. Cet apprentissage peut être vu sous deux angles qui re-

[1] *in* "Le crépuscule des idoles", chapitre "La « raison » dans la philosophie".

joignent la problématique déjà évoquée plus haut, à savoir apprentissage statistique ou "analytique" (tous deux naturellement inconscients, et, probablement, combinés). Dans le premier cas, il s'agit pour l'enfant, d'observer lesquelles de ses productions linguistiques aboutissent au but recherché et, de manière évolutive, accumuler une sorte de comptabilité de ce qui réussit et de ce qui échoue pour aboutir à une sélection de *possibles*. Ce type d'apprentissage peut être modélisé par des réseaux de neurones faisant intervenir différentes couches plus ou moins explicites reliant le lexique, des concepts et des sons ; l'intention (le but recherché) étant alors un chemin particulier au sein du réseau. Le succès ou l'échec se traduisent par un renforcement ou un affaiblissement des connexions. Cependant, aucune simulation n'est encore parvenue à simuler le langage humain dans son étendue sémantique et comportementale. La question de la convergence de l'apprentissage se pose aussi bien que celle de la réduction de la combinatoire à des parcours possibles. On se reportera tout de même avec le plus grand intérêt au modèle Caramel et à ses implémentations (Sabbah, 1996; Sabbah et Popescu-Belis, 1999). Dans le second cas, l'apprentissage consiste en un raffinement progressif de la valeur des paramètres de la grammaire universelle autorisant la production, et la compréhension, des énoncés qu'elle peut *générer*.

Remarquons que, dans le premier cas, il est toujours possible de relever, une fois le système stabilisé, des régularités structurelles, nécessaires ne serait-ce que pour assurer une homogénéité dans le temps et une compréhension mutuelle entre deux interlocuteurs, desquelles une grammaire peut être déduite. Concernant le second point de vue, des interrogations sur sa pertinence sont soulevées qui concernent à la fois l'ensemble des productions s'écartant de *la* grammaire (dans la lignée de la linguistique de corpus) mais pourtant *compréhensibles* (au sens où un enfant peut tout à coup construire incorrectement une phrase qui aurait pourtant dû respecter une forme syntaxique qu'il maîtrise par ailleurs)[2] et les conséquences linguistiques propres à certains troubles du langage (dysphasie développementale) difficiles à modéliser[3]. Ici encore, les interprétations sont multiples mais l'on pourra retenir celle exprimée par Jakubowicz (2007) :

> [...] *je propose que les différences observées par rapport au système adulte, transitoires chez l'enfant sain, de durée plus longue ou indéfinie chez l'enfant atteint d'un trouble spécifique du langage, ne relèvent en fait pas de l'engin syntaxique lui-même mais d'opérations post-syntaxiques par lesquelles les représentations construites par cet engin sont matérialisées.*

Les découvertes récentes de l'imagerie médicale fonctionnelle[4] ne permettent pas à

[2] Si les modèles numériques sont par nature *robustes*, ça n'est pas le cas de la plupart des modèles symboliques. Face à l'incapacité des grammaires chomskyennes à considérer les énoncés qu'elles ne peuvent générer (et qui sont donc en ce sens en dehors du langage), d'autres grammaires ont été proposées telles que les grammaires de propriétés (Blache, 2000, 2003; Guénot, 2006). Elles fournissent "[...] *une vision de la langue comme un tout, dans laquelle chaque domaine* [phonologie, syntaxe, sémantique...] *possède son propre système de description (ou grammaire), les interactions étant contrôlées par un niveau supérieur.* [...] *La possibilité de rendre compte de la langue en situation et dans une perspective communicationnelle tout en se situant dans une approche formelle confère ainsi pleinement un aspect cognitif à la théorie.*

[3] D'autres interrogations, plus philosophiques, trouvent au moins en partie leur réponse dans (Pollock, 2007) qui ajoute que l'étude neurophysiologique des activités cérébrales ne peut suffire à elle seule pour étudier le langage sous peine de se priver d'une cinquantaine d'années de travaux aux résultats "*formellement précis*".

[4] Dans le domaine de l'imagerie anatomique, certains auteurs vont jusqu'à spécifier des aires corticales

elles seules de décider de la validité de cette hypothèse. Si l'on parvient à déterminer au moins approximativement les zones du cerveau propres à la prononciation mentale (aire de Broca comme étape pré-articulatoire), la détection de phonèmes (partie antérieure de l'aire de Wernicke) ou encore les aires visuelles distinctes pour la reconnaissance des visages, la représentation mentale des lieux ou des mots écrits (différentes zones qui sont en jeu dans la lecture) (Dehaene, 2007), il est plus délicat de déterminer une localisation précise d'un quelconque *engin syntaxique*. Ce pas est en partie franchi par certains chercheurs qui relèvent une forte activation de l'aire de Broca dans le traitement de mouvement syntaxique — mouvement de l'objet au cours du traitement des relatives objet par rapport aux phrases complétives — (Grodzinsky, 2007).

Finalement, les propositions et modèles qui découlent des travaux de N. Chomsky ne paraissent pas incompatibles avec une vision behavioriste relativisée (voir par exemple Harth (1993) cité par Sabbah, 2006a). Au lieu d'opposer modélisations neuronales et grammaire générative avec structures innées, il est parfaitement possible de postuler l'existence de réseaux neuronaux ayant une structure initiale propice à l'acquisition du langage, à la fois en respect avec une grammaire universelle et en accord avec la nécessité d'interconnexion avec les aires cognitives non spécifiques au langage : le réseau forme bien un *tout*. Cette approche est alors en partie (mais en partie seulement) conforme au modèle d'une conscience humaine définie comme une *dynamique neuronale* (Edelman, 2007) dans laquelle l'apprentissage par renforcement (processus de sélection) est contraint par le code génétique tout en étant sensible à l'expérience. G. Edelman s'oppose toutefois fermement aux propositions chomskyennes sans les nommer explicitement (p. 77) :

> *Nous formons la seule espèce dotée d'un langage fondé sur une syntaxe. De nombreux chercheurs ont suggéré que le langage est un trait dû à l'évolution biologique ; certains ont même proposé de penser que nous possédons un dispositif spécifique d'acquisition du langage ; nous en aurions hérité et il nous permettrait d'effectuer et de reconnaître les déclarations correctes quant à la syntaxe. La théorie de la sélection des groupes de neurones rejette cette conception.*

Toutefois, cette opposition ne paraît pas définitive puisqu'à peine plus loin (p. 78), il ajoute :

> *L'interaction des ganglions de la base des aires motrices, sensorielles et préfrontales du cortex a pu donner lieu à une capacité généralisée de détecter les séquences sensorimotrices, formant ainsi une sorte de "syntaxe de base". Le cas échéant, un langage vrai fondé sur une syntaxe est apparu, invention s'appuyant sur des capacités déjà évoluées.* [...] *notre compréhension de la façon dont le cerveau rend possible le langage en est au stade de l'enfance.*

J.P. Changeux (2002) va dans ce sens lorsqu'il écrit :

> *Le développement du langage chez l'enfant se démarque de la conception naïve de l'acquisition des mots comme une simple mise en connexion un par un des signifiants et des signifiés, mais s'accorde, en revanche, avec le schéma de la communication inférentielle : il s'agit de la dynamique de développement du processus d'acquisition du sens des mots.*

spécifiques à certains verbes d'action en liaison avec les zones neuronales du système moteur (Rizzolatti et Sinigaglia, 2008). Ainsi certaines neurones discriminent l'information sensorielle suivant les possibilités d'action qu'elle offre (formes, dimensions...). L'activation de neurones miroirs potentiellement associés à des actes moteurs est liée à des actions (*intentions*) et non à des mouvements particuliers du corps.

> *[Elle] ne suit pas une croissance progressive du nombre de mots puis de syntagmes [...] puis de groupes de mots. [c'est le contraire qui se produit] [...] On peut concevoir qu'une restriction et une spécification progressive des relations entre son et sens [...] s'effectuent par un processus de sélection par récompense partagée. Tout récemment, L. Rizzi [a montré que] l'enfant explore un grand nombre de règles syntaxiques possibles, qui sont éphémères [...] Ensuite il omet ou oublie, c'est à dire élimine, les constructions qui ne sont pas en accord avec les connaissances grammaticales courantes qu'il tire de son environnement social immédiat.*

En tout état de cause, les modèles neuronaux (et les modèles numériques en général) apportent un plus essentiel qui est la possibilité d'associer des valeurs continues aux objets manipulés, qu'ils soient conceptuels ou non. Ainsi, plutôt que d'opposer de manière irréductible le modèle aristotélicen qui va permettre de catégoriser des objets en fonction de leurs propriétés communes et le modèle qui procède au classement selon des ressemblances avec des objets "prototypes", on préférera associer, à tout *objet*, une valeur d'appartenance à chaque catégorie. Ainsi un verre, défini comme étant un objet dans lequel on peut mettre un liquide, sera toujours un verre, même s'il est cassé ; mais il le sera *un peu moins*. Nous suivons en ce sens les positions avancées par Sabbah (2006a) qui préconise de laisser de côté la quête plus ou moins vaine de définir le sens d'un texte dans l'absolu (sauf pour des cas particuliers de domaines spécifiques ou normalisés comme pour le Web Sémantique) pour se pencher vers une sémantique subjective :

> *Ignorer l'existence de sensations néglige le fait que l'incarnation est productrice de sens et qu'il est impossible d'avoir des états intentionnels sans expérience subjective. Une sémantique subjective centrée sur l'individu, serait ainsi plus utile pour modéliser de façon plus analogique les processus de compréhension. On ne chercherait pas à représenter le sens comme un état du monde de référence, mais comme une modification d'un état de connaissance, c'est à dire comme un effet sur le contexte cognitif du système.* [...] *Il semble donc fondamental d'examiner de plus près les liens entre les significations et les perceptions.* [...] *pour une confrontation continuelle entre les énoncés reçus et les connaissances antérieures, stockés dans une mémoire non seulement associative, mais aussi prospective et réflexive. Ainsi l'intelligence artificielle purement symbolique semble-t-elle prendre le problème à l'envers.* [...]

Nous n'avons pas la prétention de trancher entre une approche ou une autre quant au choix de celle qui correspond le mieux au fonctionnement du cerveau humain. Cependant nous croyons que l'un des principaux acquis du traitement automatique des langues de ces dernières années et qu'il est nécessaire, au moins en l'état des connaissances et des capacités computationnelles des machines, de ne pas se cantonner soit dans une approche entièrement numérique soit (j'oserai dire *encore moins*) dans une approche uniquement symbolique.

Ainsi, nous proposons d'être attentifs simultanément aux travaux issus des neurosciences et de la linguistique. À ce titre, nous serons particulièrement sensibles à la robustesse des modèles face aux phénomènes linguistiques rencontrés, qu'ils soient *corrects* ou non. Les méthodes d'apprentissage automatique basées sur le renforcement (Munos, 1997) et la réentrance, par opposition à un apprentissage supervisé ou à des méthodes de prise de décision markoviennes où un plan d'action doit être décidé en chaque état, seront approfondies et adaptées en profitant de leur apparente adéquation

supérieure avec les processus cognitifs qui régissent notre cerveau. Le travail débuté avec la thèse de Laurianne Sitbon (chapitre 4), la capacité des modèles numériques neuronaux à rendre compte de certaines déficiences cognitives et l'expérience acquise avec les moteurs de questions-réponses (chapitre 1) sont autant d'arguments en faveur de cette voie pluridisciplinaire centrée sur *l'individu* que nous souhaitons continuer à suivre.

Troisième partie

Annexes

Chapitre 6

Aide à l'apprentissage de la lecture

Dans le cadre de nos travaux autour de la dyslexie, deux groupes d'étudiants en Master Informatique ont développé une plateforme logicielle d'aide à la création d'exercices dédiés à la conscience phonologique, utile en phase d'apprentissage de la lecture. Ce travail a été réalisé en collaboration avec une *CLasse d'Intégration Scolaire* (CLIS) d'enfants dyslexiques.

Les figures suivantes présentent quelques unes des interfaces graphiques permettant recherches phonétiques et variations graphiques au sein de lexiques adaptés aux enfants. Cette plateforme fera l'objet d'une communication au colloque Majecstic 2008 (Rubino et Lavalley, 2008).

FIG. 6.1: *Interface web du logiciel d'aide à la création d'exercices éducatifs : module de recherche phonétique au sein du lexique Manulex-Infra*[1] *(Peereman et al., 2007). Ce logiciel a été développé par R. Rubino et R. Lavalley dans le cadre de leur projet de Master. La recherche peut se faire en combinant ou excluant phonèmes et graphèmes correspondants, en fonction de la catégorie morpho-syntaxique des mots ou encore suivant le niveau de difficulté défini en classes de niveaux suivant qu'ils sont rencontrés ou non dans les livres scolaires de CP, CE1 ou CE2.*

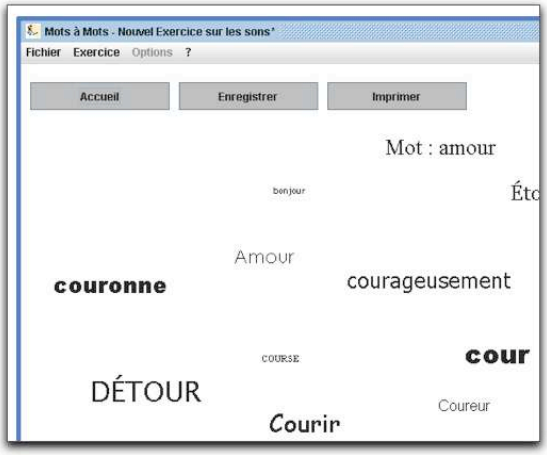


FIG. 6.2: *Exemple d'exercice de lecture créé avec le logiciel Mot à Mot : différents mots, automatiquement extraits de la base Manulex et contenant chacun les phonèmes "ou-r" sont affichés selon une typographie variée. L'élève doit identifier le mot de référence, ici amour. Ce logiciel a été développé par G. Cheylac et S. Masurel dans le cadre de leur projet de Master.*

IPA	codes	Exemples	IPA	codes	Exemples
Voyelles			Consonnes		
i	i	lire, vie	p	p	loupe, pain
u	u	joue, ours	t	t	terre, vite
y	y	bulle, sud	k	k	qui, bec
e	e	fée, nez	b	b	cube, brosse
ɛ	E	jouet, aile	d	d	danse, aide
a	a	date, plat	g	g	gare, bague
ɑ	A	tâche, bois	f	f	foule, phare
ø	2	deux, peu	s	s	tasse, cerf
œ	9	neuf, fleuve	ʃ	S	chat, vache
ə	*	le, ancre	v	v	vent, rêve
ɔ	o	roche, sol	z	z	zéro, rose
o	O	jaune, mot	ʒ	Z	gel, juge
ɔ̃	§	nom, pont	m	m	main, femme
ɛ̃	5	cinq, plein	n	n	nage, laine
ɑ̃	@	vent, blanc	ɲ	N	ligne, peigne
œ̃	1	un, brun	l	l	lune, pull
Semi-Voyelles			ʁ	R	rue, air
j	j	feuille, lieu			
w	w	soie, watt	ŋ	G	viking, ring
ɥ	8	huit, fruit			

FIG. 6.3: *Liste des phonèmes du français avec leurs caractères IPA et leurs codes dans Manulex-Infra. Tableau repris de la documentation de Manulex-Infra, Bases de Données Descriptives pour l'Etude de la Lecture et de l'Ecriture chez les Enfants de l'Ecole Elémentaire (http://leadserv.u-bourgogne.fr/bases/manulex/manulex_infra/indexFR.htm) (Peereman et al., 2007)*

Grapheme	Exemple	Grapheme	Exemple	Grapheme	Exemple
a	affiche	er	amer	on	ronde
à	çà	eu	deux	o	bol
â	âge	eû	jeûner	ô	hôpital
aen	caen	ey	poneys	ooin	shampooing
ai	aigle	ez	parlez	oo	alcool
aî	aîné	f	fond	ou	joue
aim	daim	ff	affable	où	où
ain	train	ge	bougeoir	oû	août
am	tambour	g	gare	ow	clown
an	ange	gg	agglomération	oy	troyes
aon	faon	gn	agneau	ph	dauphin
aô	saône	gu	aiguiser	p	pont
au	épaule	h	hiver	pp	grippe
aw	crawl	i	rire	q	cinq
ay	crayon	î	île	qu	quille
b	balcon	ï	héroïque	r	roue
bb	abbé	il	ail	rr	nourrice
cch	pinocchio	illi	serpillière	sch	schéma
cc	accord	ill	accastillage	sc	crescendo
ch	chaise	im	timbre	sç	acquiesça
c	car	in	vin	sh	washington
ck	hockey	în	devînt	ss	mousse
cqu	acquéreur	ïn	coïncidence	s	rose
ç	hameçon	j	jour	th	panthère
d	date	k	kilo	t	vite
dd	bouddha	kk	drakkar	tt	roulotte
ea	leader	le	scrabble	ue	bruegel
ean	jean	l	loup	um	humble
eau	agneau	ll	actuelle	un	aucun
e	adverbe	m	meuble	ü	capharnaüm
é	abbé	mm	pomme	u	rue
è	mystère	ng	boeing	û	bûche
ê	tête	n	nom	v	vent
ë	noël	nn	abonné	wh	whisky
ee	jeep	oa	goal	w	kiwi
ei	veine	oeu	bœuf	x	boxe
eim	reims	oe	moelle	y	yoga
ein	plein	oê	poêle	ym	cymbale
em	assembler	oin	loin	yn	larynx
emm	dilemme	oi	oiseau	z	lézard
en	dent	oî	benoît	zz	pizza
enn	solennel	om	ombre		

FIG. 6.4: *Liste des 125 graphèmes du français. Tableau repris de la documentation de Manulex-Infra, Bases de Données Descriptives pour l'Etude de la Lecture et de l'Ecriture chez les Enfants de l'Ecole Elémentaire (*`http://leadserv.u-bourgogne.fr/bases/manulex/manulex_infra/indexFR.htm`*) (Peereman et al., 2007)*

Chapitre 7

Recherche d'information suivant le modèle vectoriel

Le modèle vectoriel représente les documents du corpus et les requêtes par des vecteurs de mots clés. Ces mots clés sont eux-mêmes extraits des textes lors de la phase d'indexation. Pour chaque document, un poids est attribué à chacun des mots clés qu'il contient. Dans le modèle vectoriel, le vecteur requête est représenté dans le même espace que les vecteurs documents. Le vecteur requête peut alors être comparé à chacun des vecteurs documents. Cette comparaison correspond au calcul d'une similarité (ou distance) entre les vecteurs documents et le vecteur requête. Les différentes valeurs de similarité permettent d'ordonner les documents trouvés.

Si deux documents partagent le même nombre de mots communs avec la requête, seule une pondération non binaire des mots permet de différencier (et donc d'ordonner) ces documents. D'une manière générale, la pondération permet d'établir, aussi bien dans les documents que dans la requête, un ordre d'importance entre les mots sur lequel se base le calcul de similarité. L'étude des pondérations est donc particulièrement importante et constitue une partie majeure des travaux en recherche documentaire. Un document d (resp. une requête q) est représenté par un vecteur de la forme :

$$\vec{d} = \begin{pmatrix} w_{m_1,d} \\ w_{m_2,d} \\ \vdots \\ w_{m_n,d} \end{pmatrix}$$

avec m_i le i-ème mot de la liste de tous les mots retenus et $w_{i,d}$ le poids de m_i dans le document d (resp. q). Dans le cas où les m_i n'appartient pas au document d (resp. à la requête q) : $w_{i,d} = 0$.

Les mots sont généralement supposés comme étant mutuellement indépendants. Ceci est une forte simplification puisqu'il est clair que la possibilité de présence ou d'absence d'un mot dans un texte dépend des autres mots qui constituent ce texte.

7.0.3 Mesures de similarité

Les similarités sont calculées en tenant compte des poids des mots dans les documents et dans la requête. Lorsque aucune technique de regroupement de mots ou d'enrichissement n'est utilisée, la valeur de similarité tient uniquement compte des mots communs à la requête q et au document d. Parmi les nombreuses mesures de similarité possibles, le produit scalaire est couramment utilisé :

$$s(\vec{d}, \vec{q}) = \sum_{i=1}^{i=n} w_{m_i,d} \cdot w_{m_i,q} \tag{7.1}$$

Ainsi, un document qui n'a pas un seul mot commun avec la requête ne peut pas être trouvé par le système de recherche alors qu'il peut malgré tout être un document pertinent ou bien, dans le cas de questions-réponses, contenir la réponse cherchée.

7.0.4 Pondération des entrées de l'index : les critères tf et idf

Lorsque les poids w valent systématiquement 1 pour les mots contenus dans d ou q, la mesure de similarité (7.1) revient à dénombrer le nombre de mots communs à la requête et au document. Une pondération plus fine permet de distinguer, dans un document et dans une requête, les mots importants de ceux qui le sont moins. Dans une optique de recherche documentaire, un mot important est un mot discriminant qui permet de fortement caractériser un document vis à vis des requêtes pouvant être posées et vis à vis de l'ensemble des documents de la collection.

Globalement, favoriser les mots qui apparaissent souvent dans les documents revient à augmenter le nombre de documents trouvés et donc à favoriser le rappel. Inversement, favoriser les mots qui apparaissent rarement, revient à sélectionner des documents spécifiques et a donc tendance à améliorer la précision (pour les définitions du rappel et de la précision, se reporter à la section 10, page 157).

Si les mots les plus importants ne sont pas ceux qui apparaissent dans un grand nombre de documents, ce ne sont pas non plus seulement ceux qui apparaissent dans un faible nombre de documents (adopter cette solution pour seul critère reviendrait à privilégier fortement les fautes de frappe ou les orthographes rares – transcriptions de noms propres non courantes par exemple –). Il est raisonnable de penser que plus un mot apparaît dans un document, plus le sens de ce mot influe sur la thématique du document. Ceci est essentiellement valable pour certaines catégories de mots (noms propres, substantifs…). On peut alors entrevoir la possibilité de pondérer différemment les mots en fonction de leur type : en fait, cette distinction est implicite pour la plupart des pondérations du fait de fréquences d'apparition différentes des catégories morpho-syntaxiques.

Finalement, pour un document donné, un mot important est un mot qui est fréquent dans ce document mais qui ne se retrouve que dans un faible nombre de documents

dans le corpus. Cette définition de l'importance d'un mot correspond à introduire les facteurs tf (*term frequency*) et idf (*inverse document frequency*) exprimant respectivement le nombre d'occurrences d'un mot donné dans un document et le nombre de documents qui contiennent ce mot dans le corpus. Ces critères sont généralement combinés (Salton, 1975) selon une expression du type :

$$w_i = tf_i \cdot idf_i \tag{7.2}$$

Normalisation des poids. Si la pondération w tient compte du critère tf, une mesure de similarité telle que le produit scalaire (7.1) a tendance à favoriser les documents longs par rapport aux documents courts puisque les mots ont plus de chances d'y apparaître souvent. La normalisation des poids en fonction de la taille des vecteurs permet de minimiser l'avantage des documents longs. Parmi les deux normalisations les plus employées, citons :

$$w_{i,d} = \frac{w_{i,d}}{\sum_{j=1}^{n} w_{j,d}} \tag{7.3}$$

et la norme euclidienne :

$$w_{i,d} = \frac{w_{i,d}}{\sqrt{\sum_{j=1}^{n} w_{j,d}^2}} \tag{7.4}$$

Lorsque la norme euclidienne est choisie pour normaliser les poids, le calcul de la similarité (7.1) se ramène à celui du cosinus :

$$s(\vec{d},\vec{q}) = \sum_{i=1}^{i=n} \frac{w_{i,d}}{\sqrt{\sum_{j=1}^{n} w_{j,d}^2}} \cdot \frac{w_{i,q}}{\sqrt{\sum_{j=1}^{n} w_{j,q}^2}} = \frac{\vec{d}\cdot\vec{q}}{\|\vec{d}\|_2 \cdot \|\vec{q}\|_2} = \cos(\vec{d},\vec{q}) \tag{7.5}$$

Notons toutefois que comme cela a été montré dans (Singhal et al., 1996), le cosinus a tendance, dans la pratique, à privilégier les documents courts par rapport aux documents longs. Afin de mieux prendre en compte le critère «taille des documents», plusieurs autres mesures de pondération ont vu le jour à l'occasion des évaluations TREC. Les deux plus importantes sont celles proposées dans le système Okapi (voir la formule 8.37 décrite page 151) et dans (Singhal et al., 1996), donnée comme suit :

$$w_{i,d} = \frac{1 + \frac{\log tf_{i,d}}{1+pivot}}{(1-slope)\cdot pivot + slope \cdot dl_d} \tag{7.6}$$

avec $slope$ et $pivot$ des constantes fixées expérimentalement et dl_dla longueur du document d en nombre de mots.

Pondérations différentes pour les requêtes et les documents. Parmi les très nombreuses pondérations proposées dans la littérature, l'une des plus performantes en moyenne est la suivante qui distingue les poids dans les documents d et dans la requête q :

$$w_{i,d} = \frac{tf_{i,d} \cdot \log \frac{N}{n(m_i)}}{\sqrt{\sum_{m_j \in d} \left(tf_{i,d} \cdot \log \frac{N}{n(m_j)} \right)^2}} \tag{7.7}$$

et, pour la requête :

$$w_{i,q} = \left(0,5 + 0,5 \frac{tf_{i,q}}{\max_{m_j \in d} tf_{j,q}} \right) \cdot \log \frac{N}{n(m_i)} \tag{7.8}$$

Chapitre 8

Recherche d'information suivant le modèle probabiliste

Le modèle probabiliste (Robertson et Sparck-Jones, 1976) permet de représenter le processus de recherche documentaire comme un processus de décision : le *coût*, pour l'utilisateur, associé à la récupération d'un document doit être minimisé. Autrement dit, un document n'est proposé à l'utilisateur que si le coût associé à cette proposition est inférieur à celui de ne pas le retrouver :

$$EC_{retr}(d) < EC_{\bar{retr}}(d) \tag{8.1}$$

avec :

$$EC_{retr}(d) = P(\text{pert.}|\vec{d})C_{\text{retrouvé,pert.}} + P(\overline{\text{pert.}}|\vec{d})C_{\text{retrouvé,}\overline{\text{pert.}}} \tag{8.2}$$

où $P(\text{pert.}|\vec{d})$ désigne la probabilité qu'un document d est pertinent sachant ses caractéristiques $\vec{d}$, $P(\overline{\text{pertinent}}|\vec{d})$ qu'il ne le soit pas et $C_{\text{retrouvé,pert.}}$ le coût associé au fait de retrouver (ramener) un document pertinent et $C_{\text{retrouvé,}\bar{\text{pert.}}}$ de retrouver un document non pertinent.

La règle de décision devient alors : retrouver un document s seulement si :

$$P(\text{pert.}|\vec{d})C_{\text{retr.,pert.}} + P(\overline{\text{pert.}}|\vec{d})C_{\text{retr.,}\overline{\text{pert.}}} < P(\text{pert.}|\vec{d})C_{\overline{\text{retr.}},\text{pert.}} + P(\overline{\text{pert.}}|\vec{d})C_{\overline{\text{retr.}},\overline{\text{pert.}}} \tag{8.3}$$

soit :

$$\frac{P(\text{pert.}|\vec{d})}{P(\bar{\text{pert.}}|\vec{d})} > \frac{C_{\text{retrouvé,}\overline{\text{pert.}}} - C_{\overline{\text{retrouvé}},\overline{\text{pert.}}}}{C_{\overline{\text{retrouvé}},\text{pertinent}} - C_{\text{retrouvé,pert.}}} = \text{constante} = \lambda \tag{8.4}$$

La valeur de la constante λ dépend du type de recherche effectuée : désire-t-on privilégier le rappel ou la précision *etc.*

Une autre manière de voir le modèle probabiliste est de considérer que celui-ci cherche à modéliser l'ensemble des documents pertinents, autrement dit à estimer la probabilité qu'un mot donné apparaisse dans de tels documents.

8.0.5 Le modèle de pertinence binaire

Soit q une requête et d_j un document. Le modèle probabiliste essaye d'estimer la probabilité que l'utilisateur trouve intéressant le document d_j sachant la requête q. On suppose qu'il existe alors l'ensemble R des documents intéressants (on parle d'ensemble *idéal*) et que ces documents désignent l'ensemble des documents *pertinents*. Soit $\overline{R}$ le complément de R. Le modèle attribue à chaque document d_j sa probabilité de pertinence de la façon suivante :

$$d_j \leftrightarrow \frac{P(d_j \text{ est pertinent})}{P(d_j \text{ n'est } \textit{pas} \text{ pertinent})} \tag{8.5}$$

$$sim(d_j, q) = \frac{P(R|\vec{d}_j)}{P(\overline{R}|\vec{d}_j)} \tag{8.6}$$

Ainsi, si la probabilité que d_j soit pertinent est grande mais que la probabilité qu'il ne le soit pas est grande également, la similarité $sim(d_j, q)$ sera faible. Cette quantité ne pouvant être calculée qu'à la condition de savoir définir la *pertinence* d'un document en fonction de q (ce que l'on ne sait faire), il est nécessaire de la déterminer à partir d'exemples de documents pertinents.
Selon la règle de Bayes : $P(R|\vec{d}_j) = \dfrac{P(R) \cdot P(\vec{d}_j|R)}{P(\vec{d}_j)}$, la similarité est égale à :

$$sim(d_j, q) = \frac{P(\vec{d}_j|R) \times P(R)}{P(\vec{d}_j|\overline{R}) \times P(\overline{R})} \sim \frac{P(\vec{d}_j|R)}{P(\vec{d}_j|\overline{R})} \tag{8.7}$$

$P(\vec{d}_j|R)$ correspond à la probabilité de sélectionner aléatoirement d_j dans l'ensemble des documents pertinents et $P(R)$ la probabilité qu'un document choisi aléatoirement dans la collection est pertinent. $P(R)$ et $P(\overline{R})$ sont indépendants de q, leur calcul n'est donc pas nécessaire pour ordonner les $sim(d_j, q)$.
Il est alors possible de définir un seuil λ en-deça duquel les documents ne sont plus considérés pertinents.

En faisant l'hypothèse que les mots apparaissent indépendamment les uns des autres dans les textes (hypothèse naturellement fausse... mais réaliste à l'usage !), les probabilités se réduisent à celles des sacs de mots.

$$P(\vec{d}_j|R) = \prod_{i=1}^{i=n} P(d_{j,i})|R) = \prod_{i=1}^{i=n} P(w_{m_i,d_j})|R) \tag{8.8}$$

$$P(\vec{d}_j|\overline{R}) = \prod_{i=1}^{i=n} P(d_{j,i})|\overline{R} = \prod_{i=1}^{i=n} P(w_{m_i,d_j})|\overline{R})) \tag{8.9}$$

Dans le modèle probabiliste, les poids des entrées m_i de l'index sont binaires :

$$w_{m_i,d_j} = \{0,1\} \tag{8.10}$$

La probabilité de sélectionner aléatoirement d_j dans l'ensemble des documents pertinents est égal au produit des probabilités d'appartenance des mots de d_j dans un document de R (choisi aléatoirement) et des probabilités de non appartenance à un document de R (choisi aléatoirement) des mots non présents dans d_j :

$$sim(d_j,q) \sim \frac{\left(\prod_{m_i \in d_j} P(m_i|R)\right) \times \left(\prod_{m_i \notin d_j} P(\bar{m}_i|R)\right)}{\left(\prod_{m_i \in d_j} P(m_i|\overline{R})\right) \times \left(\prod_{m_i \notin d_j} P(\bar{m}_i|\overline{R})\right)} \tag{8.11}$$

avec $P(m_i|R)$ la probabilité que le mot m_i soit présent dans un document sélectionné aléatoirement dans R et $P(\bar{m}_i|R)$ la probabilité que le mot m_i ne soit *pas* présent dans un document sélectionné aléatoirement dans R.

Cette équation peut être coupée en deux parties suivant que le mot appartient ou non au document d_j :

$$sim(d_j,q) \sim \prod_{m_i \in d_j} \frac{P(m_i|R)}{P(m_i|\overline{R})} \times \prod_{m_i \notin d_j} \frac{P(\bar{m}_i|R)}{P(\bar{m}_i|\overline{R})} \tag{8.12}$$

Soit $p_i = P(m_i \in d_j|R)$ la probabilité que le i^e mot de d_j apparaisse dans un document pertinent et soit $q_i = P(m_i \in d_j|\overline{R})$ la probabilité que le i^e mot de d_j apparaisse dans un document *non* pertinent. Il est clair que $1 - p_i = P(m_i \notin d_j|R)$ et $1 - q_i = P(m_i \notin d_j|\overline{R})$. Il est enfin généralement supposé que, pour les mots n'apparaissant pas dans la requête : $p_i = q_i$ (Fuhr, 1992). Dans ces conditions :

$$\begin{aligned}
sim(d_j,q) &\sim \prod_{m_i \in d_j} \frac{p_i}{q_i} \times \prod_{m_i \notin d_j} \frac{1-p_i}{1-q_i} \\
&\sim \prod_{m_i \in d_j \cap q} \frac{p_i}{q_i} \times \prod_{m_i \in d_j, m_i \notin q} \frac{p_i}{q_i} \times \prod_{m_i \notin d_j, m_i \in q} \frac{1-p_i}{1-q_i} \times \prod_{m_i \notin d_j, m_i \notin q} \frac{1-p_i}{1-q_i} \\
&\sim \prod_{m_i \in d_j \cap q} \frac{p_i}{q_i} \times \prod_{m_i \notin d_j, m_i \in q} \frac{1-p_i}{1-q_i} \\
&= \prod_{m_i \in d_j \cap q} \frac{p_i}{q_i} \times \frac{\prod_{m_i \in q} \frac{1-p_i}{1-q_i}}{\prod_{m_i \in d_j \cap q} \frac{1-p_i}{1-q_i}} \\
&= \prod_{m_i \in d_j \cap q} \frac{p_i(1-q_i)}{q_i(1-p_i)} \times \prod_{m_i \in q} \frac{1-p_i}{1-q_i}
\end{aligned} \tag{8.13}$$

Le deuxième terme de ce produit est indépendant du document (tous les mots de la requête sont pris en compte, indépendamment de d_j). Ce qui nous intéresse étant uniquement d'ordonner les documents, ce terme peut être ignoré.

Soit, en passant au logarithme[1] :

$$sim(d_j, q) \sim \sum_{m_i \in d_j \cap q} \log \frac{p_i(1-q_i)}{q_i(1-p_i)} = RSV(d_j, q) \tag{8.15}$$

$sim(d_j, q)$ est souvent dénommée le *RSV* (*Retrieval Status Value*) de d_j pour la requête q.

En gardant les notations précédentes :

$$sim(d_j, q) \sim \sum_{m_i \in q \cap d_j} \left(\log \frac{P(m_i|R)}{1 - P(m_i|R)} + \log \frac{P(m_i|\overline{R})}{1 - P(m_i|\overline{R})} \right) \tag{8.16}$$

Si l'on intègre le nombre d'occurrences $f(m_i, d_j)$ des m_i dans d_j, on obtient :

$$sim(d_j, q) \sim \sum_{m_i \in d_j \cap q} f(m_i, d_j) \cdot \log \frac{p_i(1-q_i)}{q_i(1-p_i)} \tag{8.17}$$

8.0.6 Estimation des paramètres

Lors de la première itération, aucun document pertinent n'a encore été trouvé, il est nécessaire de poser les valeurs de $P(m_i|R)$ et de $P(m_i|\overline{R})$. On suppose ainsi qu'il y a une chance sur deux qu'un mot quelconque de l'index soit présent dans un document pertinent et que la probabilité qu'un mot soit présent dans un document non pertinent est proportionnelle à sa distribution dans la collection (étant donné que le nombre de documents non pertinents est généralement bien plus grand que celui des pertinents) :

$$P(m_i|R) = 0,5 \tag{8.18}$$

$$P(m_i|\overline{R}) = \frac{n_i}{N} \tag{8.19}$$

avec n_i le nombre de documents qui contiennent m_i dans la collection et N le nombre total de documents de la collection. Ces valeurs doivent être estimées lors de chaque itération en fonction des documents qu'elles permettent de trouver (et, éventuellement de la sélection de ceux qui sont pertinents par l'utilisateur).

À partir de ces valeurs initiales, il est possible de calculer $sim(d_j, q)$ pour tous les documents de la collection et de ne retenir que ceux dont la similarité est supérieure à λ. Le choix de λ peut se ramener au choix d'un *rang* r au-delà duquel les documents

[1] D'autres démonstrations font intervenir le calcul des probabilités selon une distribution binaire. Une telle distribution (également dite de Bernouilli), décrit la probabilité d'un événement binaire (le mot appartient ou n'appartient pas) en fonction de la valeur de la variable et de la probabilité de cette valeur :

$$\beta(x; p) = p^x (1-p)^{1-x} \tag{8.14}$$

qui donne la probabilité que x vaut 1 ou 0 en fonction de p. Le paramètre p peut être interprété comme la probabilité que x vaut 1 ou comme le pourcentage de fois où $x = 1$.

sont écartés. Soit V_i le nombre des documents dans le sous-ensemble des documents retenus qui contiennent m_i (V désigne alors le nombre de documents retenus). $P(m_i|R)$ et de $P(m_i|\overline{R})$ sont alors calculées récursivement :

$$P(m_i|R) = \frac{V_i}{V} \tag{8.20}$$

$$P(m_i|\overline{R}) = \frac{n_i - V_i}{N - V} \tag{8.21}$$

ou encore (pour éviter un problème avec les valeurs $V = 1$ et $V_i = 0$) :

$$P(m_i|R) = \frac{V_i + 0.5}{V + 1} \tag{8.22}$$

$$P(m_i|\overline{R}) = \frac{n_i - V_i + 0.5}{N - V + 1} \tag{8.23}$$

et, plus souvent :

$$P(m_i|R) = \frac{V_i + \frac{n_i}{N}}{V + 1} \tag{8.24}$$

$$P(m_i|\overline{R}) = \frac{n_i - V_i + \frac{n_i}{N}}{N - V + 1} \tag{8.25}$$

De nombreuses méthodes d'apprentissage ont été proposées, à partir d'approches bayésiennes (Bookstein, 1983), de modèle 2-Poisson ou bien de mixture de n distributions de Poisson.

À partir du modèle probabiliste originel, Robertson et l'équipe du *Centre for Interactive Systems Research* de City University (Londres) y ont intégré la possibilité de tenir compte de la fréquence d'apparition des mots dans les documents et dans la requête ainsi que de la longueur des documents. Cette intégration correspondait originellement à l'intégration du modèle 2-poisson de Harter (utilisé par ce dernier pour sélectionner les bons termes d'indexation et non pour les pondérer) dans le modèle probabiliste. À partir du modèle 2-poisson et de la notion d'ensemble d'*élite* E pour un mot (selon Harter, l'ensemble des documents les plus représentatifs de l'usage du mot ; plus généralement : l'ensemble des documents qui contiennent le mot), sont dérivées les probabilités conditionnelles $p(E|R), p(\bar{E}|R), p(E|\bar{R})$ et $p(\bar{E}|\bar{R})$ donnant un nouveau modèle probabiliste dépendant de E et de $\bar{E}$. Avec la prise en compte d'autres variables telles la longueur des documents et le nombre d'occurrences du mot au sein du document, ce modèle a donné lieu à une famille de pondérations dénommées *BM* (*Best Match*).

De manière générale, la prise en compte des *poids* w des mots dans les documents et dans la requête s'exprime par :

$$sim(d_j, q) = \sum_{m_i \in d_j \cap q} w_{m_i,d_j} \cdot w_{m_i,q} \cdot \log \frac{p_i(1 - q_i)}{q_i(1 - p_i)} \tag{8.26}$$

Lorsque l'on n'a pas d'informations sur l'ensemble R des documents pertinents, il est d'usage de transformer cette égalité en un classique produit scalaire :

$$sim(d_j, q) = \sum_{m_i \in d_j \cap q} w_{m_i,d_j} \cdot w_{m_i,d_j} \tag{8.27}$$

Il reste maintenant à estimer les poids w. Ceci peut être fait en intégrant des modèles statistiques ou bien en s'inspirant des pondérations du modèle vectoriel.

Intégration du modèle 2-poisson de Harter

On obtient une loi de Poisson lorsque le nombre d'événements est assez grand et que la probabilité élémentaire est faible (exemples : défauts dans une chaîne de fabrication, erreurs de frappe dans une page). Certaines expériences ont montré que seule la distribution de 50 % (mais jusqu'à 70 % selon d'autres) des mots peut s'apparenter à un modèle de Poisson (Fuhr, 1992).

Définition 5 (Distribution de Poisson) *Sachant μ_i, le nombre d'occurrences moyen du mot m_i par document dans un ensemble de documents R, la probabilité que le nombre d'occurrences $f(m_i, d)$ de m_i dans un document d soit égal à k est de :*

$$P(f(m_i, d) = k|R) = e^{-\mu_i} \cdot \frac{\mu_i^k}{k!} \tag{8.28}$$

L'espérance et la variance d'une telle distribution sont égales à μ_i.

Les hypothèses formulées par Harter disent qu'un mot non particulièrement informatif est distribué sur la collection selon un modèle de Poisson et qu'inversement, un mot très spécifique ne respecte pas un tel modèle. Pour déterminer la quantité d'informations liée à un mot, il faut alors évaluer à quel point sa distribution sur la collection s'écarte du modèle. Une seconde hypothèse dit alors qu'un mot spécifique suit en fait un modèle de Poisson mais seulement pour un sous-ensemble de la collection : l'ensemble d'élite (avec une moyenne associée à ce modèle plus grande que celle du modèle de la collection complète pour le même mot).

Définition 6 (Modèle 2-Poisson) *On suppose qu'il existe deux classes de documents associées à un mot "plein" (par opposition à un mot outil i.e. non discriminant pour la requête en cours) : la classe "non privilégiée" (elle contient des documents pour lesquels la présence du mot, souvent faible, est accidentelle ; le mot devant alors considéré comme un mot outil) et la classe "privilégiée" (elle contient des documents dans lesquels le mot est central et apparaît souvent). Soient π la probabilité qu'un document soit dans la classe privilégiée (et $1 - \pi$ qu'il n'y soit pas) et μ_i et λ_i les nombres moyens d'occurrences de m_i dans ces deux classes :*

$$P(f(m_i, d) = k|R) = \pi e^{-\mu_i} \cdot \frac{\mu_i^k}{k!} + (1 - \pi) e^{-\lambda_i} \cdot \frac{\lambda_i^k}{k!} \tag{8.29}$$

La formulation de RSV devient alors :

$$\begin{aligned} RSV(d_j, q) = sim(d_j, q) &\sim \sum_{m_i \in q \cap d_j} \left(\log \frac{P(m_i|R)}{1 - P(m_i|R)} + \log \frac{P(m_i|\overline{R})}{1 - P(m_i|\overline{R})} \right) \\ &= \sum_{m_i \in q \cap d_j} \left(\log \frac{e^{-\mu_{m_i}} \cdot \frac{\mu_{m_i}^{f(m_i,d_j)}}{f(m_i,d_j)!}}{1 - e^{-\mu_{m_i}} \cdot \frac{\mu_{m_i}^{f(m_i,d_j)}}{f(m_i,d_j)!}} + \log \frac{e^{-\overline{\mu_{m_i}}} \cdot \frac{\overline{\mu_{m_i}}^{f(m_i,d_j)}}{f(m_i,d_j)!}}{1 - e^{-\overline{\mu_{m_i}}} \cdot \frac{\overline{\mu_{m_i}}^{f(m_i,d_j)}}{f(m_i,d_j)!}} \right) \\ &= \sum_{m_i \in q \cap d_j} f(m_i, d_j) \cdot \log \frac{\mu_i}{\overline{\mu_i}} \end{aligned} \tag{8.30}$$

Si l'on intègre la longueur des documents comme paramètre (il est en effet plus probable que le nombre d'occurrences d'un mot soit plus important dans un document long que dans un document court), la probabilité que le nombre d'occurrences $f(m_i, d)$ de m_i dans un document d soit égal à k devient :

$$P\left(f(m_i, d) = k|R\right) = e^{-\mu \cdot l_{d_j}} \cdot \frac{(\mu \cdot l_{d_j})^k}{k!} \tag{8.31}$$

avec l_{d_j} la longueur de d_j en nombre d'occurrences de mots.

Intégration d'un modèle gaussien

Si l'on considère que les mots sont maintenant distribués selon une loi normale, la similarité proposé en 1982 par Bookstein (1983) est :

$$RSV(d_j, q) = \sum_{m_i \in q \cap d_j} f(m_i, d_j) \left[\left(\frac{\mu_{m_i}}{\sigma_{m_i}^2} - \frac{\overline{\mu_{m_i}}}{\overline{\sigma_{m_i}}} \right) - \frac{f(m_i, d_j)}{2} \cdot \left(\frac{1}{\sigma_{m_i}^2} - \frac{1}{\overline{\sigma_{m_i}}} \right) \right] \tag{8.32}$$

avec μ et σ les moyennes et les écarts-types dans R et dans $\bar{R}$.

Les pondérations Okapi

Lors des différentes campagnes d'évaluation TREC, de nombreuses pondérations ont été testées au sein des systèmes Okapi (Robertson et al., 1994) et Inquery (Allan et al., 1996) dont la célèbre BM25 (8.37) fondée sur le modèle 2-poisson. C'est également ce type de pondération qui est utilisé par le module de recherche documentaire de nombreux systèmes de questions-réponses (par exemple Brill et al., 2001).

Une manière courante de définir la composante IDF (*Inverse Document Frequency*) avec N le nombre de documents dans la collection et $n(m_i)$ le nombre de documents

contenant m_i dans la collection est[2] :

$$IDF(m_i) = \log\left(\frac{N - n(m_i) + 0.5}{n(m_i) + 0.5}\right) \tag{8.34}$$

Le nombre d'occurrences $f(m_i, d_j)$ est généralement normalisé suivant la longueur moyenne $\bar{l}$ des documents de la collection et $l(d_j)$ la taille (en nombre d'occurrences de mots) de d_j. Avec K une constante réelle, habituellement choisie entre 1 et 2, une possibilité consiste à définir la composante TF de telle sorte de favoriser les documents courts :

$$TF(m_i, d_j) = \frac{(K+1) \cdot f(m_i, d_j)}{f(m_i, d_j) + K \cdot (l(d_j)/\bar{l})} \tag{8.35}$$

Un grand nombre de pondérations ont été testées dont les premiers résultats ont été publiés à l'issue des campagnes d'évaluation TREC-2 et TREC-3. La définition de ces nouvelles pondérations fut concomitante de la généralisation de l'expansion automatique de requête à partir des premiers documents trouvés.

Soient :
- N le nombre de documents dans la collection ;
- $n(m_i)$ le nombre de documents contenant le mot m_i ;
- R le nombre de documents connus comme étant pertinents pour la requête q ;
- $r(m_i)$ le nombre de documents de R contenant le mot m_i ;
- $tf(m_i, d_j)$ le nombre d'occurrences de m_i dans d_j ;
- $tf(m_i, q)$ le nombre d'occurrences de m_i dans q ;
- $l(d_j)$ la taille (en nombre de mots) de d_j ;
- $\bar{l}$ la taille moyenne des documents de la collection ;
- k_i et b des paramètres dépendants de la requête et, si possible, de la collection.

Le poids w d'un mot m_i est défini par :

$$w(m_i) = \log \frac{(r(m_i) + 0.5)/(R - r(m_i) + 0.5)}{(n(m_i) - r(m_i) + 0.5)/(N - n(m_i) - R + r(m_i) + 0.5)} \tag{8.36}$$

Définition 7 (BM25) *La pondération **BM25** est définie de la manière suivante :*

$$sim(d_j, q) = \sum_{m_i \in q} \left(w(m_i) \times \frac{(k_1+1) \cdot tf(m_i, d_j)}{K + tf(m_i, d_j)} \times \frac{(k_3+1) tf(m_i, q)}{k_3 + tf(m_i, q)} \right) \tag{8.37}$$

avec :

$$K = k_1 \cdot \left((1 - b) + b \cdot \frac{l(d_j)}{\bar{l}} \right) \tag{8.38}$$

[2]Etant donné que $n(m_i)$ est généralement petit par rapport à N, on peut parfois simplifier cette définition en :

$$IDF(m_i) = \log\left(\frac{N + 0.5}{n(m_i) + 0.5}\right) \tag{8.33}$$

Lorsqu'on n'a pas d'informations sur R et $r(m_i)$, cette définition se réduit à (pondération utilisée dans le système Okapi durant TREC-1) :

$$w(m_i) = \log \frac{N - n(m_i) + 0.5}{n(m_i) + 0.5} \tag{8.39}$$

avec $R = r(m_i) = 0$. Ce sont ces valeurs qui sont utilisées dans les deux exemples suivants.

Lors de la campagne TREC-8, le système *Okapi* a été utilisé avec les valeurs : $k_1 = 1.2$, $b = 0.75$ (des valeurs inférieures de b sont parfois intéressantes) et pour les longues requêtes, k_3 est positionné soit à 7 soit à 1000 :

$$sim(d_j, q) = \sum_{m_i \in q} \frac{2.2 \cdot tf(m_i, d_j)}{0.3 + 0.9 \cdot \frac{l(d_j)}{\bar{l}} + tf(m_i, d_j)} \times \frac{1001 \cdot tf(m_i, q)}{1000 + tf(m_i, q)} \times \log_2 \frac{N - n(m_i) + 0.5}{n(m_i) + 0.5} \tag{8.40}$$

Le système *Inquery* (Allan et al., 1996) utilise BM25 avec $k_1 = 2$, $b = 0.75$ et $\forall i$: $tf(m_i, q) = 1$:

$$sim(d_j, q) = \sum_{m_i \in q} \frac{tf(m_i, d_j)}{0.5 + 1.5 \cdot \frac{l(d_j)}{\bar{l}} + tf(m_i, d_j)} \times \frac{\log_2 \frac{N+0.5}{n(m_i)}}{log_2(N+1)} \tag{8.41}$$

De nombreuses variantes. Certains auteurs ont suggéré de pondérer les mots en fonction de l'écart de leur distribution observée en corpus par rapport à des distributions aléatoires dont le principal avantage est de ne pas nécessiter de paramètres dont les valeurs doivent être apprises sur les données (Amati et Van Rijsbergen, 2002). Cela a conduit à la génération de plusieurs modèles dont certains ont obtenu des résultats meilleurs que BM25 sur les collections TREC *ad-hoc*. Il a également été proposé de tenir compte d'un plus grand nombre de caractéristiques que celles énoncées précédemment, par exemple la variance des occurrences dans la collection (Greiif et al., 2002).

Chapitre 9

La méthode d'enrichissement de requêtes de Rocchio

Dans le modèle vectoriel (p. 141), le processus d'enrichissement défini par Rocchio (1971) correspond à une modification du vecteur requête.

Définition 8 (Expansion selon Rocchio) *Soient q_0 la requête initiale, $|d|$ le nombre de documents pertinents, n le nombre de documents non pertinents, d_i un document pertinent, N_i un document non pertinent. La requête enrichie q_1 se calcule de la manière suivante (les facteurs β et γ sont généralement choisis proportionnellement à $\frac{1}{|d|}$ et à $\frac{1}{n}$ respectivement) :*

$$\overrightarrow{q_1} = \alpha \overrightarrow{q_0} + \beta \sum_{i=1}^{|d|} \overrightarrow{d_i} - \gamma \sum_{j=1}^{n} \overrightarrow{N_j} \tag{9.1}$$

La formulation de Rocchio s'adapte également au modèle probabiliste (page 145). Selon la définition initiale du modèle probabiliste, il est possible d'obtenir un premier jeu de documents qui seront examinés par l'utilisateur selon les paramètres initiaux de la formule 8.19 (p. 148).

Après sélection d'un ensemble de documents pertinents et d'un ensemble de documents non pertinents par l'utilisateur, les probabilités deviennent :

$$P(k_i|d) = \frac{|d_i|}{|d|} \tag{9.2}$$

$$P(k_i|\overline{d}) = \frac{n_i - |d_i|}{N - |d|} \tag{9.3}$$

avec d_i le nombre de documents pertinents contenant k_i et $|d|$ le nombre total de documents pertinents.

Pour éviter certains problèmes avec des valeurs faibles de $|D_i|$ et de $|D|$, Yu et al. (1983) ont proposé les définitions suivantes :

$$P(k_i|d) = \frac{|d_i| + \frac{n_i}{N}}{|d| + 1} \tag{9.4}$$

$$P(k_i|\bar{d}) = \frac{n_i - |d_i| + \frac{n_i}{N}}{N - |d| + 1} \tag{9.5}$$

Dans le cas de plusieurs itérations, et contrairement au modèle vectoriel, il n'est pas tenu compte des probabilités des itérations précédentes. En outre, les fréquences d'apparition des mots ne sont pas considérées et il n'est pas possible d'ajouter de nouveaux mots à la requête. Pour toutes ces raisons l'enrichissement fonctionne généralement moins bien dans le modèle probabiliste que dans le modèle vectoriel (Baeza-Yates, 1999). Croft (1983) a toutefois proposé une extension du modèle probabiliste qui tient compte des fréquences des mots dans les documents. Selon ce modèle, la similarité entre un document et une requête est définie comme suit :

Définition 9 (Mesure de Croft)

$$sim(d_j, q) = \sum_{m_i \in q} w_{i,q} \cdot w_{i,d_j} \cdot \left(\log \frac{P(k_i|d)}{1 - P(k_i|d)} + \log \frac{1 - P(k_i|\bar{d})}{P(k_i|\bar{d})} + C \right) \cdot \left(K + (1 - K) \frac{f_{i,j}}{max(f_{i,j})} \right) \tag{9.6}$$

avec m_i un mot, C et K deux constantes et $f_{i,j}$ le nombre d'occurrences de k_i dans d_j.

Enrichissement automatique. Les moteurs de recherche ont tendance à ramener plus de documents pertinents en tête de liste qu'en queue. En fonction de ce constat, il est possible d'effectuer un enrichissement automatique en « pariant » que les x premiers documents sont des documents pertinents et en procédant ensuite exactement comme pour l'enrichissement interactif (la rétroaction négative en moins).

En ce qui concerne l'extraction automatique de mots proches de ceux de la requête, on peut l'étendre à l'ensemble des documents trouvés (voire de la collection entière) et non plus uniquement aux documents sélectionnés par l'utilisateur. Une autre manière simple d'enrichir une requête est de considérer que les mots des documents pertinents qui ne sont pas trop éloignés des mots de la requête (au sens de leur position dans la phrase) sont de bons candidats pour une nouvelle recherche. En réalité les mots sémantiquement les plus proches de ceux de la requête ne sont pas seulement déterminés en fonction de leur localisation mais plutôt suivant le calcul d'un score d'association. Cette option s'inspire de travaux en extraction de terminologie (Jacquemin, 1997; Bourigault et C., 2000; Daille, 2002).

Monz (2003) a montré qu'un tel processus dégradait les performances de son système de questions-réponses (avec recherche documentaire vectorielle) alors qu'il était efficace sur la seule tâche recherche documentaire ad-hoc de TREC-8. La solution proposée (mais non testée à ma connaissance) est d'employer une analyse locale pour l'enrichissement tel que décrit dans (Xu et Croft, 1996).

Chapitre 10

Mesures d'évaluation en recherche documentaire et en questions-réponses

L'objectif de cette partie est à la fois de rappeler les critères d'évaluation utilisés en RI mais aussi de faire le lien avec la tâche questions-réponses en présentant des mesures adaptées.

Le souci d'évaluer les performances des systèmes de recherche d'information (SRI) date des années 50 (Berry, 1955) et les techniques d'évaluation sont elles-mêmes évaluées, entre autres dans (Saracevic, 1995). L'évaluation a été abordée selon deux angles principaux : l'efficience et l'efficacité. L'efficience mesure la capacité d'un SRI à répondre rapidement à une requête de l'utilisateur. L'efficacité, quant à elle, mesure la capacité d'un SRI à sélectionner uniquement des documents pertinents.

L'évaluation de la qualité d'un SRI en termes de pertinence des résultats est une tâche subjective puisqu'elle dépend de référentiels (listes des bonnes réponses) construits par des individus qui s'accordent dans seulement 70 % ou 80 % des cas sur la pertinence ou non d'un document (Harman, 1995). D'autre part, la notion de pertinence d'un document dépend des objectifs (compréhension d'un problème ou simple description, compléments d'information ou recherche d'une présentation générale) et du point du vue de l'utilisateur. Malgré la complexité de cette tâche, plusieurs métriques reconnues et largement utilisées en RI ont été proposées, les plus importantes sont détaillées dans Fluhr (2004) et résumées ici.

Précision. La précision mesure la proportion de documents pertinents dans la liste de documents retournés par le SRI.

Rappel. Le rappel rend compte de la quantité de bonnes réponses par rapport au nombre de documents pertinents dans le corpus. Autrement dit, le rappel est le taux de documents pertinents trouvés par rapport au nombre de documents pertinents à trouver. Le nombre total de documents pertinents dans une collection peut être obtenu

si les documents du corpus sont connus et jugés par des individus. Pour questions-réponses, il correspond au nombre total de documents qui contiennent une réponse exacte et supportée.

La **F-mesure** correspond à une moyenne harmonique combinant rappel et précision.

Courbe rappel / précision. La précision peut être calculée à différents niveaux de rappel. Supposons par exemple qu'à une requête donnée correspondent deux documents pertinents, que le SRI propose 100 documents et que les documents pertinents sont en position 2 et 8. Pour atteindre un rappel de 50 %, il faut atteindre la position 2 de la liste. À ce rang, la précision est aussi de 50 %. Pour atteindre 100 % de rappel, il faut aller jusqu'à la position 8 : la précision est de 25 % (2 documents sur 8 sont pertinents).

La précision moyenne. Elle tient compte à la fois du rappel et de la précision. Elle correspond à la moyenne des précisions non interpolées associées aux positions – dans la liste des documents trouvés – de chaque document pertinent à trouver. Lorsqu'un document pertinent du référentiel n'est pas présent dans la liste des documents rapportés, la valeur qui lui est associée est nulle. La précision moyenne est la moyenne arithmétique de toutes ces valeurs.

10.0.7 Mesures spécifiques à la tâche questions-réponses

Plusieurs critères d'évaluation ont été proposés pour mesurer la qualité des systèmes de questions-réponses depuis la fin des années 1990 et les premières évaluation TREC QA (Gillard et al., 2006b).

Questions factuelles et booléennes. Parmi les critères les plus courants, citons le MRR, le CWS et la K-Mesure.

Définition 10 (*Mean Reciprocal Rank* – MRR) *Soit un système de questions-réponses qui retourne à l'utilisateur une liste de réponses ordonnées selon ses propres critères de la plus probable à la moins probable. Le score $s(q_i)$ correspondant à une question q_i est l'inverse du rang de la première réponse jugée correcte dans cette liste. S'il n'y en a aucune, le score est nul. Le meilleur système est celui qui obtient le score le plus élevé. Le MRR désigne la moyenne des scores sur l'ensemble $\mathbb{Q}$ des questions q_i :*

$$0 \leqslant MRR = \frac{1}{|\mathbb{Q}|} \sum_{i=1}^{i=|\mathbb{Q}|} s(q_i) \leqslant 1 \tag{10.1}$$

Définition 11 (*Confidence Weighted Score* – CWS) *Soit un système qui répond à un ensemble $\mathbb{Q}$ de questions q_i et qui est capable d'estimer la confiance qu'il a en l'exactitude de ses réponses. Il retourne alors à l'utilisateur* une seule réponse par question, *en ordonnant les réponses par ordre décroissant de confiance : la première réponse fournie correspond à la question pour laquelle il est le plus sûr de sa réponse et ainsi de suite. Le CWS récompense le système sur ses capacités à répondre correctement aux questions et, simultanément, à estimer leur exactitude. Il correspond à la moyenne arithmétique du nombre de questions $n(i)$ auxquelles il a*

correctement répondu dans les i premiers rangs (le meilleur système est celui qui obtient la valeur la plus élevée) :

$$0 \leqslant CWS = \frac{1}{|\mathbb{Q}|} \sum_{i=1}^{i=|\mathbb{Q}|} n(i) \leqslant 1 \tag{10.2}$$

La mesure CWS a tendance à être mise de côté ces dernières années à cause de sa tendance à favoriser les stratégies d'ordonnancement par rapport aux capacités à répondre correctement au plus grand nombre de questions possible. Pourtant, la capacité d'un système à estimer correctement la qualité de ses réponses est un enjeu primordial pour les années à venir tant les questions de la pertinence et de l'exactitude d'une réponse sont au centre de nombreuses utilisations réelles des moteurs de recherche.

Dans la suite de CWS et pour répondre aux critiques formulées, la **K-mesure** a été introduite afin de tenir compte de l'estimation de la confiance mais aussi pour pénaliser plus fortement les réponses incorrectes (Herrera et al., 2004). Dans ces conditions, il est préférable de ne pas répondre aux questions si le score de confiance est trop faible plutôt que de proposer une réponse inexacte.

Questions "Listes". Les questions appelant non pas une réponse mais une *liste de réponses* sont évaluées en utilisant le critère de précision moyenne (*non interpolated average precision*) tel qu'utilisé dans les campagnes en recherche documentaire *ad-hoc*. Ce critère tient compte à la fois du rappel et de la précision mais aussi de la position des bonnes réponses dans la liste. Si une bonne réponse se trouve en position 10, la précision qui lui est associée pour le calcul de la précision moyenne correspond au taux de bonnes réponses parmi les 10 premières réponses. Pour chaque réponse à trouver qui n'est pas présente dans la liste évaluée, une précision nulle lui est associée (c'est de cette manière que le rappel est pris en compte). Ainsi pour une question q_i de type "liste" à laquelle le système répond par n réponses, $rep_1, \cdots, rep_n$ et pour laquelle R bonnes réponses sont à trouver, la précision moyenne vaut :

$$0 \leqslant précisionMoyenne(q_i) = \frac{\sum_{j=1}^{n} \mathcal{I}(rep_j) \cdot précision(j)}{R} \leqslant 1 \tag{10.3}$$

avec :

$$\mathcal{I}(rep_j) = \begin{cases} 1 \text{ si } rep_j \text{ est une bonne réponse} \\ 0 \text{ sinon} \end{cases}$$

et $précision(j)$, le nombre de bonnes réponses *différentes* jusqu'au rang j :

$$précision(j) = \frac{\sum_{k=1}^{k=j} \mathcal{I}(rep_k)}{j} \tag{10.4}$$

Pour un ensemble $\mathbb{Q}$ de questions "listes" q_i, il est possible de calculer la moyenne des précisions moyennes :

$$précisionMoyenne_{\mathbb{Q}} = \frac{\sum_{\mathbb{Q}} précisionMoyenne(q_i)}{|\mathbb{Q}|} \tag{10.5}$$

Exemples : soit la question q pour laquelle 3 (bonnes) réponses sont à trouver :
- si la liste à évaluer présente les 3 réponses à trouver dans les 3 premières positions, la précision moyenne vaut : $(1/1 + 2/2 + 3/3)/3 = 1$;
- si seulement 2 bonnes réponses sont trouvées et qu'elles se trouvent en positions 1 et 2, la précision moyenne vaut : $(1/1 + 2/2)/3 = 2/3$;
- si seulement 1 bonne réponse est trouvée et qu'elle se trouve en position 5, la précision moyenne vaut : $(1/5)/3$.

Inconvénients et difficultés :
- Les mauvaises réponses positionnées après la dernière bonne réponse trouvée n'ont pas d'influence : les systèmes ont toujours intérêt à proposer le maximum autorisé de réponses ;
- l'évaluateur doit se méfier des réponses correctes équivalentes qui seraient proposées plusieurs fois dans la liste...

Questions définitoires. Pour ce dernier type de questions, une mesure d'évaluation automatique a été proposée par Marton (2006) à partir des définitions des mesures ROUGE utilisées pour la tâche de résumé automatique (voir (Minel, 2004) *in* (Chaudiron, 2004a) pour une présentation des mesures spécifiques au résumé automatique).

10.0.8 Mesures de recherche documentaire spécifique à QR

Au-delà des mesures classiquement utilisées en questions-réponses, certaines sont définies pour mesurer uniquement la phase de recherche documentaire : le *Mean Reciprocal Document Rank* (MRDR) donne une indication sur le rang moyen du premier document contenant une bonne réponse (Prager et al., 1999). Le *Reciprocal Document Rank* (RDR) vaut 1 si le premier document contient la bonne réponse, $\frac{1}{2}$ si le premier ne contient pas de bonne réponse mais le second oui *etc.* Si aucun document trouvé ne contient de bonne réponse le RDR vaut 0. Le MRDR est la moyenne de ces valeurs.

Chapitre 11

De TREC à CLEF en passant par EQUER : synthèse des résultats

Nous dressons dans cette partie une synthèse des résultats que nous avons obtenus lors des différentes évaluations auxquelles nous avons participé ainsi que les architectures du système SQuaLIA.

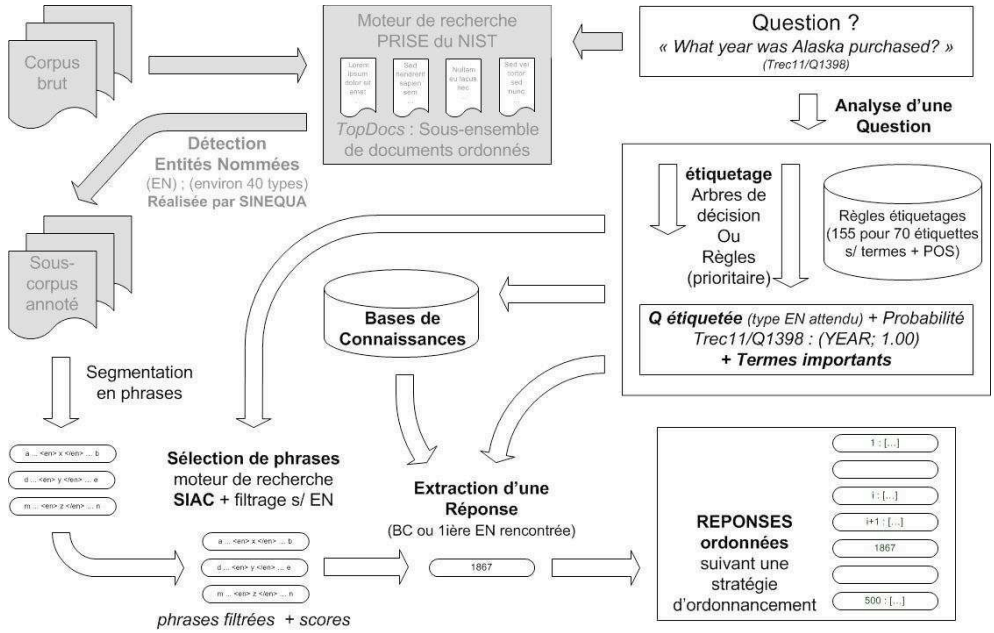


FIG. 11.1: *Schéma fonctionnel du système de questions-réponses développé durant notre participation à la campagne TREC-11 en 2002. Ce prototype utilisait le module de reconnaissance d'entités nommées de la société Sinequa.*

Évaluation		*Nombre de réponses correctes* (pour des questions) *Définitoires (33)*	*Factuelles (400)*	*Booléennes (31)*	*Total (464)* en #	en %	*MRR*	*Nombre de questions répondues (464)*
Run1	***Passages***	17	153	12	182	39,2	0,33	388
	Courtes	9	118	12	139	29,5	0,25	388
Run2	***Passages***	14	137	11	162	34,9	0,29	354
	Courtes	8	111	11	130	27,6	0,23	354

TAB. *11.1:* *Résultats officiels de notre participation à EQueR. Nombre de réponses correctes et MRR par soumission (Run1 et Run2), par type d'évaluation (sur les passages, sur les réponses courtes) et par famille de questions (définitoires, factuelles, et booléennes).*

#Réponses Correctes par Type		**#Q Définitoires (33)**	**#Q Factuelles (400)**	**#Q Booléennes (31)**	**#TOTAL réponses correctes**	**soit %**	**#réponses fournies (464)**
OFFICIEL	PASSAGE	17	153	12	182	39,2	388
	COURTE	9	118	12	139	29,5	388
APRÈS CORR.	PASSAGE	31	248	13	292	62,9	416
	COURTE	26	196	13	235	50,4	416

TAB. *11.2:* *Résultats de notre participation à EQueR après correction d'un certain nombre de bugs lors de la soumission officielle. Nombre de réponses correctes et MRR par type d'évaluation (sur les passages, sur les réponses courtes) et par famille de questions (définitoires, factuelles, et booléennes). Cette seconde évaluation a été faite manuellement avec l'aide des organisateurs d'EQUER dans un cadre strictement identique.*

runs	Right	Fact.	Def.	Temp.	R NIL	NIL answered	ineXact	Unsupported
FR-FR1	88	56 (+1)	32	0	2	30	7 (-1)	2
EN-FR1	67	40 (+1)	27	0	5	34	7 (-1)	2

TAB. *11.3:* *Evaluation de SQuaLIA sur les questions factuelles et définitoires de CLEF 2006 pour les tâches FR-FR et EN-FR. Pour FR-FR, nous avons su répondre à 5 questions "listes" supplémentaires, ce qui porte le taux de bonnes réponses à 93 sur 200. Notons la bonne performance sur les questions définitoires : 32 bonnes réponses sur 41.*

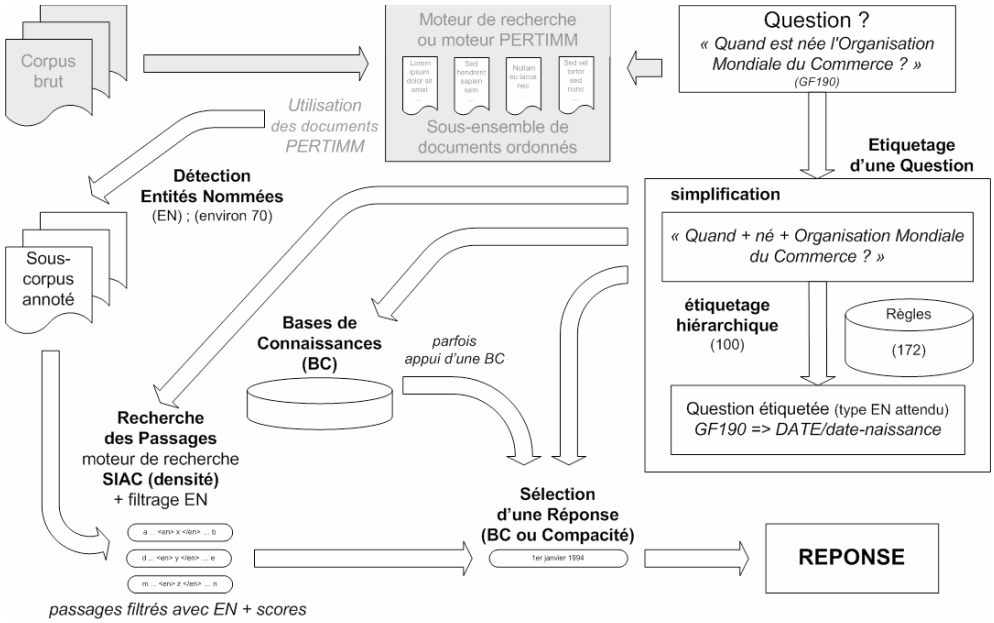


FIG. *11.2:* *Schéma fonctionnel du système de questions-réponses développé durant notre participation à la campagne EQUER en 2004.*

Chapitre 12

Quelques méthodes de segmentation linéaire non supervisées

DotPlotting (Reynar). Cette approche est une adaptation pour la segmentation de la méthode des nuages de points présentée par (Helfman, 1994) pour la recherche d'information. Il se base sur une représentation graphique du texte (figure 12.1) par les positions des occurrences des mots du texte à segmenter. Lorsqu'un mot apparaît à deux positions du texte x et y, les quatre points $(x,x),(x,y),(y,x)$ et (y,y) sont représentés sur un graphe, ce qui permet de déterminer visuellement les zones du texte où les répétitions sont nombreuses. Cette méthode est adaptée par Reynar (2000) à la segmentation thématique de textes. Les positions de début et de fin des zones les plus denses du graphe sont les limites des segments thématiquement cohérents. La densité est calculée pour chaque unité d'aire en divisant le nombre de points d'une région par l'aire de cette région. À partir de là, deux algorithmes peuvent déterminer les frontières thématiques : identifier les limites en maximisant la densité au sein des segments, ou repérer la configuration qui minimise la densité des zones entre les segments.

C99 (Choi). C99 utilise une mesure de similarité entre chaque unité textuelle. L'idée de base de cette méthode est que les valeurs des mesures de similarité entre des segments de textes courts sont statistiquement insignifiantes, et que seul des classements locaux sont à considérer pour appliquer un algorithme de catégorisation sur la matrice de similarité.

Dans un premier temps, une matrice de similarité est construite, représentant la similarité (cosinus) entre toutes les phrases du texte prises deux à deux. On effectue ensuite un classement local, en déterminant pour chaque paire de phrases, le rang de sa similarité par rapport à ses $m \times n - 1$ voisins, mn étant le *masque* de classement choisi. Le rang est le nombre d'éléments voisins ayant une mesure de similarité plus faible, conservé sous la forme d'un ratio r (égal au rapport entre le rang et le nombre de voisins

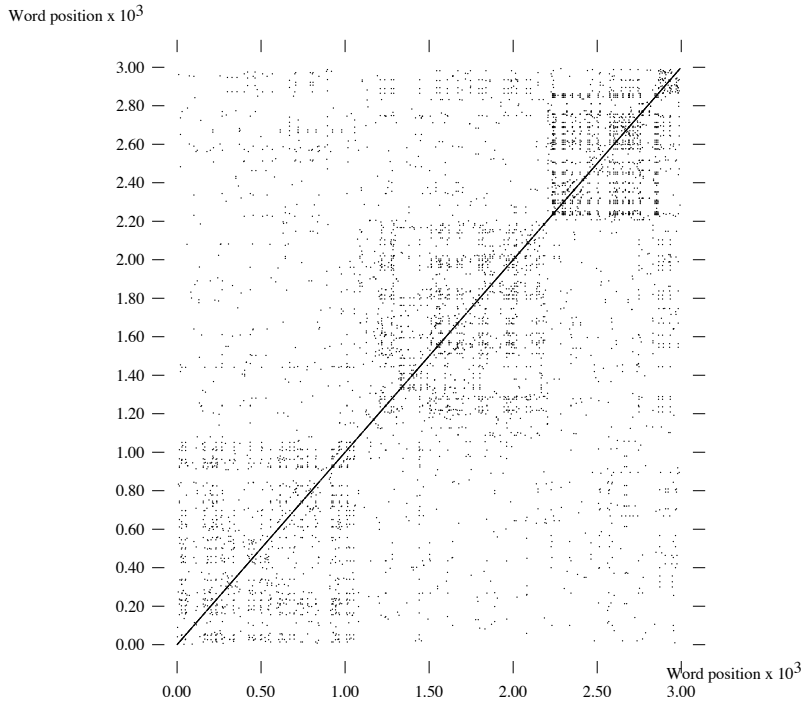


FIG. 12.1: *Méthode de segmentation DotPlotting (Reynar, 2000) : chaque point représente un mot et les axes la localisation dans le texte. Sur cette figure, on distingue 3 zones denses qui sont autant de segments.*

dans le masque) afin de prendre en compte les effets de bord. Enfin, la dernière étape détermine les limites de chaque segment. Sur la figure 12.2, on peut voir une illustration de cette méthode où différentes répartitions de segments sont testées. Les segments sont représentés par des carrés le long de la diagonale de la matrice de similarité modifiée avec les classements locaux. Pour chaque segment de la répartition proposée à une étape donnée, on considère son aire notée a_k ainsi que son poids s_k qui est la somme des tous les rangs des phrases qu'il contient[1]. On calcule alors la densité D avec :

$$D = \frac{\sum_k s_k}{\sum_k a_k} \tag{12.1}$$

La subdivision des segments s'arrête lorsque la densité est suffisamment faible, ou lorsqu'est atteint le nombre de segments désirés.

Dias et Alves (2005) ont proposé d'utiliser une similarité basée sur la position relative des mots dans les segments plutôt que sur une pondération de type $tf.idf$. Selon cette méthode et si x et y sont les deux segments dont on veut estimer la similarité, le poids d'un mot w dans x est la somme Σ des écarts en nombre de mots au sein de x

[1] Si le segment k commence à la phrase i et se termine à la phrase j, son aire a_k est égale à $(j-i+2)^2$.

entre w et les autres mots w' qui appartiennent simultanément à x et à y :

$$\Sigma_y(x,w) = \sum_{\forall w' \neq w, w' \in x \cap y} |pos_x(w) - pos_x(w')| \tag{12.2}$$

où $pos_x(w)$ désigne la position de w dans le segment x.

La similarité entre x et y peut être ensuite calculée classiquement à l'aide d'un cosinus entre les vecteurs $\overrightarrow{\Sigma_y}$ et $\overrightarrow{\Sigma_x}$

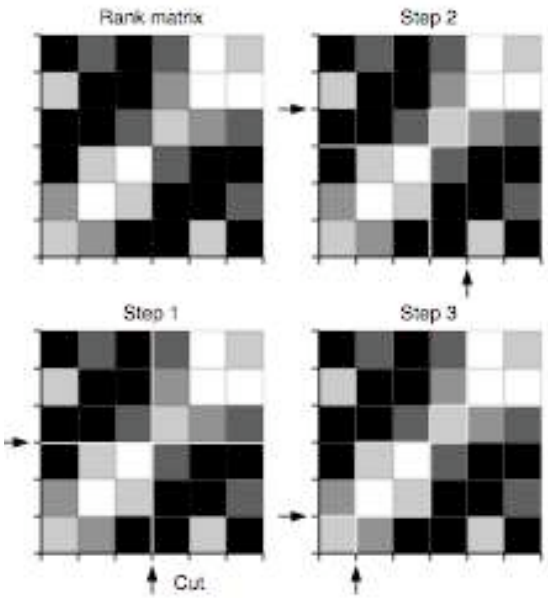


FIG. 12.2: *Méthode de segmentation C99 (Choi, 2000) : la segmentation est itérative en fonction de la densité des segments considérés, elle-même calculée à partir des similarités des phrases prises deux à deux.*

Text-Tiling (Hearst). Elle étudie la distribution des mots dans les textes selon plusieurs critères. Un score de cohésion est attribué à chacun des segments en fonction du segment qui le suit. Ce score dépend lui-même d'un second score, attribué à chaque paire de phrases en fonction de la paire de phrases qui la suit. Ce deuxième score est calculé en tenant compte des mots communs, du nombre de mots nouveaux, et du nombre de chaînes lexicales actives dans les phrases considérées (voir figure 12.3). Le score d'un segment est finalement le produit scalaire normalisé des scores de chaque paire de phrases qu'il contient. Si l'écart entre le score d'un segment et les scores des segments qui l'entourent est grand, une frontière est apposée (voir figure **??**).

Segmenter (Kan). Il effectue une segmentation linéaire basée sur les chaînes lexicales présentes dans le texte. Ces chaînes relient les occurrences des mots dans les phrases. Une chaîne est rompue si le nombre de phrases séparant deux occurrences est trop

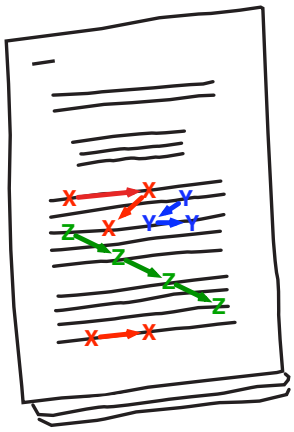


FIG. 12.3: *Les chaînes lexicales réunissent les occurrences d'un même mot au sein d'un document. La longueur maximale d'une chaîne est définie en fonction du hiatus, longueur du saut maximal entre deux occurrences. Sur cette figure, 4 chaînes sont représentées, deux pour le mot X, une pour Y et une pour Z. Pour une phrase donnée, une chaîne est dite active si sa localisation correspond au moins en partie à celle de la phrase.*

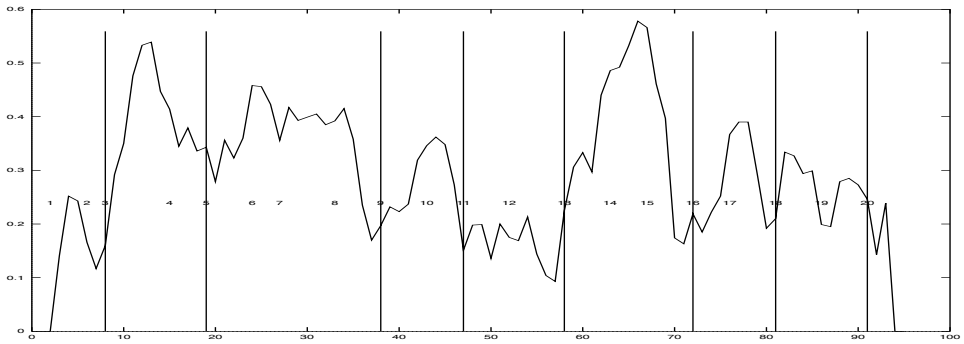


FIG. 12.4: *Méthode de segmentation Text-Tiling (Hearst, 1997). La rupture entre deux segments est située dans une zone du texte entourée de zones présentant des valeurs de cohésion très différentes de la sienne.*

important. Ce nombre dépend de la catégorie syntaxique du mot considéré. Une fois tous les liens établis, un poids leur est assigné en fonction de la catégorie syntaxique des mots en jeu et de la longueur du lien. Un score est ensuite donné à chaque paragraphe en fonction des poids et des origines des liens qui le traversent ou qui y sont créés. Les marques de segmentation sont alors apposées au début des paragraphes ayant les scores maximaux.

Chapitre 13

Curriculum vitae

PBellot

Déroulement de carrière

- 2000- : Maître de conférences ;
- 2004- : titulaire de la Prime d'Encadrement Doctoral et de Recherche ;
- 2007- : chargé de cours à l'Ecole Supérieure d'Ingénieurs de Luminy (Univ. Aix-Marseille II) ;
- 2008 : chargé de cours en recherche d'informations à l'Université de Dalat (Vietnam) ;
- 1999-2000 : Attaché Temporaire d'Enseignement et de Recherche (ATER) ;
- 1996-1999 : Moniteur de l'Enseignement Supérieur – Allocataire de Recherche.

Formation

- Habilitation à Diriger des Recherches (2008) ;
- Doctorat en Informatique (2000) ;
- DEA en Informatique (1996) ;
- Ingénieur en Informatique (1996) - ESIL (Univ. Aix-Marseille II) ;
- Licence de Mathématiques (1993).

Responsabilités

Administratives

- Membre du Conseil des Etudes et de la Vie Universitaire (CEVU) de l'Université d'Avignon (2006-)
- Membre du conseil d'administration de l'IUP GMI (2005-)
- Membre du conseil de perfectionnement de l'IUP GMI (2006-)
- Membre du conseil de la documentation de la Bibliothèque Universitaire de l'Université d'Avignon (2006-)

Enseignement

- Responsable 2009- de la mention Informatique du Master Science & Technologie de l'Université d'Avignon
- Responsable 2004-2008 du Master Informatique TAIM (Traitement Automatique de l'Information Multimédia) : Parcours Recherche et Professionnels
- Responsable 2002-2004 du DESS Traitement Automatique de l'Information sur Internet (TAII)
- Responsable 2000-2004 de l'option "Génie Logiciel", 3è année IUP GMI

Recherche

- 2004-2008 : Membre de la commission de spécialistes mixte Informatique et Linguistique (sections 7, 27 et 61) de l'Université d'Avignon
- 2008- : Membre de la commission de sélection en Informatique MCF de l'Université Aix-Marseille II
- 2008- : Membre du comité directionnel et éditorial des Editions de l'Université d'Avignon

Activités d'encadrement

Thèses en cours (2)

- 75 % : Thierry Waszak (*Méthodes d'apprentissage de modèles numériques multi-niveaux et acquisition automatique de lexiques pour l'aide à la catégorisation de textes*), co-encadrée avec M. El-Bèze (LIA), convention CIFRE avec la société SYLLABS depuis février 2008 ;
- 50 % : Rémi Lavalley (*Identification de syntagmes discriminants pour la classification automatique de textes et la détection de nouveauté. Application à la gestion de la relation client et en particulier à l'analyse de réponses à des questions ouvertes d'enquêtes de satisfaction*), co-encadrée avec M. El-Bèze (LIA), soumission CIFRE en cours avec EDF R&D : démarrage prévu en janvier 2009.

Thèses soutenues (3)

- 75 % : Laurianne Sitbon (*Robustesse des méthodes symboliques et numériques en recherche d'informations pour l'assistance de personnes handicapées*) co-encadrée avec P. Blache (LPL, CNRS Aix-en-Provence), financement BDI CNRS, soutenue en novembre 2007, actuellement post-doctorante en Australie ;
- 50 % : Laurent Gillard (*Quelles méthodes pour les systèmes de Questions/Réponses ? Une avancée vers le tout numérique*) co-encadrée avec M. El-Bèze (LIA), financement : allocataire de recherche, soutenue en octobre 2007, actuellement en post-doc au CEA ;
- 50 % : Benoît Favre (*Résumé automatique de parole pour un accès efficace aux bases de données audio*) co-encadrée avec J.-F. Bonastre (LIA), financement CIFRE avec Thalès, soutenue en mars 2007, actuellement en post-doc aux Etats-Unis.

Encadrement de Masters Recherche et DEA Informatique (8)

- T. Waszak (Détection automatique de citations) – 2007
- I. Temou (Recherche d'information et ontologie) en collaboration avec l'INRA - 2007
- N. Flavier (Similarités entre requêtes en langage naturel) - 2006

– L. Sitbon (Méthodes de segmentation thématique) - 2004

Co-encadrement à 50 % des Master Recherche / DEA de :
– R. Lavalley (Classification automatique – Acquisition terminologique) - 2008
– B. Favre (Moteur de recherche multimédia) - 2003
– L. Gillard (Indexation de documents) - 2002
– C. Raymond (Enrichissement de requêtes) - 2001

Coopération internationale

Série 4 fois 3 heures de cours/séminaires à l'Université de DaLat (Vietnam) autour du thème *"Natural Language Processing - From Text Mining to Automatic Summarization"* en avril 2008.

Projets ANR et autres

Responsable scientifique pour le LIA des projets
– ANR : PIITHIE (RNTL 2006 pour la période 2007-2009) : **Détection de plagiats et suivi informationnel** (avec Univ. Nantes LINA, Sinequa, Advestigo, Syllabs)
– Technolangue : **EQUER (évaluation en questions-réponses)** et **OURAL (segmentation thématique de documents)** (2003-2005)
– Projet Ontofruit 2006-2009 **(Recherche d'informations et ontologies pour le domaine de l'écophysiologie végétale)** avec l'INRA d'Avignon : encadrement de 3 stages de Master Informatique, publication en cours pour la revue « Document numérique », prototype fonctionnel en phase de test.
– RIP-WEB (2003-2005) (Recherche d'informations précises sur le Web) : projet collaboratif initié par B. Grau (LIMSI-CNRS) ayant donné lieu à la parution d'un ouvrage collectif publié chez Hermès-Sciences.

Autres projets :
– Projet ENCORE, 2007-2008 **(Recherche d'informations spécialisées en chimie organique)** avec le laboratoire de Chimie Organique de Synthèse de l'Université de Namur (Belgique).

Jurys de thèses

– Membre des jurys de thèse de doctorat en informatique :
  – Mehdi Embarek, "*Un système de questions-réponses dans le domaine médical*" (Université Marne-la-Vallée, CEA - sous la direction de C. Fluhr et de O. Ferret) - juillet 2008 ;
  – Samir Abdou, « *Recherche d'Information Plurilingue* » (Université de Neuchâtel, Suisse – sous la direction de J. Savoy) – juin 2007 ;
  – Amélie Imafouo, « *Etude de l'influence du passage à l'échelle sur les modèles de recherche d'information* » (Ecole des Mines de Saint-Etienne et Université Jean Monnet, sous la direction de M. Beigbeder) ;
  – Laura Perret, « *Extraction automatique d'information : génération de résumé et question-réponse* » (Université de Neuchâtel, Suisse – sous la direction de J. Savoy) – mars 2005.
– Membre du jury de thèse de doctorat en linguistique :
  – Alain Régnier, « *Analyse et représentation formelle du discours pour la classification automatique des textes* » (Université de Provence, LPL – sous la direction de P. Blache) – décembre 2007.

Autres activités

- **Président des comités de programme des conférences** « Jeunes chercheurs » RECITAL 2008 et RJCRI 2008 (sponsorisée par ACM) ;
- **Membre des comités de programme des conférences** RIAO 2007, CORIA 2009, 2008, 2007, 2006, 2005, TALN 2008, 2007, 2005, atelier Q&A TALN 2004, RECITAL 2004 ;
- **Relectures** pour les revues *Traitement Automatique des Langues (TAL)* dont numéros spéciaux en Questions-Réponses et Traitement Automatique des Langues et Handicaps, *Pattern Recognition Letters* (PATREC), *Information Retrieval* (Kluwer), *Information Processing & Management* (IPM) et *IEEE Transactions on Knowledge and Data Engineering* ;
- **Expertises** ANR : Technologies Logicielles (2006), Masse de données (2007), Programme Blanc (2008), Contenus et Interactions (2008)
- **Membre des comités d'organisation** de la conférence internationale RIAO 2004 et de la conférence francophone JEP-TALN 2008
- Participation aux **campagnes d'évaluation** :
  - en recherche documentaire et recherche d'informations : TREC-Enterprise 2008, TREC ad-hoc 1998, Amaryllis 1996 et 1999 ;
  - en questions-réponses : NIST TREC 2002 (Anglais) , Technolangue EQUER 2004 (Français), CLEF 2006 (Français et Anglais) ;
  - en résumé automatique : DUC 2006 (Anglais) ;
  - en segmentation automatique : DEFT 2005 (Français).
- Membre des associations scientifiques ACM (*Association for Computing Machinery*), ARIA et ATALA ;
- **Vulgarisation** : conférence invitée dans le cadre du Café des Sciences d'Avignon sur les Handicaps en février 2008 sous le titre : « *Modèles cognitifs et modèles informatiques pour le traitement automatique des langues* ».

13.1 Publications personnelles

Publications en français dans des Revues de rang A avec comité et audience internationaux

1. L. Sitbon, P. Bellot, P. Blache, "Éléments pour adapter les systèmes de recherche d'information aux dyslexiques", Traitement Automatique des Langues (TAL), vol. 48-2, 2008
2. P. Zweigenbaum, B. Grau, A.-L. Ligozat, I. Robba, S. Rosset, X. Tannier, A. Vilnat (LIMSI) & P. Bellot (Univ. Avignon), "Apports de la linguistique dans les systèmes de recherche d'informations précises", RFLA (Revue Française de Linguistique Appliquée),XIII (1), 2008.
3. Laurent Gillard, Laurianne Sitbon, Patrice Bellot, Marc El-Bèze, "Dernières évolutions de SQuALIA, le système de Questions/Réponses du LIA", 2006 Traitement Automatique des Langues (TAL), vol. 46, num. 3, Hermès.
4. P. Bellot, M. El-Bèze, « Classification locale non supervisée pour la recherche documentaire », Traitement Automatique des Langues (TAL), vol. 42, num.2, Hermès, p. 335 à 366, 2001.
5. P. Bellot, M. El-Bèze, « Classification et segmentation de textes par arbres de décision », Technique et Science Informatiques (TSI), Editions Hermès, volume 20, num. 3, p. 397 à 424, 2001.

Autres revues et LNCS

1. Laurent Gillard , Laurianne Sitbon , Eric Blaudez , Patrice Bellot, Marc El-Bèze, « Relevance Measures for Question Answering, The LIA at QA@CLEF-2006 », Lecture Notes in Computer Science,4730/2007, « Evaluation of Multilingual and Multi-modal Information Retrieval », p. 440 à 449, 2007.
2. P.-F. Marteau, C. De Loupy, P. Bellot, M. El-Bèze, « Le Traitement Automatique du Langage Naturel, Outil d'Assistance à la Fonction d'Intelligence Economique », Systèmes et Sécurité, Vol. 5, num.4, p. 8-41, 1999.
3. P. Bellot, M. El-Bèze, « Query length, number of classes and routes through clusters : experiments with a clustering method for information retrieval », Lecture Notes in Computer Science (LNCS 1746), Springer-Verlag, IEEE Int. Conf. Comp. Science, Hong-Kong, pp. 196-205, 1999.

Chapitres de livres

1. P. Bellot, M. Boughanem, "Recherche d'information et systèmes de questions-réponses", 2008 in " La recherche d'informations précises : traitement automatique de la langue, apprentissage et connaissances pour les systèmes de question-réponse (Traité IC2, série Informatique et systèmes d'information)", sous la direction de B.Grau, Hermès-Lavoisier, chapitre 1, p. 5-35
2. Patrice Bellot, "Classification de documents et enrichissement de requêtes", 2004 Méthodes avancées pour les systèmes de recherche d'informations (Traité des sciences et techniques de l'information) sous la dir. de IHADJADENE M., chapitre 4, p.73 à 96, Hermès
3. J.-C. Meilland, P. Bellot, "Extraction automatique de terminologie à partir de libellés textuels courts", 2005 in "La Linguistique de corpus" sous la direction de G. Williams, Presses Universitaires de Rennes, p. 357 à 370, 2005

Conférences internationales avec comité de lecture

1. Laurianne Sitbon, Patrice Bellot, « A readability measure for an information retrieval process adapted to dyslexics » Second international workshop on Adaptive Information Retrieval (AIR 2008) (in conjunction with IIiX 2008), octobre 2008.
2. Laurianne Sitbon, Patrice Bellot, « How to cope with questions typed by dyslexic users », second ACM workshop on Analytics for noisy unstructured text data (AND at SIGIR 2008), ACM, Singapour, 2008.
3. Laurianne Sitbon, Patrice Bellot, Philippe Blache, "Evaluation of lexical resources and semantic networks on a corpus of mental associations", 6th edition of the Language Resources and Evaluation Conference (LREC 2008), Marrakech (Maroc), mai 2008.
4. Laurianne Sitbon, Patrice Bellot, Philippe Blache, "A corpus of real-life questions for evaluating robustness of QA systems", 6th edition of the Language Resources and Evaluation Conference (LREC 2008), Marrakech (Maroc), mai 2008.
5. Laurianne Sitbon, Patrice Bellot, Philippe Blache, "Phonetic based sentence level rewriting of questions typed by dyslexic spellers in an information retrieval context", Interspeech 2007, Anvers (Belgique), 2007.
6. Benoît Favre, Jean-François Bonastre, Patrice Bellot, "An Interactive Timeline for Speech Database Browsing", Interspeech 2007, Anvers (Belgique), 2007

7. Laurianne Sitbon, Patrice Bellot, "Topic segmentation using weighted lexical links (WLL)", ACM SIGIR 07, ACM Press, Amsterdam (Pays-Bas), 2007 (papier court, session poster)
8. Laurent Gillard, Patrice Bellot, Marc El-Bèze, "Question Answering Evaluation Survey", actes de la 5ième conférence Language Resources and Evaluation Conference (LREC), Gênes (Italie), 24-26 mai 2006.
9. Laurianne Sitbon, Patrice Bellot, "Tools and methods for topic segmentation of texts and contextual evaluation", Fifth International Conference on Language Resources and Evaluation (LREC 2006), Italie, 2006.
10. L. Sitbon, P. Bellot, "Adapting and comparing linear segmentation methods for french", actes de la 7è conférence RIAO, Avignon, France, p.623 à 637 ; 2004
11. K. Lavenus, J. Grivolla, L. Gillard, P. Bellot, "Question-answer matching : two complementary methods", actes de la 7ième conférence en Recherche d'Information Assistée par Ordinateur (RIAO), Avignon (France), 26-28 avril 2004, pages 244 à 259
12. Benoît Favre, Patrice Bellot, Jean-François Bonastre, "Information retrieval on mixed written and spoken documents", actes de la 7è conférence RIAO, Avignon, France, p. 826 à 835, 2004
13. P. Bellot, M. El-Bèze, « Clustering by means of decision trees without learning or hierarchical and K-Means like algorithms », actes de RIAO'2000, Paris, p. 344-363, 2000.
14. C. De Loupy, P. Bellot, « Evaluation of Document Retrieval Systems and Query Difficulty », Actes du LREC'2000 Satelitte Workshop : "Using Evaluation within HLT Programs", Athènes, 2000.

Conférences internationales sans comité de lecture

1. Eric San Juan, Patrice Bellot, "The LIA at TREC-Enterprise 2008", Text REtrieval Conference, NIST Special publication, 2008
2. Benoît Favre, Frédéric Béchet, Patrice Bellot, Florian Boudin, Marc El-Bèze, Laurent Gillard, Guy Lapalme, Juan-Manuel Torres-Moreno, "The LIA-Thales summarization system at DUC-2006", Actes du workshop Document Understanding Conference (DUC-2006) durant HLT-NAACL'06, New York (USA), 8-9 juin 2006.
3. P. Bellot, E. Crestan, M. El-Bèze, L. Gillard, C. de Loupy, « Coupling Named Entity Recognition, Vector-Space Model and Knowledge Bases for TREC-11 Question-Answering Track », actes de 11th Text REtrieval Conference, NIST Special publication 500-251, 2003.
4. C. De Loupy, P. Bellot, M. El-Bèze, P.-F. Marteau, « Query Expansion and Automatic Classification », actes de 7th Text REtrieval Conference, NIST Special Publication 500-242, p. 443-450, 1999.

Conférence invitée

1. P. Bellot, "Traitement automatique des langues et classification automatique : méthodes et applications pour la recherche d'informations", RIAs 2006, Lyon, mars 2006

Conférences francophones avec comité de lecture

1. Poulard Fabien, Waszak Thierry, Hernandez Nicolas , Bellot Patrice, « Repérage de citations, classification des styles de discours et identification des constituants citationnels en écrits journalistiques », TALN 2008, Avignon, juin 2008.

2. Laurent Gillard, Patrice Bellot, Marc El-Bèze, « Quelles combinaisons de scores et de critères numériques pour un système de Questions/Réponses ? », TALN 2008, Avignon, juin 2008.
3. Laurianne Sitbon, Patrice Bellot, Philippe Blache, "Lisibilité et recherche d'information : vers une meilleure accessibilité", 5è Conférence en Recherche d'Informations et Applications (CORIA) soutenue par l'ACM, Trégastel (France), mars 2008 **(NB : cette publication a reçu le prix « Jeune Chercheur »)**.
4. Laurent Gillard, Patrice Bellot, Marc El-Bèze, "D'une compacité positionnelle à une compacité probabiliste pour un système de Questions/Réponses", 4è Conférence en Recherche d'Informations et Applications (CORIA) soutenue par l'ACM, Saint-Etienne (France), mars 2007
5. Laurianne Sitbon, Patrice Bellot, Philippe Blache, "Traitements phrastiques phonétiques pour la réécriture de phrases dysorthographiées", Actes de TALN 2007, Toulouse, 2007.
6. Laurent Gillard, Patrice Bellot, Marc El-Bèze, "Analyse des échecs d'une méthode pour traiter les questions définitoires soumises à un système de Questions/Réponses", actes de TALN, Toulouse (France), 2007.
7. Nicolas Flavier, Patrice Bellot, "Vers un appariement automatique de questions extraites de courriers électroniques", Conférence Francophone sur l'Apprentissage Automatique (CAp 2007), Grenoble (France), 2007.
8. L. Sitbon, J. Grivolla, L. Gillard, P. Bellot, P. Blache, "Vers une prédiction automatique de la difficulté d'une question en langue naturelle", 13ième conférence Traitement Automatique des Langues Naturelles (TALN), Louvain (Belgique), 10-13 avril 2006, pages 337 à 346.
9. Laurent Gillard, Patrice Bellot, Marc El-Bèze, "Influence de mesures de densité pour la recherche de passages et l'extraction de réponses dans un système de questions-réponses", actes de la 3ième Conférence en Recherche d Informations et Applications (CORIA) soutenue par l'ACM, Lyon (France), 15-17 mars 2006, pages 193-204.
10. Laurent Gillard, Patrice Bellot, Marc El-Bèze, "Questions Booléennes : Oui ou Non, des Questions et des Réponses", actes de la 13ième conférence Traitement Automatique des Langues Naturelles (TALN), Louvain (Belgique), 10-13 avril 2006, pages 159 à 166.
11. Benoît Favre, Jean-François Bonastre, Patrice Bellot, François Capman, "Accès aux connaissances orales par le résumé automatique", 6è journées francophones "Extraction et Gestion des Connaissances" EGC 2006, Lille (France), janvier 2006.
12. L. Sitbon, P. Bellot, "Segmentation thématique par chaînes lexicales pondérées", Actes de TALN 2005, Dourdan, France, 2005.
13. Benoît Favre, Jean-François Bonastre, Patrice Bellot, "Recherche d'information dans un mélange de documents écrits et parlés", Journées d'Etude de la Parole, Fèz (Maroc), 2004.
14. L. Sitbon, P. Bellot, "Evaluation de méthodes de segmentation thématique linéaire non supervisées après adaptation au français ", Actes de la conférence TALN, Fez (Maroc), p. 441 à 450, avril 2004.
15. K. Lavenus, J. Grivolla, L. Gillard, P. Bellot, "Deux pistes complémentaires pour améliorer l'appariement Question Réponse", 11è conférence TALN, Fez (Maroc), p. 403 à 412, 2004.
16. L. Gillard, P. Bellot, M. El-Bèze, « Bases de connaissances pour asseoir la crédibilité des réponses d'un prototype de question réponse », actes de la conférence Traitement Automatique des Langues Naturelles, Nantes, 2003
17. C. Raymond, P. Bellot, M. El-Bèze, « Enrichissement de requêtes pour la recherche documentaire selon une classification non-supervisée », 13ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et d'Intelligence Artificielle (RFIA'2002) - Angers, volume 2, p. 625 à 632, 2002.

18. J.-C. Meilland, P. Bellot, « Extraction automatique de terminologie - Application à des libellés courts issus de la grande distribution » , 2è Journées "Linguistique de Corpus" - Lorient - Septembre 2002
19. P. Bellot , « Structuration dynamique de textes pour la recherche documentaire », Ecole thématique "Nouveaux défis en science de l'information" - GDR I3 - Marseille, septembre 2000.
20. P. Bellot, M. El-Bèze, « Un Algorithme de Segmentation Automatique de Corpus - Le Système S.I.A.C. », Premières Journées Scientifiques et Techniques (JST 97), p. 113-117, Avignon, 1997.

Conférences nationales sans comité de lecture

1. Laurent Gillard, Patrice Bellot, Marc El-Bèze, "Le LIA à EqueR (campagne Technolangue des systèmes Questions-réponses)", actes de l'atelier Evaluation en Question Réponse (EQueR) de la 12è conférence Traitement Automatique des Langues Naturelles (TALN), volume 2, Dourdan (France), 6-10 juin 2005, pages 81 à 84.
2. P. Bellot, C. De Loupy, « SIAC et IndeXal à Amaryllis'99 », Atelier de la deuxième campagne Amaryllis (AUF), 2000, Paris.
3. P. Bellot, M. El-Bèze, « Description du Système I.A.C.S. », Première campagne d'évaluation Amaryllis (AUPELF-UREF), Avignon, 1997.

Séminaires et ateliers francophones... sans acte papier

1. Aventurier, P. ; Leiser, H. ; Richard, H. ; Bellot, P., OntoFruit : Ecophysiologie végétale de l'arboriculture fruitière –un référentiel documentaire indexé par une ontologie du domaine, Séminaire Texte et Connaissance, INRA, Paris, 2008
2. P. Bellot, Les moteurs de recherche sur Internet, Journée de rencontres Enseignants-Chercheurs, Rectorat de l'académie Aix-Marseille, mars 2004
3. P. Bellot, Quelques modèles probabilistes et statistiques en recherche d'informations. Application aux moteurs Questions-Réponses, Journée d'Etude RIP-WEB, Paris, décembre 2003
4. E. Crestan, L. Gillard, M. El-Bèze, P. Bellot, C. de Loupy, « Entités nommées pour les systèmes de question/réponse », Journée ATALA "Des requêtes aux questions : nouvelle perspective pour la recherche d'information ?", Paris, 17 mai 2003.
5. P. Bellot, « Méthodes de classification et de segmentation pour la recherche documentaire », Workshop "Fouille de textes", GDRI3, Ecole Polytechnique, Paris, 1999.

PBellot

Index

BM25, 152

ciQA, 50
CLEF, 22, 33, 43, 45, 120
Compacité, 40
Croft (expansion), 156
CWS, 158

densité Δ, 36
DUC, 67, 87
dyslexie, 92, 94, 98, 104, 105
dysorthographie, 107

EQUER, 22, 23, 31, 38, 40, 42, 43, 50, 108, 110

F-mesure, 158

graphème, 95, 99, 102, 106, 111, 113, 119, 138

K-mesure, 159

Mean Reciprocal Document Rank, voir MRDR
MMR, 84, 85
Modèle 2-Poisson, 150
MRDR, 160
MRR, 35, 38, 158

okapi, 151

phonème, 95, 97, 99, 101, 104, 112, 113, 119, 138
Poids (chaîne lexicale), 55
précision , 158
précision moyenne, 158

rappel, 157
Rocchio (expansion), 155

SDR, 81

TREC, 12, 18, 20, 21, 27, 32, 34, 38, 42, 45, 50, 59, 81, 118, 121, 127
TREC-Enterprise, 39

WindowDiff, 56

Abdou, S. 189
Abney, S. 211
Aiken, Alex 211
Al-Onaizan, Y. 189
Allan, J. 189, 203, 210
Alves, E. 196
Amati, Gianni 189
Amento, Brian 214
Amini, Massih 213
An, J. 204
Anden, F. 202
Arguin, Martin 198
Ashby, J. 189
Aslam, Javed A. 190, 206
Aubergé, V. 215
Aubin, S. 190
Ault, T. 215
Aussenac-Gilles, N. 190, 193
Auzanne, C.G.P. 199
Aventurier, P. 190
Ayache, C. 190

Bacchiani, M. 211
Bacchiani, Michiel 214
Bachimont, Bruno 190
Badulescu, A. 206
Baecker, Ron 196, 207
Baeza-Yates, R.A. 190
Bagein, M. 215
Bai, S. 194
Bailly, G. 215
Bakker, E.M. 190
Bakshi, K. 204
Balakrishnan, Ravin 209
Ballesteros, L. 189
Banko, M. 193
Barnes, M.A. 211
Barry, C. 192
Baziz, M. 190
Beckwith, R. 206
Beeferman, D. 190
Belew, R. K. 190
Belkin, Nicholas J. 191
Belkin, Nick 202
Bellegarda, Jerome R. 191
Bellengier, Emmanuel 191
Bellot, P. 190, 191, 196, 197, 199, 200, 205, 212, 215
Bellot, Patrice 191, 196, 198–200, 203, 204, 209, 211, 212
Benamara, F. 191
Bentin, S. 198
Beretta, A. 191
Berger, A. 190
Berrut, Catherine 191
Berry, M. 192
Besançon, R. 192
Besner, D. 195
Biébow, B. 190
Blache, Philippe 191, 192, 212
Blair-Goldensohn, Sasha 209
Blaudez, Eric 200
Bleasdale, F.A. 192
Bodner, G.E. 192
Boissière, Philippe 192
Bolohan, O. 206
Bonastre, J.-F. 197, 199, 202, 207
Bonin, P. 192
Bookstein 192
Boudin, F. 197
Bouffier, A. 206
Boughanem, M. 190, 192, 193
Boughanem, Mohand 191
Boula de Mareil, P. 215
Bourigault, D. 193
Brill, E. 193
Brill, Eric 193
Brown, E. 209
Brown, E. W. 193
Brown, G.D.A. 193
Brown, Gordon D. A. 193
Bruandet, M.-F. 193
Bruner, J. 193
Brunswick, N. 208
Bruza, P. 193
Brysbaert, M. 207
Buckley, C. 193, 211, 215
Buckley, Chris 210
Bunescu, R. 201
Burger, J.D. 193
Burges, C.J. 193
Byrd, D. 189
Béchet, F. 197, 207, 215
Béchet, Frédéric 194

C., Jacquemin 193
Caelen, J. 194
Cai, Rui 194
Callan, J. 200, 207

Callan, James P. 189, 194, 215
Calvé, A.L. 211
Campbell, Nick 201
Campolini, Claire 194
Cao, G. 199
Capman, F. 197
Cappa, S.F. 208
Carbonell, J. 200
Cardie, C. 213
Carletta, J. 207
Carmel, D. 194
Carpineto, C. 189
Castells, Pablo 194
Cha, J. 204
Chaffee, J. 199
Chalard, M. 192
Chang, Y. 194
Changeux, J.P. 194
Chanoine, V. 208
Chappelier, J.-C. 192
Charlet, J. 193
Chatterjee, Anjan 212
Chaudiron, S. 194
Chaumette, Cédric 214
Chen, F. 203
Chevalier, M. 194
Chevallet, J.-P. 193, 194, 200
Chidlovskii, Boris 194
Chignell, Mark 209
Cho, B. H. 204
Choi, F.Y.Y. 194
Choi, John 214
Chomsky, Noam 194, 195, 208
Chong, H. A. 193
Choukri, K. 199
Chrisment, C. 201
Chua, Tat-Seng 195
Ciura, Marcin G. 196
Clark, C. 206
Clarke, C. L. A. 195
Clarke, Charles L. A. 195
Claveau, Vincent 206
Clough, P.D. 195
Collier, Rem 204
Collins, M. 211
Coltheart, M. 195, 209
Coltheart, V. 195
Colé, P. 205, 213
Coppola, B. 203
Cormack, G. V. 195
Courtois, Fabienne 214
Crestan, E. 191
Croft, W. Bruce 189, 195, 205, 208, 215
Cronen-Townsend, S. 195
Cui, Hang 195
Cutting, D. 195
Czuba, K. 209

Daille, Brigitte 195
d'Alessandro, C. 215
Dang, H. 195
Danon-Boileau, Laurent 195
Dave, Kushal 196
Davelaar, E. 195
de Cormis, C. 209
de Loupy, C. 191, 196, 205
De Luca, Maria 215
De Mori, Renato 200, 203
Deerwester, S.C. 196
Degerstedt, L. 202
Dehaene, S. 196
Delbecque, T. 205
Demonet, J.F. 208
Dennis, S. 193
Denos, Nathalie 191
Deorowicz, Sebastian 196
Desclès, Jean-Pierre 207
Detweiler, S.R. 197
Deutsch, G.K. 213
Di Cristo, Philippe 214
Di Pace, Enrico 215
Dias, G. 196
Dickinson, Anna 196
Diehl, C.-P. 199
Donovan, R. E. 196
Dours, Daniel 192
Dowdall, James 209
Draffan, E.A. 202
Du, Y. 215
Ducrot, S. 196
Ducrot, Stéphanie 205
Dufour, Christine 196
Dulucq, S. 196
Dumais, S. 193, 196
Dunnion, John 204
Dupoux, E. 205
Dutoit, T. 196
Déjerine, J. 196

E., Folser-Lussier 199
Echihabi, Abdessamad 197
Edelman, G.E. 197
Eglin, V. 194, 197
Egret, D. 201
El-Bèze, M. 191, 196, 197, 199, 200, 205
El-Bèze, Marc 199, 200, 209, 213
Ellis, A.W. 207
Embarek, Mehdi 197
Emptoz, H. 197

Fairweather, P.G. 197
Fan, W. 209
Fan, Weiguo 209
Favre, B. 197
Fazio, F. 208
Fellbaum, C. 206
Feng, F. 205
Fernandes, Aaron 213
Fernandez, Miriam 194
Ferrand, Ludovic 197, 200, 207
Ferrandez, A. 213
Ferret, O. 198
Fiorentino, R. 191
Fiset, Daniel 198
Fiset, Stéphanie 198
Flavier, Nicolas 198
Flesch, R. 198
Flohic 205
Fluhr, C. 198
Flycht-Eriksson, A 202
Foote, J. T. 202
Foote, Jonathan 198
Foucault, M. 198
Foukia, S. 215
Fournier, R. 206
Fox, E. 198, 210
Frank, E. 214
Franz, M. 202
Fredouille, C. 198, 202, 207
Friedman, N. 198
Friedman, R.B. 205
Frith, C.D. 208
Frith, U. 198, 208
Frost, R. 198
Fuhr, N. 198
Furnas, G.W. 196

Gabbay, D.M. 203
Gabrieli, J.D. 213
Gaizauskas, R. 195
Galan, J.B. 211
Galley, M. 199
Galliano, S. 199
Gallinari, Patrick 208, 213
Gao, J. 199
Garay, Michael R. 199
Garofolo, J.S. 199
Gasperini, Filippo 215
Gatford, M. 210
Gauch, S. 199
Gaussier, E. 199
Geoffrois, E. 199
Gernsbacher, M.A. 199
Getoor, L. 199
Gibson, D. 199
Gillard, L. 191, 197, 199, 200
Gillard, Laurent 199, 200, 203, 204, 212
Girju, R. 201
Girju, S. 206
Goldman, J. F. 215
Goldstein, J. 200
Goldzsmidt, M. 198
Gong, Y. 200
Graesser, A.C. 200
Grainger, J. 200
Grau, B. 190, 198, 200, 215
Gravier, G. 199
Grefenstette, G. 200
Gregor, Peter 196
Greiif, Warren R. 200
Grewal, A. 209
Grivolla, Jens 200, 203, 204, 212
Grodzinsky, Y. 200
Gross, D. 206
Guénot, M.-L. 200

Habert, B. 201
Habib, Michel 201, 208
Hamon, T. 190, 201
Hancock-Beaulieu, M. 210
Hanson, V.L. 197
Harabagiu, S. 201, 206, 207
Harm, M.W. 201
Harman, Donna K. 201, 214
Harshman, R.A. 196

Harth, E. 201
Hearst, M.A. 201, 208
Hermjakob, U. 201
Hernandez, Nathalie 201
Hernandez, Nicolas 209
Herrera, J. 201
Hersh, W. 201
Hess, Michael 206, 209
Higuchi, Fumito 201
Hindle, D. 211
Hindle, Don 214
Hino, Y. 208
Hinton, G.E. 210
Hirschberg, Julia 214
Hollard, S. 194
Hovy, E. 201
Huang, X. 215
Hubert, J. 200
Hurault-Plantet, M. 198
Huynh, D. 204

Ihadjadene, M. 201
Iida, Akemi 201
Illouz, G. 198
Ingwersen, Peter 202
Isenhour, Philip 214
Istrate, D. 202
Ittycheriah, A. 202

Jackiewicz, Agata 202
Jacobs, A.M. 209, 215
Jacquemart, P. 202
Jacquemin, C. 199, 202
Jakubowicz, C. 202
James, Abi 202
Jansen, B.J. 202
Jenner, B. 202
Jeong, K. 214
Jin, H. 189
Jing, H. 199
Johnson, David S. 199
Johnson, S.E. 202
Jolion, J.M. 202
Jonasson, J.T. 195
Jones, G. J. F. 202
Jones, rck 202
Jonsson, A 202
Jourlin, Pierre 200, 202
Judica, Anna 215
Julien, C. 194
Jumel, B. 203
Jung, H. 204

Kaljurand, K. 209
Kan, M. Y. 203
Kan, Min-Yen 195
Kandel, L. 203
Karger, D. R. 204
Karlsson, M. 209
Kaszkiel, Marcin 203
Katz, B. 204
Katz, Boris 213
Katz, L. 198
Keller, E. 215
Kelly, D. 195, 203
Kemkes, G. 195
Kempson, R. 203
Kim, D. 204
Kisman, D. I. E. 195
Klavans, J.L. 203
Kleinberg, J. 199
Knight, K. 189
Kobler, J. 202, 203
Kosseim, L. 208
Kouylekov, M. 203
Kraaij, W. 192
Kremer, P. 204
Kroner, G.K. 202
Kuhn, R. 203
Kumaran, G. 203
Kupiec, J. 203
Kwak, B. K. 204
Kwok, K. L. 203

Labadié, A. 203
Lacutusu, F. 206
Lafferty, J. 190
Lalmas, M. 203
Lam, K. 215
Landauer, T.K. 196
Langdon, R. 195
Lapalme, G. 197, 208
Lapalme, Guy 204
Laszlo, M. 195
Lattimer, C.W. 215
Lavenus, Karine 203, 204
Lavrenko, V. 189
Lawrence, Steve 196
Lee, C. 204
Lee, G. G. 204

Lee, Joon Ho 204
Lee, M.H. 207
Lee, S. 204
Lee, W.S. 215
Leek, T. 204
Lehnert, W. 204
Leiser, H. 190
Lespinasse, K. 204
Lew, Michael S. 190
Lewis, Jonathan 196
L'Homme, M.C. 193
Li, Keya 195
Li, X. 204
Liberman, M.Y. 211
Ligozat, M.-L. 215
Lillis, David 204
Lin, C. Y. 201
Lin, J. 193, 195, 203, 204
Lin, Jimmy 204, 213
Linares, G. 198, 207
Litkowski, K. C. 204
Liu, X. 200
Liu, Xiaoyong 205
Llopis, F. 213
Locker, L. 215
Loosemore, Richard P. W. 205
Lu, Lie 194
Lucas, Emmanuel Giguet 205
Lupker, S.J. 208
Lynam, T. R. 195
Lytinen, S. 205, 213
Lété, B. 205, 208
Lété, Bernard 205
M., El-Bèze 200
Maarek, Y. 194
Maedche, A. 213
Magnini, B. 203
Maiorano, S. 206
Malaisé, V. 205
Mandelbrod, M. 194
Manmatha, R. 205
Marchand, Y. 205
Marcu, Daniel 197
Marteau, P.-F. 196, 205
Marton, G. 205
Marton, Gregory 213
Mass, Y. 194
Masson, M.E.J. 192
Massonié, D. 207
Matos, R. 207
Matrouf, D. 198, 207
Maybury, M. T. 205
Mc Arthur, R. 193
Mc Crory, E. 208
McClelland, J.L. 211
McKenzie, P. 202
McKeown, K. 199, 203
Mehler, J. 205
Meignier, S. 207
Meir, R. 205
Merkel, M. 202
Merzenich, M.M. 213
Messerschmitt 205
Meunier, J.G. 213
Meyer-Viol, W. 203
Mihalcea, R. 201
Miller, G.A. 205, 206
Miller, K. 206
Miller, S.L. 213
Minel, Jean-Luc 206
Minsky, Marvin 206
Mitra, M. 193, 211
Mittal, V. 200
Mizzaro, S. 206
Moldovan, D. 201, 206
Moles, A. 203
Molla, Diego 206, 209
Monaghan, Padraic 206
Monceaux, L. 198
Mondary, T. 206
Montague, Mark 190, 206
Monz, C. 206
Moore, Robert C. 193, 213
Morarescu, P. 201, 206
Moreau, Fabienne 206
Morgan, William T. 200
Moriceau, Véronique 206
Morin, Emmanuel 195
Morrison, C.M. 207
Morton, J. 207
Mostefa, D. 199
Mothe, Josiane 201, 207
Mourad, Ghassan 207
Munteanu, Cosmin 207
Murray, G. 207
Myaeng, S.H. 207
Méot, A. 192

Nadine 205

Nazarenko, A. 201, 206
Negri, M. 203
New, Boris 197, 207
Newell, Alan F. 196
Ng, A. 193
Nguyen, N. 196
Nie, J. Y. 192, 199
Nietzsche, F. 207
Nobata, C. 211
Nocera, P. 197, 198
Nocéra, P. 207
Norberg, S. 202
Novischi, A. 206
Nyberg, E. 207

Ogilvie, Paul 207
Oh, H.-J. 207
O'Regan, J.K. 207
O'Shaughnessy, D. 215
Over, P. 201

Pagel, V. 215
Pallier, C. 207
Papert, Seymour 206
Pasca, M. 201, 206, 207
Paulesu, E. 208
Pedersen, J. 203
Pederson, J. 203
Pedler, Jennifer 208
Peereman, R. 208
Penas, A. 201
Penn, Gerald 207
Pennock, David M. 196
Perea, M. 208
Pereira, Alvaro R. 208
Pereira, F. 211
Pereira, Fernando 214
Perry, C. 195, 208
Person, N. 200
Pevzner, L. 208
Pexman, P.M. 208
Piaget, Jean 208
Piao, S.L. 195
Piwowarski, B. 208
Plamondon, L. 208
Poeppel, D. 191
Poibeau, Thierry 208
Polanyi, L. 208
Poldrack, R.A. 213
Pollock, J.-Y. 208
Ponte, Jay M. 200, 208
Popescu-Belis, A. 210
Poulard, Fabien 209
Prade, H. 190
Prager, J. 209
Prager, John 209
Pretschner, P. 199

Qi, H. 209
Qi, Hong 209
Quan, D. 204
Quinlan, J. 209

Radev, D. 209
Radev, D. R. 209
Radev, Dragomir R. 209
Raghavan, P. 199
Rajman, M. 192
Ranjan, Abhishek 209
Rasolofo, Y. 209
Rastle, Kathleen 195, 209
Rath, T. 205
Ratsch, G. 205
Rauzy, Stéphane 191, 192
Raymond, Christian 209
Rayner, K. 189
Renals, S. 207
Rey, A. 209
Rey, V. 209
Reynar, J.C. 209
Richard, H. 190
Rinaldi, Fabio 209
Rizzi, L. 209
Rizzolatti, G. 210
Robba, I. 198, 215
Robertson, S.E. 210
Robertson, Stephen 210
Rocchio, J.J. 210
Romano, G. 189
Romero, Y. 203
Rosa, E. 208
Rosenberg, Aaron 214
Rosenberg, C. R. 211
Rosset, S. 215
Rossignol, C. 210
Roth, D. 204
Roukos, S. 202
Rumelhart, D.E. 210
Rus, V. 201

Sabater, C. 209

Sabbah, G. 210
Saint Dizier, P. 191
Salton, G. 193, 210, 211, 215
Salton, Gerard 210
Samuelson, P. 210
Sannier, F. 215
Saracevic, T. 210
Savoy, J. 189, 209, 211
Scheffer, N. 202
Schibler, D. 204
Schleimer, Saul 211
Schmidt-Weigand, F. 209
Schmitt, L. 204
Schoning, U. 203
Schultz, J.M. 211
Schwartz, R. 204
Schwerdtfeger, R.S. 197
Schwitter, R. 206, 209
Scott, N.G. 211
Seidenberg, M.S. 201, 211
Sejnowski, T. J. 211
Sekine, S. 211
Senator, T.E. 211
Seo, J. 204
Shasha, D. 214, 215
Shaw, J.A. 198
Shillcock, Richard 206
Si, Luo 214
Simpson, G.B. 215
Singhal, A. 193, 211
Singhal, Amit 214
Sinha, V. 204
Sinigaglia, C. 210
Sista, S. 204
Sitbon, L. 203, 212
Sitbon, Laurianne 200, 211, 212
Smucker, Mark D. 204
Soffer, A. 194
Somasundaran, Swapna 212
Soulé-Dupuy, C. 192, 194
Southwood, M. Helen 212
Sp, K. 202
Sparck-Jones, K. 202, 210, 212
Spinelli, Donatella 215
Spooner, Roger 212
Sprenger-Charolles, L. 205, 208, 213
Spriet, Thierry 213
Staab, S. 213
Stark, Litza 214
Statman, R. 215
Stead, Larry 214
Stone, G.O. 215
Stoyanov, V. 213
Stoyanov, Veselin 212
Stroop, J.R. 213
Stéfanini, M.-H. 199
Sudo, K. 211
Sun, Renxu 195
Surdeanu, M. 201, 206
Sutcliffe, R.F.E. 213
Swan, R. 189
Szulman, S 190
Sébillot, Pascale 206

Tallal, P. 213
Tamine, L. 192
Tanenhaus, M.K. 211
Tanguy, L. 207
Tannier, X. 215
Taylor, Michael 210
Tebri, H. 193
Tellex, Stefanie 213
Temple, E. 213
Terra Egidio, L. 195
Tichit, L. 196
Tilker, P.L. 195
Tmar, M. 193
Tombros, Anastasios 203
Toms, Elaine G. 196
Tomuro, N. 205, 213
Toolan, Fergus 204
Toran, J. 202, 203
Torres-Moreno, J.-M. 197, 213
Toutanova, Kristina 213

Usunier, Nicolas 213

Vallet, David 194
Van Rijsbergen, Cornelis Joost 189
Velazquez-Morales, P. 213
Verdejo, F. 201
Vicedo, J. L. 213
Vilnat, A. 190, 198, 208, 215
Voorhees, E. M. 213, 214
Voorhees Ellen, M. 199, 214
Vrajitoru, D. 211
Véronis, J. 215
Véronis, Jean 214

W3C 214
Walker, S. 210
Wang, J.T.L. 214
Wang, Mengqiu 214
Waszak, Thierry 209, 214
Waters, G.S. 211
Watson, F.L. 193
Werbos, P.J. 214
Whittaker, Steve 214
Wiebe, Janyce 212
Wilkerson, Daniel S. 211
Wilkinson, Ross 214
Wilks, Y 195
Williams, R.J. 210
Wilson, Theresa 212
Wise, M. 214
Witten, I.H. 214
Woodland, P.C. 202
Wu, G. 199
Wu, H. 209, 210
Wu, L. 215

Xu, H. 194
Xu, J. 189, 215

Yang, Y. 215
Yasumura, Michiaki 201
Yates, M. 215
You, L. 215
Young, S. J. 202
Yu, C.T. 215
Yvon, F. 215

Zamchick, Gary 214
Zaragoza, Hugo 210
Zellner, B. 215
Zhang, D. 215
Zhang, Hong-Jiang 194
Zhang, K. 214, 215
Zhang, Y. 215
Zhang, Yuecheng 207
Zhang, Zhu 209
Zheng, Zhiping 209
Zhou, Y. 195, 215
Ziegler, Johannes C. 195, 208, 209, 215
Ziviani, Nivio 208
Zobel, Justin 203
Zoccolotti, Pierluigi 215
Zorzi, M. 208
Zweigenbaum, P. 198, 199, 201, 202, 205, 215

Liste des illustrations

1.1 Processus de catégorisation d'une question précise . . . 19
1.2 Des arbres de décision pour la catégorisation de questions . . . 22
1.3 Exemple de courrier électonique utilisé pour l'identification de questions . . . 26
1.4 Exemple de questions regroupées après calcul de similarité . . . 29
1.5 Illustration de la mesure de densité pour la segmentation . . . 37
1.6 Illustration de la mesure de densité : un score en fonction des occurrences les plus proches . . . 37
1.7 Architecture générale de notre système de questions-réponses SQuaLIA . . . 39
1.8 Progression des résultats de SQuaLIA entre 2002 et 2007 . . . 44
1.9 Le module d'annotations du projet Ontofruit . . . 48
1.10 Le module de recherche du projet Ontofruit . . . 49

2.1 Lissage des fréquences locales pour la segmentation . . . 55
2.2 Schéma de fonctionnement général de LIA_SEG . . . 56
2.3 Exemple d'arbre de décision utilisé pour l'expansion de requête . . . 63
2.4 Deux requêtes enrichies selon des arbres de décision non supervisés . . . 64
2.5 Evaluation de l'expansion sur les données Amaryllis . . . 65
2.6 Exemple de FSM pour l'identification de citations (couple source / relateur) . . . 70
2.7 Exemple de citations automatiquement détectées . . . 71
2.8 Exemple de texte source . . . 73
2.9 Exemple de texte cible proche . . . 74
2.10 Coloration du texte source . . . 74
2.11 Coloration du texte cible . . . 75
2.12 Construction des structures de base de VSearch . . . 75
2.13 Mise en correspondance des occurrences des mots pendant la recherche de sous-chaînes communes à deux textes pour la recherche automatique de copies *verbatim*. 76
2.14 Exemple de structure propositionnelle . . . 77
2.15 Coloration des segments communs de taille ≥ 5 pour la recherche de copies *verbatim*. . . . 78

3.1 Différences entre les valeurs d'idf pour les modalités parole et texte . . . 83
3.2 Valeurs d'idf pour les modalités "texte" et "parole" au sein du corpus mélangé . . . 83
3.3 Recherche de documents audio : l'interface de SCAN . . . 85
3.4 Interface de navigation au sein de "webcasts" . . . 86
3.5 Interface de navigation du système de B. Favre . . . 87
3.6 Indices linguistiques utilisées pour le résumé automatique . . . 88
3.7 Sélection de phrases par Maximal Marginal Relevance . . . 88
3.8 Exemple de requête issue de la campagne DUC . . . 89
3.9 Scores des . . . 90

3.10 Scores des 90

4.1 Lecture experte : le modèle à double voie enrichi 96
4.2 Lecture experte : le modèle connexionniste parallèle distribué (PDP) 97
4.3 Un modèle connexionniste de la lecture 98
4.4 Voisinages orthographiques et phonologiques d'un mot 101
4.5 Physiologie de la dyslexie 105
4.6 Dyslexie et rééducation 107
4.7 Les 20 questions d'EQUER utilisées pour évaluer la robustesse de SQuaLIA . . . 109
4.8 Exemple de questions réelles pour évaluer la robustesse de SQuaLIA 110
4.9 Nombre de bonnes réponses obtenues avec SQuaLIA pour chaque utilisateur . . 111
4.10 Synthèse et reconnaissance vocale pour la correction de questions 114
4.11 Exemple de graphe d'hypothèses de mots généré par Aspell 114
4.12 Exemple d'hypothèses phonétiques produites par LIAPhon 115
4.13 Réordonnancement des documents trouvés en fonction de la lisibilité 119
4.14 Critères envisagés afin d'estimer la difficulté de lecture d'un mot 120
4.15 Pondération BM25 : évolution non linéaire du poids selon la composante TF . . . 122

6.1 Interface de recherche phonétique dans le lexique Manulex 138
6.2 Exemple d'exercice de lecture créé avec le logiciel Mot à Mot 138
6.3 Liste des phonèmes du français issue de Manulex-Infra 139
6.4 Liste des graphèmes du français issue de Manulex-Infra 140

11.1 Schéma général de SQuaLIA pour la campagne TREC-11 161
11.2 Schéma général de SQuaLIA pour la campagne EQUER 163

12.1 DotPlotting 166
12.2 C99 167
12.3 Chaînes lexicales pour la segmentation thématique 168
12.4 Text-Tiling 168

Liste des tableaux

1.1 Evaluation du module de catégorisation des questions 22
1.2 Capacité de SQuaLIA à répondre à certains types de questions 23
1.3 Prédiction de la difficulté d'une question 25
1.4 Identification automatique de questions dans des courriels 27
1.5 Evaluation de la densité pour questions-réponses 38
1.6 Evaluation de SQuaLIA durant EQUER 41
1.7 Taux de couverture des bases de connaissance sur les questions CLEF 43

2.1 Evaluation de la segmentation à base de chaînes lexicales pondérées 57
2.2 Évaluation des méthodes de repérage de citations directes et indirectes. 71

4.1 Les trois premières réécritures retenues pour la question "*kel aje a la Bepierre*" après phonétisation et re-transcription. 115
4.2 Taux de lemmes corrects et pourcentage de phrases identiques à la référence après lemmatisation et filtrage, sur les questions initiales ou réécrites à l'aide de Aspell (Asp) ou de notre système (PhonTrans), pour une ou trois hypothèses. . . . 116

11.1 Résultats officiels de notre participation à EQueR 2004 162
11.2 Evaluation de SQuaLIA sur les données EQueR 2004 après correction 162
11.3 Evaluation de SQuaLIA sur les données CLEF 2006 162

Bibliographie

(Abdou et Savoy, 2007) S. Abdou et J. Savoy, 2007. Considérations sur l'évaluation de la robustesse en recherche d'informations. Dans les actes de *CORIA 2007*, Saint-Etienne (France), 5–20.

(Al-Onaizan et Knight, 2001) Y. Al-Onaizan et K. Knight, 2001. Translating named entities using monolingual and bilingual resources. Dans les actes de *Annual Meeting on Association for Computational Linguistics (ACL'01)*, USA, 400–408.

(Allan, 2001) J. Allan, 2001. Perspectives on information retrieval and speech. Dans A. Coden, E. Brown, et S. Srivinvasen (Eds.), *Information Retrieval Techniques for Speech Applications*. London (UK) : Springer-Verlag.

(Allan, 2002) J. Allan, 2002. *Topic Detection and Tracking : Event-based Information Organization*, Volume 12 de *The Kluwer International Series on Information Retrieval*. Kluwer Academic Publishers.

(Allan, 2005) J. Allan, 2005. Hard track overview in trec 2005. Dans les actes de *TREC 2005*. NIST Special Publication.

(Allan et al., 1996) J. Allan, J. P. Callan, W. B. Croft, L. Ballesteros, D. Byrd, R. Swan, et J. Xu, 1996. Inquery does battle with trec-6. Dans les actes de *The Sixth Text REtrieval Conference (TREC 6)*, Volume NIST Special Publication n° 500-240, USA, 169–206.

(Allan et al., 2000) J. Allan, V. Lavrenko, et H. Jin, 2000. First story detection in tdt is hard. Dans les actes de *ACM CIKM 2000*.

(Allan et al., 2002) J. Allan, V. Lavrenko, et R. Swan, 2002. Explorations within topic tracking and detection. Dans J. Allan (Ed.), *Topic Detection and Tracking – Event-based Information Organization*. Kluwer.

(Amati et al., 2004) G. Amati, C. Carpineto, et G. Romano, 2004. Query difficulty, robustness and selective application of query expansion. Dans les actes de *ECIR'04 - Lecture Notes in Computer Science*, Sunerland, 127–137. Springer.

(Amati et Van Rijsbergen, 2002) G. Amati et C. J. Van Rijsbergen, 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst. 20*(4), 357–389. 582416.

(Ashby et Rayner, 2004) J. Ashby et K. Rayner, 2004. Representing syllable information during silent reading : Evidence from eye movements. *Language and Cognitive Processs 19*(3), 391–426.

(Aslam et Montague, 2000) J. A. Aslam et M. Montague, 2000. Bayes optimal metasearch : a probabilistic model for combining the results of multiple retrieval systems (poster session). Dans les actes de *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, Athens, Greece, 379–381. ACM. 345665.

(Aslam et Montague, 2001) J. A. Aslam et M. Montague, 2001. Models for metasearch. Dans les actes de *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, New Orleans, Louisiana, United States, 276–284. ACM. 384007.

(Aubin et Hamon, 2006) S. Aubin et T. Hamon, 2006. Improving term extraction with terminological resources. Dans T. Salakoski, F. Ginter, S. Pyysalo, et T. Pahikkala (Eds.), *5th International Conference on NLP (FinTAL 2006)*, Volume Advances in Natural Language Processing - LNAI 4139, 380–387. Springer Verlag.

(Aussenac-Gilles et al., 2000) N. Aussenac-Gilles, B. Biébow, et S. Szulman, 2000. Revisiting ontology design : a method based on corpus analysis. knowledge engineering and knowledge management : methods, models and tools. Dans R. Dieng et O. Corby (Eds.), *12th International Conference on Knowledge Engineering and Knowledge Management*, Volume Lecture Notes in Artificial Intelligence Vol 1937, Juan-Les-Pins, France, 171–186. Springer Verlag.

(Aventurier et al., 2008) P. Aventurier, H. Leiser, H. Richard, et P. Bellot, 2008. Ontofruit : Ecophysiologie végétale de l'arboriculture fruitière –un référentiel documentaire indexé par une ontologie du domaine. Dans les actes de *Séminaire Texte et Connaissance*, Paris. INRA.

(Ayache et al., 2005) C. Ayache, B. Grau, et A. Vilnat, 2005. Campagne d'évaluation equer-evalda : évaluation en question-réponse. Dans les actes de *TALN 2005*, Dourdan, France, 6–10.

(Bachimont, 2003) B. Bachimont, 2003. L'indexation multimédia. Dans E. Gaussier et M.-H. Stéfanini (Eds.), *Assistance intelligente à la recherche d'informations*, 153–184. Paris : Hermès.

(Baeza-Yates, 1999) R. Baeza-Yates, 1999. *Modern Information Retrieval*. ACM Press, Addison-Wesley.

(Bakker et Lew, 2002) E. Bakker et M. S. Lew, 2002. Semantic video retrieval using audio analysis. Dans les actes de *1st International Conference on Image and Video Retrieval*, London, 262–270. Springer-Verlag.

(Baziz et al., 2005) M. Baziz, M. Boughanem, et N. Aussenac-Gilles, 2005. A conceptual indexing approach based on document content representation. Dans les actes de *COLIS 2005 Context : nature, impact and role*, Glasgow, Grande-Bretagne, 171–186. LNCS 3507, Springer-Verlag.

(Baziz et al., 2007) M. Baziz, M. Boughanem, et H. Prade, 2007. Une approche de représentation de l'information en ri basée sur les sous-arbres.

(Beeferman et al., 1997) D. Beeferman, A. Berger, et J. Lafferty, 1997. Text segmentation using exponential models. Dans les actes de *2nd conference on Empirical Methods in Natural Language Processing (EMNLP)*, USA.

(Belew, 1989) R. K. Belew, 1989. Adaptive information retrieval : using a connectionist representation to retrieve and learn about documents. Dans les actes de *12th annual international ACM SIGIR conference on Research and development in information retrieval*, Cambridge, Massachusetts, United States, 11–20. 75337.

(Belkin, 2008) N. J. Belkin, 2008. Some(what) grand challenges for information retrieval. *SIGIR Forum 42*(1), 47–54. 1394261.

(Bellegarda, 2005) J. R. Bellegarda, 2005. Unsupervised, language-independent grapheme-to-phoneme conversion by latent analogy. *Speech Communication 46*(2), 140–152.

(Bellengier et al., 2004) E. Bellengier, P. Blache, et S. Rauzy, 2004. Pca : Un système de communication alternative évolutif et réversible. Dans les actes de *ISAAC'04 (International Society for Augmentative and Alternative Communication)*, Neuchâtel, Suisse, 78–85.

(Bellot, 2000a) P. Bellot, 2000a. *Méthodes de classification et de segmentation locales non supervisées pour la recherche documentaire*. Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse.

(Bellot, 2000b) P. Bellot, 2000b. Structuration dynamique de textes pour la recherche documentaire. Dans les actes de *Ecole thématique "Nouveaux défis en science de l'information" - GDR I3*, Marseille.

(Bellot et Boughanem, 2008) P. Bellot et M. Boughanem, 2008. Recherche d'informations et question-réponse. Dans B. Grau et J.-P. Chevallet (Eds.), *La recherche d'informations précises*, 31–68. Lavoisier, Hermes.

(Bellot et al., 2003) P. Bellot, E. Crestan, M. El-Bèze, L. Gillard, et C. de Loupy, 2003. Coupling named entity recognition, vector-space model and knowledge bases for trec-11 question-answering track. Dans les actes de *TREC-2002*, Volume NIST Special publication 500-251.

(Bellot et El-Bèze, 1997) P. Bellot et M. El-Bèze, 1997. Un algorithme de segmentation automatique de corpus - le système s.i.a.c. Dans les actes de *Premières Journées Scientifiques et Techniques (JST 97)*, Avignon (France), 113–117.

(Bellot et El-Bèze, 2000a) P. Bellot et M. El-Bèze, 2000a. Clustering by means of unsupervised decision trees or hierarchical and k-means-like algorithm. *Proc. RIAO 2000 Conf*, 344–363.

(Bellot et El-Bèze, 2000b) P. Bellot et M. El-Bèze, 2000b. Query length, number of classes and routes through clusters : experiments with a clustering method for information retrieval. *Lecture Notes in Computer Science (LNCS) 1746*(IEEE Int. Conf. Comp. Science, Hong-Kong), 196–205.

(Bellot et El-Bèze, 2001a) P. Bellot et M. El-Bèze, 2001a. Classification et segmentation de textes par arbres de décision. *Technique et Science Informatiques (TSI), Hermes 20*, 397–424.

(Bellot et El-Bèze, 2001b) P. Bellot et M. El-Bèze, 2001b. Classification locale non supervisée pour la recherche documentaire. *Traitement Automatique des Langues (TAL) 42*(2), 335–366.

(Benamara et Saint Dizier, 2004) F. Benamara et P. Saint Dizier, 2004. Advanced relaxation for cooperative question answering. Dans M. T. Maybury (Ed.), *New Directions in Question Answering*, 263–274. The MIT Press.

(Beretta et al., 2005) A. Beretta, R. Fiorentino, et D. Poeppel, 2005. The effects of homonymy and polysemy on lexical access : An meg study. *Cognitive Brain Research 24*, 57–65.

(Berrut et Denos, 2003) C. Berrut et N. Denos, 2003. Filtrage collaboratif. Dans E. Gaussier et M.-H. Stéfanini (Eds.), *Assistance intelligente à la recherche d'informations*, 255–284. Paris : Hermès.

(Berry, 1955) M. Berry, 1955. Operational criteria for designing information retrieval systems. *American Documentation 6*, 93–101.

(Besançon et al., 2003) R. Besançon, M. Rajman, et J.-C. Chappelier, 2003. Représentation vectorielle de connaissances sémantiques pour la recherche d'information. Dans E. Gaussier et M.-H. Stéfanini (Eds.), *Assistance intelligente à la recherche d'informations*, 133–152. Paris : Hermès.

(Blache, 2000) P. Blache, 2000. Le rôle des contraintes dans les théories linguistiques et leur intérêt pour l'analyse automatique : les grammaires de propriétés. Dans les actes de *TALN 2000*, Lausanne (Suisse).

(Blache, 2003) P. Blache, 2003. Vers une théorie cognitive de la langue basée sur les contraintes. Dans les actes de *TALN 2003*, Batz-sur-Mer (France).

(Blache et Rauzy, 2007) P. Blache et S. Rauzy, 2007. Le module de reformulation iconique de la plateforme de communication alternative. Dans les actes de *Atelier « Reconstruire la langue dans les communications alternatives et augmentées» (TALN-07)*, Toulouse, France.

(Blache et Rauzy, 2008) P. Blache et S. Rauzy, 2008. Le moteur de prédiction de mots de la plateforme de communication alternative. *Traitement Automatique des Langues (TAL) 48*(2), 47–70.

(Bleasdale, 1987) F. Bleasdale, 1987. Concreteness-dependent associative priming : Separate lexical organization for concrete and abstract words. *Journal of Experimental Psychology : Learning, Memory and Cognition 13*, 582–594.

(Bodner et Masson, 1997) G. Bodner et M. Masson, 1997. Masked repetition priming of words and nonwords : Evidence for a non lexical basis for priming. *Journal of Memory and Language 37*, 268–293.

(Boissière et Dours, 2000) P. Boissière et D. Dours, 2000. Vitipi : A universal writing interface for all. Dans les actes de *6th ERCIM Workshop "User Interfaces for All"*.

(Bonin, 2007) P. Bonin, 2007. Le mot "dragon" est-il toujours traité plus rapidement que le mot "taxe" ? impact de l'âge d'acquisition des mots et de la fréquence objective dans des tâches lexicales. Dans E. Demont, J.-E. Gombert, et M. Lutz (Eds.), *Acquisition du langage : approche intégrée*. Marseille, France : Solal.

(Bonin et al., 2004) P. Bonin, C. Barry, A. Méot, et M. Chalard, 2004. The influence of age of acquisition in word reading and other tasks : A never ending story ? *Journal of Memory and Language 50*, 456–473.

(Bookstein, 1983) Bookstein, 1983. Information retrieval : A sequential learning process. *Journal of the American Society for Information Science*.

(Boughanem et al., 2004a) M. Boughanem, W. Kraaij, et J. Y. Nie, 2004a. Modèles de langue pour la recherche d'information. Dans M. Ihadjadene (Ed.), *Les systèmes de recherche d'informations*, 163–182. Hermes-Lavoisier.

(Boughanem et Soulé-Dupuy, 1997) M. Boughanem et C. Soulé-Dupuy, 1997. Mercure at trec6. Dans les actes de *The Sixth Text REtrieval Conference (TREC 6)*, 321–328. NIST Special Publication 500-240.

(Boughanem et Tamine, 2004) M. Boughanem et L. Tamine, 2004. Connexionisme et génétique pour la recherche d'informations. Dans M. Ihadjadene (Ed.), *Les systèmes de recherche d'informations*, 77–104. Paris : Hermès-Lavoisier.

(Boughanem et al., 2004b) M. Boughanem, M. Tmar, et H. Tebri, 2004b. Filtrage d'information. Dans M. Ihadjadene (Ed.), *Méthodes avancées pour les systèmes de recherche d'informations*, 137–162. Paris : Hermès.

(Bourigault et al., 2004) D. Bourigault, N. Aussenac-Gilles, et J. Charlet, 2004. Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle (RIA) , " Techniques Informatiques et structuration de terminologiques 18*(1), 87–110.

(Bourigault et C., 2000) D. Bourigault et J. C., 2000. Construction de ressources terminologiques. Dans J. Pierrel (Ed.), *Ingénierie des Langues*, 215–230. Hermès Sciences.

(Bourigault et al., 2001) D. Bourigault, M. L'Homme, et J. C., 2001. *Recent advances in computational terminology*. Amsterdam : John Benjamins Publishing Company.

(Brill et al., 2001) E. Brill, J. Lin, M. Banko, S. Dumais, et A. Ng, 2001. Data-intensive question answering. Dans les actes de *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*, 393–400. NIST Special Publication 500-250.

(Brill et Moore, 2000) E. Brill et R. C. Moore, 2000. An improved error model for noisy channel spelling correction. Dans les actes de *38th Annual Meeting of the ACL*, 286–293.

(Brown et Chong, 1997) E. W. Brown et H. A. Chong, 1997. The guru system in trec-6. Dans les actes de *Proceedings of TREC6*, 535–540. NIST Special Publication 500-240.

(Brown et Watson, 1987) G. Brown et F. Watson, 1987. First in, first out : Word learning age and spoken word frequency as predictors of word familiarity and word naming latency. *Memory and Cognition 15*(3), 208–216.

(Brown, 1997) G. D. A. Brown, 1997. Connectionism, phonology, reading, and regularity in developmental dyslexia. *Brain and Language 59*(2), 207–235.

(Bruandet et Chevallet, 2003) M.-F. Bruandet et J.-P. Chevallet, 2003. Utilisaton et construction de bases de connaissances pour la recherche d'information. Dans E. Gaussier et M.-H. Stéfanini (Eds.), *Assistance intelligente à la recherche d'informations*, 99–132. Paris : Hermès.

(Bruner, 1983) J. Bruner, 1983. *Child's talk, learning to use language*. New-York (USA) : W.W. Norton and Company Inc.

(Bruza et al., 2000) P. Bruza, R. Mc Arthur, et S. Dennis, 2000. Interactive internet search : Keyword, directory and query reformulation mechanisms compared. Dans les actes de *ACM-SIGIR 2000*, 280–288. ACM Press.

(Buckley, 1985) C. Buckley, 1985. Implementation of the smart information retrieval system. Rapport technique, Department of Computer Science, Cornell University.

(Buckley et al., 1996) C. Buckley, A. Singhal, M. Mitra, et G. Salton, 1996. New retrieval approaches using smart : Trec-4. Dans les actes de *Fourth Text REtrieval Conference (TREC-4)*, Gaithersburg (USA), 25–48. NIST Special publication.

(Burger, 2003) J. Burger, 2003. Mitre's quanda at trec-12. Dans les actes de *TREC-12*, Gaithersburg, USA.

(Burges, 1998) C. Burges, 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* *2*(2), 121–167.

(Béchet, 2001) F. Béchet, 2001. Liaphon : Un système complet de phonétisation de textes. *Traitement Automatique des Langues (TAL)* *42*(1), 47–67.

(Caelen et al., 2003) J. Caelen, V. Eglin, et S. Hollard, 2003. Evaluation de documents par oculométrie. Dans E. Gaussier et M.-H. Stéfanini (Eds.), *Assistance intelligente à la recherche d'informations*, 285–315. Paris : Hermès.

(Cai et al., 2003) R. Cai, L. Lu, et H.-J. Zhang, 2003. Using structure patterns of temporal and spectral feature in audio similarity measure. Dans les actes de *Proceedings of the eleventh ACM international conference on Multimedia*, Berkeley, CA, USA, 219–222. ACM. 957056.

(Callan, 1994) J. P. Callan, 1994. Passage-level evidence in document retrieval. Dans les actes de *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, Dublin, Ireland, 302–310. Springer-Verlag New York, Inc. 188589.

(Campolini, 2004) C. Campolini, 2004. *Troubles du Developpement et Langage*. Ellipse.

(Carmel et al., 2003) D. Carmel, Y. Maarek, M. Mandelbrod, Y. Mass, et A. Soffer, 2003. Searching xml documents via xml fragments. Dans les actes de *26th ACM SIGIR Conference*, 151–158.

(Castells et al., 2007) P. Castells, M. Fernandez, et D. Vallet, 2007. An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Transactions on Knowledge and Data Engineering* *19*(2), 261–272. 1191759.

(Chang et al., 2003) Y. Chang, H. Xu, et S. Bai, 2003. Trec 2003 question answering track at cas-ict. Dans les actes de *TREC-12*, Gaithersburg, USA.

(Changeux, 2002) J. Changeux, 2002. *L'homme de vérité*. Paris : Odile Jacob (2008 pour la trad. française).

(Chaudiron, 2004a) S. Chaudiron, 2004a. *Evaluation des systèmes de traitement de l'information*. Traité des Sciences et Technoqies de l'Information. Paris : Hermès Lavoisier.

(Chaudiron, 2004b) S. Chaudiron, 2004b. La place de l'usager dans l'évaluation des systèmes de recherche d'informations. Dans S. Chaudiron (Ed.), *Evaluation des systèmes de traitement de l'information*, 287–310. Paris : Hermès.

(Chevalier et al., 2007) M. Chevalier, C. Julien, et C. Soulé-Dupuy, 2007. Prise en compte de l'usager dans la recherche d'information. Dans les actes de *PeCUSI (Prise en Compte de l'Utilisateur dans les Systèmes d'Information), atelier de Inforsid 2007*, Perros Guirec (France), 274–284.

(Chevallet, 2004) J.-P. Chevallet, 2004. Modélisation logique pour la recherche d'informations. Dans les actes de *Les systèmes de recherche d'informations*, 105–138. Paris : Hermès.

(Chidlovskii, 2003) B. Chidlovskii, 2003. Métarecherche : la recherche distribuée d'informations. Dans E. Gaussier et M.-H. Stéfanini (Eds.), *Assistance intelligente à la recherche d'informations*, 187–218. Paris : Hemès.

(Choi, 2000) F. Choi, 2000. Advances in domain independent linear text segmentation. Dans les actes de *1st Meeting of the North American Chapter of the Association for Computational Linguistics*, USA.

(Chomsky, 1964) N. Chomsky, 1964. Current issues in linguistic theory. Dans J. Fodor et B. Katz (Eds.), *The structure of language*. Prentice Hall.

(Chomsky, 1981) N. Chomsky, 1981. *Lectures in Government and Binding*. Dordrecht : Foris Publications.

(Chomsky, 2000) N. Chomsky, 2000. *News Horizons in the Study of Language and Mind*. Cambridge University Press.

(Clarke et al., 2003) C. L. A. Clarke, G. V. Cormack, G. Kemkes, M. Laszlo, T. R. Lynam, L. Terra Egidio, et P. Tilker, 2003. Statistical selection of exact answers (multitext experiments for trec 2002). Dans les actes de *The Eleventh Text Retrieval Conference (TREC 2002)*, Volume NIST Special Publication 500-251.

(Clarke et al., 2000) C. L. A. Clarke, G. V. Cormack, D. I. E. Kisman, et T. R. Lynam, 2000. Question answering by passage selection (multitext experiments for trec-9). Dans les actes de *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*, 673–684. NIST Special Publication 500-249.

(Clough et al., 2002) P. Clough, R. Gaizauskas, S. Piao, et Y. Wilks, 2002. Measuring text reuse. Dans les actes de *40th Meeting for the Association for Computational Linguistics (ACL)*.

(Coltheart, 1978) M. Coltheart, 1978. Lexical access in simple reading task. Dans G. Underwood (Ed.), *Strategies of information processing*, 151–216. London : Academic Press.

(Coltheart et Coltheart, 1997) M. Coltheart et V. Coltheart, 1997. Reading comprehension is not exclusively reliant upon phonological representation. *Cognitive Neuropsychology 14*, 164–175.

(Coltheart et al., 1977) M. Coltheart, E. Davelaar, J. Jonasson, et D. Besner, 1977. Access to the internal lexicon. Dans S. Dornic (Ed.), *Attention and Performance VI*, 535–555. London : Academic Press.

(Coltheart et al., 2001) M. Coltheart, K. Rastle, C. Perry, R. Langdon, et J. C. Ziegler, 2001. Drc : A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review 108*, 204–256.

(Croft, 1983) W. B. Croft, 1983. Experiments with representation in a document retrieval system. *Information Technology : Research Development 2*(1), 1–21.

(Cronen-Townsend et al., 2002) S. Cronen-Townsend, Y. Zhou, et W. B. Croft, 2002. Predicting query performance. Dans les actes de *SIGIR 2002*, 299–306. ACM Press.

(Cui et al., 2005) H. Cui, R. Sun, K. Li, M.-Y. Kan, et T.-S. Chua, 2005. Question answering passage retrieval using dependency relations. Dans les actes de *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, Salvador, Brazil, 400–407. ACM Press. 1076103.

(Cutting,) D. Cutting. The lucene search engine.

(Daille, 2002) B. Daille, 2002. *Découvertes linguistiques en corpus*. Habilitation à diriger des recherches, Université de Nantes, France.

(Daille et Morin, 2000) B. Daille et E. Morin, 2000. Reconnaissance automatique des noms propres de la langue écrite : les récentes réalisations. *Traitement Automatique des Langues (TAL) - Hermès 41*(3), 601–622.

(Dang et al., 2006) H. Dang, J. Lin, et D. Kelly, 2006. Overview of the trec 2006 question answering track. *TREC 2006*.

(Danon-Boileau, 2004) L. Danon-Boileau, 2004. *Les troubles du langage et de la communication chez l'enfant*. Que sais-je ? PUF.

(Dave et al., 2003) K. Dave, S. Lawrence, et D. M. Pennock, 2003. Mining the peanut gallery : Opinion extraction and semantic classification of product reviews. Dans les actes de *Twelfth International World Wide Web Conference (WWW'03)*.

(de Loupy et Bellot, 2000) C. de Loupy et P. Bellot, 2000. Evaluation of document retrieval systems and query difficulty. Dans les actes de *LREC'2000 Satellite Workshop "Using Evaluation within HLT Programs : Results and trends"*, Athènes, Grèce, 31–38.

(de Loupy et al., 1999) C. de Loupy, P. Bellot, M. El-Bèze, et P.-F. Marteau, 1999. Query expansion and automatic classification. Dans les actes de *7th Text REtrieval Conference*, 443–450. NIST Special Publication 500-242.

(Deerwester et al., 1990) S. Deerwester, S. Dumais, T. Landauer, G. Furnas, et R. Harshman, 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science 41*(6), 391–407.

(Dehaene, 2007) S. Dehaene, 2007. *Les neurones de la lecture*. Paris : Odile Jacob.

(Deorowicz et Ciura, 2005) S. Deorowicz et M. G. Ciura, 2005. Correcting spelling errors by modelling their causes. *International Journal of Applied Mathematics and Computer Science 15*(2), 275–285.

(Dias et Alves, 2005) G. Dias et E. Alves, 2005. Unsupervised topic segmentation based on word co-occurrence and multi-word units for text summarization. Dans les actes de *ELECTRA Workshop associated to 28th Annual International ACM SIGIR Conference*, Salvador, Brazil, 41–48.

(Dickinson et al., 2002) A. Dickinson, P. Gregor, et A. F. Newell, 2002. Ongoing investigation of the ways in which some of the problems encountered by some dyslexics can be alleviated using computer techniques. 638268 97-103.

(Donovan, 2003) R. E. Donovan, 2003. Topics in decision tree based speech synthesis. *Computer Speech and Language 17*(1), 43–67.

(Ducrot et Nguyen, 2003) S. Ducrot et N. Nguyen, 2003. Special issue on language disorders and reading acquisition : Introductory remarks. *Current Psychology Letter (CPL), Behaviour, Brain and Cognition, Special Issue on Language Disorders and Reading Acquisition 1*(10).

(Dufour et al., 2005) C. Dufour, E. G. Toms, J. Lewis, et R. Baecker, 2005. User strategies for handling information tasks in webcasts. Dans les actes de *CHI '05 extended abstracts on Human factors in computing systems*, Portland, OR, USA, 1343–1346. ACM. 1056912 1343-1346.

(Dulucq et Tichit, 2001) S. Dulucq et L. Tichit, 2001. À propos de la comparaison de structures secondaires d'arn. Rapport technique, Rapport interne LABRI.

(Dutoit, 1997) T. Dutoit, 1997. High-quality text-to-speech synthesis : An overview. *Journal of Electrical and Electronics Engineering (Australia) 17*, 25–36. IREE INSTITUTION OF RADIO AND ELECTRONICS.

(Déjerine, 1892) J. Déjerine, 1892. *Contribution à l'étude anatomo-pathologique et clinique des différentes variétés de cécité verbale*, Volume 4. Mémoire de la Société de Biologie.

(Echihabi et Marcu, 2003) A. Echihabi et D. Marcu, 2003. A noisy-channel approach to question answering. Dans les actes de *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, Sapporo, Japan, 16–23. Association for Computational Linguistics. 1075099.

(Edelman, 2007) G. Edelman, 2007. *La science du cerveau et de la connaissance*. Paris : Odile Jacob.

(Eglin et Emptoz, 1997) V. Eglin et H. Emptoz, 1997. Logarithmic spiral grid and gaze control for the development of strategies of visual segmentation on a document. Dans les actes de *Intl. Conference on Document Analysis and Recognition (ICDAR-97)*, 689–692.

(El-Bèze, 2006) M. El-Bèze, 2006. Les systèmes de questions-réponses. Dans G. Sabbah (Ed.), *Compréhension des langues et interaction*, 277–298. Paris : Hermès.

(Embarek, 2008) M. Embarek, 2008. *Un système de question-réponse dans le domaine médical. Le système Esculape*. Thèse de doctorat, Université Paris-Est Marne la Vallée.

(Fairweather et al., 2002) P. Fairweather, V. Hanson, S. Detweiler, et R. Schwerdtfeger, 2002. From assistive technology to a web accessibility service. Dans les actes de *ACM-Assets 2002*, Edinburgh, Scotland, 4–8. ACM-Press.

(Favre, 2003a) B. Favre, 2003a. *Indexation Multimédia - Caractérisation du déséquilibre entre les modalités texte et parole*. Mémoire de master recherche, Université d'Avignon et des Pays de Vaucluse.

(Favre, 2003b) B. Favre, 2003b. Recherche documentaire multimédia : caractérisation du déséquilibre entre les modalités texte et parole. Dans les actes de *Majecstic'03*.

(Favre, 2007) B. Favre, 2007. *Résumé automatique de parole pour un accès efficace aux bases de données audio*. Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse.

(Favre et al., 2004a) B. Favre, P. Bellot, et J.-F. Bonastre, 2004a. Information retrieval on mixed written and spoken documents. Dans les actes de *7è conférence RIAO*, Avignon, France, 826–835.

(Favre et al., 2004b) B. Favre, J.-F. Bonastre, et P. Bellot, 2004b. Recherche d'information dans un mélange de documents écrits et parlés. Dans les actes de *Journées d'Etude de la Parole (JEP 2004)*, Fez, Maroc, 403–412.

(Favre et al., 2007) B. Favre, J.-F. Bonastre, et P. Bellot, 2007. An interactive timeline for speech database browsing. Dans les actes de *Interspeech 2007*, Antwerpen (Belgique).

(Favre et al., 2006) B. Favre, J.-F. Bonastre, P. Bellot, et F. Capman, 2006. Accès aux connaissances orales par le résumé automatique. Dans les actes de *6è journées francophones "Extraction et Gestion des Connaissances" (EGC 2006)*, Lille, France.

(Favre et al., 2006) B. Favre, F. Béchet, P. Bellot, F. Boudin, M. El-Bèze, L. Gillard, G. Lapalme, et J.-M. Torres-Moreno, 2006. The lia-thales summarization system at duc-2006. Dans les actes de *Document Understanding Conference (DUC-2006)*, New York (USA).

(Favre et al., 2005) B. Favre, F. Béchet, et P. Nocera, 2005. Robust named entity extraction from large spoken archives. Dans les actes de *HLT-EMNLP'05*.

(Ferrand, 2007) L. Ferrand, 2007. *Psychologie cognitive de la lecture. Reconnaissance des mots écrits chez l'adulte.* Ouvertures psychologiques. Bruxelles, Belgique : de Boeck.

(Ferrand et New, 2003) L. Ferrand et B. New, 2003. Syllabic length effects in visual word recognition and naming. *Acta Psychologica 113*, 167–183.

(Ferret, 2002) O. Ferret, 2002. Using collocations for topic segmentation and link detection. Dans les actes de *19th international conference on Computational linguistics*, Volume 1, Taipei, Taiwan.

(Ferret et al., 2002) O. Ferret, B. Grau, M. Hurault-Plantet, G. Illouz, L. Monceaux, I. Robba, et A. Vilnat, 2002. Finding an answer based on the recognition of the question focus. Dans les actes de *The Tenth Text REtrieval Conference (TREC 2001)*, Volume NIST Special Publication SP 500-250, Gaithersburg, Maryland, USA.

(Ferret et Zweigenbaum, 2008) O. Ferret et P. Zweigenbaum, 2008. Connaissances et systèmes de question-réponse. Dans B. Grau et J.-P. Chevallet (Eds.), *La recherche d'informations précises : apprentissage, traitement automatique de la langue et connaissances pour les systèmes de question-réponse*, 133–170. Paris : Hermès.

(Fiset et al., 2006) S. Fiset, M. Arguin, et D. Fiset, 2006. An attempt to simulate letter-by-letter dyslexia in normal readers. *Brain and Language 98*(3), 251–263.

(Flavier, 2006) N. Flavier, 2006. *Détection de questions similaires extraites de textes non structurés*. Mémoire de master recherche, Université d'Avignon et des Pays de Vaucluse.

(Flavier et Bellot, 2007) N. Flavier et P. Bellot, 2007. Vers un appariement automatique de questions extraites de courriers électroniques. Dans les actes de *Conférence Francophone sur l'Apprentissage Automatique (CAP 2007)*, Grenoble, France.

(Flesch, 1948) R. Flesch, 1948. A new readability yardstick. *Journal of applied psychology 32*, 221–233.

(Fluhr, 2004) C. Fluhr, 2004. L'évaluation des systèmes de recherche d'informations textuelles. Dans S. Chaudiron (Ed.), *Evaluation des systèmes de traitement de l'information*, 27–46. Hermes Lavoisier.

(Foote, 1999) J. Foote, 1999. An overview of audio information retrieval. *ACM Springer Multimedia Systems 7*(1), 2–10. 297251.

(Foucault, 1966) M. Foucault, 1966. *Les mots et les choses - Une archéologie des sciences humaines*. Collection tel (édition 2007). Gallimard.

(Fox et Shaw, 1994) E. Fox et J. Shaw, 1994. Combination of multiple searches. Dans les actes de *2nd Text REtrieval Conference*, 243–252. NIST Special publication 500-215.

(Fredouille et al., 2004) C. Fredouille, D. Matrouf, G. Linares, et P. Nocera, 2004. Segmentation en macro-classes acoustiques d'émissions radiophoniques dans le cadre d'ester. Dans les actes de *Journée d'Etudes sur la Parole (JEP'04)*, Fez (Maroc).

(Friedman et Goldzsmidt, 1996) N. Friedman et M. Goldzsmidt, 1996. Building classifiers using bayesian networks. Dans les actes de *Thirteenth National Conference on Artificial Intelligence (NCAI)*, 1277–1284.

(Frith, 1985) U. Frith, 1985. Beneath the surface of developmental dyslexia. Dans K. Petterson, J. Marshall, et M. Coltheart (Eds.), *Surface dyslexia*. London : Routledge and Kegan Paul.

(Frost et al., 1987) R. Frost, L. Katz, et S. Bentin, 1987. Strategies for visual word recognition and orthographic depth. *Journal of Experimental Psychology : Human Perception and Performance 13*, 104–115.

(Fuhr, 1992) N. Fuhr, 1992. Probabilistics models in ir.

(Galley et al., 2003) M. Galley, K. McKeown, F.-L. E., et H. Jing, 2003. Discourse segmentation of multi-party conversation. Dans les actes de *ACL*.

(Galliano et al., 2005) S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, et G. Gravier, 2005. The ester phase 2 evaluation campaign for the rich transcription of french broadcast news.

(Gao et al., 2004) J. Gao, J. Y. Nie, G. Wu, et G. Cao, 2004. Dependence language model for information retrieval. Dans les actes de *Proceedings of the 27th annual international ACM SIGIR conference on Research and developement in information retrieval*, 170–177. ACM Press.

(Garay et Johnson, 1979) M. R. Garay et D. S. Johnson, 1979. *Computers and Intractability. A Guide to the Theory of NP-Completeness*. New York, USA : W.H. Freeman and Company.

(Garofolo et al., 2000) J. Garofolo, C. Auzanne, et M. Voorhees Ellen, 2000. The trec spoken document retrieval track : A success story. Dans les actes de *Eighth Text REtrieval Conference (TREC-8)*. NIST Special Publication.

(Gauch et al., 2003) S. Gauch, J. Chaffee, et P. Pretschner, 2003. Ontology based user profiles for search and browsing. *Special issue on user modelling for Web and hypermedia information retrieval*.

(Gaussier et al., 2003) E. Gaussier, C. Jacquemin, et P. Zweigenbaum, 2003. Traitement automatique des langues et recherche d'information. Dans E. Gaussier et M.-H. Stéfanini (Eds.), *Assistance intelligente à la recherche d'informations*, 71–96. Paris : Hermès.

(Gaussier et Stéfanini, 2003) E. Gaussier et M.-H. Stéfanini, 2003. *Assistance intelligente à la recherche d'informations*. Paris : Hermès.

(Gernsbacher, 1984) M. Gernsbacher, 1984. Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness and polysemy. *Journal of Experimental Psychology : General 113*(2), 256–281.

(Getoor et Diehl, 2005) L. Getoor et C.-P. Diehl, 2005. Link mining : A survey. *ACM SIGKDD Explorations 7*(2), 3–12.

(Gibson et al., 1998) D. Gibson, J. Kleinberg, et P. Raghavan, 1998. Inferring web communities from link topology. Dans les actes de *ACM Conference on Hypertext and Hypermedia*, 225–234.

(Gillard, 2007) L. Gillard, 2007. *Quelles méthodes pour les systèmes de Questions/Réponses ? Une avancée vers le tout numérique*. Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse, France.

(Gillard et al., 2003) L. Gillard, P. Bellot, et M. El-Bèze, 2003. Bases de connaissances pour asseoir la crédibilité des réponses d'un système de q/r. Dans les actes de *TALN 2003*, Batz sur Mer, France.

(Gillard et al., 2005) L. Gillard, P. Bellot, et M. El-Bèze, 2005. Le lia à equer (campagne technolangue des systèmes questions-réponses). Dans les actes de *atelier Évaluation en Question Réponse (EQueR) de la 12è conférence Traitement Automatique des Langues Naturelles (TALN)*, Dourdan, France, 81–84.

(Gillard et al., 2006a) L. Gillard, P. Bellot, et M. El-Bèze, 2006a. Influence de mesures de densité pour la recherche de passages et l'extraction de réponses dans un système de questions-réponses. Dans les actes de *3è Conférence en Recherche d'Informations et Applications (CORIA)*, Lyon, France, 193–204.

(Gillard et al., 2006b) L. Gillard, P. Bellot, et M. El-Bèze, 2006b. Question answering evaluation survey. Dans les actes de *5è conférence Language Resources and Evaluation Conference (LREC)*, Genova, Italie.

(Gillard et al., 2006c) L. Gillard, P. Bellot, et M. El-Bèze, 2006c. Questions booléennes : Oui ou non, des questions et des réponses. Dans les actes de *13ième conférence Traitement Automatique des Langues Naturelles (TALN)*, Louvain (Belgique), 159–166.

(Gillard et al., 2007a) L. Gillard, P. Bellot, et M. El-Bèze, 2007a. Analyse des échecs d'une méthode pour traiter les questions définitoires soumises à un système de questions/réponses. Dans les actes de *TALN 2007*, Toulouse (France).

(Gillard et al., 2007b) L. Gillard, P. Bellot, et E.-B. M., 2007b. D'une compacité positionnelle à une compacité probabiliste pour un système de questions/réponses. Dans les actes de *4è Conférence en Recherche d'Informations et Applications (CORIA)*, Saint-Etienne (France).

(Gillard et al., 2008) L. Gillard, P. Bellot, et E.-B. M., 2008. Quelles combinaisons de scores et de critères numériques pour un système de questions/réponses ? Dans les actes de *TALN 2008*, Avignon (France).

(Gillard et al., 2007a) L. Gillard, L. Sitbon, P. Bellot, et M. El-Bèze, 2007a. Dernières évolutions de squalia, le système de questions/réponses du lia. *Traitement Automatique des Langues (TAL) - Hermès 46*(3/2005), 41–70.

(Gillard et al., 2007b) L. Gillard, L. Sitbon, E. Blaudez, P. Bellot, et M. El-Bèze, 2007b. Relevance measures for question answering, the lia at qa@clef-2006. Dans les actes de *Evaluation of Multilingual and Multi-modal Information Retrieval*, Volume 4730/2007, 440–449. Springer-Verlag.

(Goldstein et al., 2000) J. Goldstein, V. Mittal, J. Carbonell, et J. Callan, 2000. Creating and evaluation multi-document sentence extract summaries. Dans les actes de *CIKM'00*, McLean (USA). ACM Press.

(Gong et Liu, 2001) Y. Gong et X. Liu, 2001. Generic text summarization using relevance measure and latent semantic analysis. Dans les actes de *SIGIR'01*, 19–25. ACM Press.

(Graesser et al., 1992) A. Graesser, N. Person, et J. Hubert, 1992. *Mechanisms that Generate Questions*. Lawrence Erlbaum Associates.

(Grainger et Ferrand, 1996) J. Grainger et L. Ferrand, 1996. Masked orthographic and phonological priming in visual word recognition and naming : Cross task comparisons. *Journal of Memory and Language 35*, 623–647.

(Grau et Chevallet, 2008) B. Grau et J.-P. Chevallet, 2008. *La recherche d'informations précises : apprentissage, traitement automatique de la langue et connaissances pour les systèmes de question-réponse*. Paris : Hermès.

(Grefenstette, 1999) G. Grefenstette, 1999. The www as a resource for example-based mt tasks. Dans les actes de *ASLIB'99 Translating and the Computer*.

(Greiif et al., 2002) W. R. Greiif, W. T. Morgan, et J. M. Ponte, 2002. The role of variance in term weighting for probabilistic information retrieval. Dans les actes de *Proceedings of the eleventh international conference on Information and knowledge management*, McLean, Virginia, USA, 252–260. ACM Press. 584836 252-259.

(Grivolla et al., 2005) J. Grivolla, P. Jourlin, et R. De Mori, 2005. Automatic classification of queries by expected retrieval performance. Dans les actes de *SIGIR'06*, Salvador. ACM Pres.

(Grodzinsky, 2007) Y. Grodzinsky, 2007. La syntaxe générative dans le cerveau. Dans J. Bricmont et J. Franck (Eds.), *Chomsky (Les Cahiers de l'Herne)*. Paris : Editions de l'Herne.

(Guénot, 2006) M.-L. Guénot, 2006. *Eléments de grammaire du français pour une théorie descriptive et formelle de la langue*. Thèse de doctorat, Université Aix-Marseille I, CNRS.

(Habert et Zweigenbaum, 2002) B. Habert et P. Zweigenbaum, 2002. Régler les règles. *Traitement Automatique des Langues (TAL) 43*(3), 83–105.

(Habib, 1997) M. Habib, 1997. *Dyslexie : le cerveau singulier*. Collection Neuropsychologie. Marseille, France : Solal.

(Hamon et Nazarenko, 2001) T. Hamon et A. Nazarenko, 2001. Detection of synonymy links between terms : experiment and results. Dans B. D., J. C., et M. L'Homme (Eds.), *Recent Advances in Computational Terminology*. John Benjamins.

(Harabagiu et al., 2001) S. Harabagiu, D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Girju, V. Rus, et P. Morarescu, 2001. Falcon : Boosting knowledge for answer engines.

(Harm et Seidenberg, 1999) M. Harm et M. Seidenberg, 1999. Phonology, reading acquisition and dyslexia : Insights from connectionist models. *Psychological Review 106*, 491–528.

(Harm et Seidenberg, 2004) M. Harm et M. Seidenberg, 2004. Computing the meaning of words in reading : Cooperative division of labor between visual and phonological processes. *Psychological Review 111*(3), 662–720.

(Harman, 1995) D. K. Harman, 1995. Overview of the third text retrieval conference (trec-3). Dans les actes de *Text REtrieval Conference TREC-3*, Volume NIST special publication n° 500-225, Gaithersburg, USA, 1–19.

(Harth, 1993) E. Harth, 1993. *The creative loop : how the brain makes a mind*. New-York : Addison-Wesley.

(Hearst, 1994) M. Hearst, 1994. Multi-paragraph segmentation of expository text. Dans les actes de *ACL'94*, Las Cruces, NM, USA.

(Hearst, 1997) M. Hearst, 1997. Text-tiling : segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 59–66.

(Hearst, 1999) M. Hearst, 1999. User interfaces and visualization. Dans R. Baeza-Yates et B. Ribeiro-Neto (Eds.), *Modern Information Retrieval*. Addison-Wesley Longman Publishing Company.

(Hernandez et al., 2006) N. Hernandez, J. Mothe, C. Chrisment, et D. Egret, 2006. Modeling context through domain ontologies. *Journal of Information Retrieval, Special issue Contextual Information Retrieval*.

(Herrera et al., 2004) J. Herrera, A. Penas, et F. Verdejo, 2004. Question answering pilot task at clef 2004. Dans les actes de *CLEF 2004*, Bath, UK.

(Hersh et Over, 2001) W. Hersh et P. Over, 2001. Interactivity at the text retrieval conference (trec). *Information Processing and Management 37*(3), 365–367.

(Hovy et al., 2001) E. Hovy, U. Hermjakob, et C. Y. Lin, 2001. The use of external knowledge in factoid qa. Dans les actes de *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*, 644–652. Proceedings of the Tenth Text REtrieval Conference (TREC-10).

(Ihadjadene, 2004) M. Ihadjadene, 2004. *Les systèmes de recherche d'informations*. Paris : Hermès-Lavoisier.

(Iida et al., 2003) A. Iida, N. Campbell, F. Higuchi, et M. Yasumura, 2003. A corpus-based speech synthesis system with emotion. *Speech Communication 40*(1-2), 161–187.

(Ingwersen et Belkin, 2004) P. Ingwersen et N. Belkin, 2004. Information retrieval in context - irix : workshop at sigir 2004 - sheffield. *SIGIR Forum 38*(2), 50–52. http ://doi.acm.org/10.1145/1041394.1041405 ACM.

(Istrate et al., 2005) D. Istrate, N. Scheffer, C. Fredouille, et J.-F. Bonastre, 2005. Broadcast news speaker tracking for ester 2005 campaign. Dans les actes de *Eurospeech'05*, Lisboa (Portugal).

(Ittycheriah et al., 2001) A. Ittycheriah, M. Franz, et S. Roukos, 2001. Ibm's statistical question answering system (trec-10). Dans les actes de *Proceedings of the Tenth Text REtrieval Conference (TREC-10)*, 258–264. NIST Special Publication 500-250.

(Ittycheriah et Roukos, 2002) A. Ittycheriah et S. Roukos, 2002. Ibm's statistical question answering system – trec11. Dans les actes de *11th-Text REtrieval Conference (TREC 11)*, Volume NIST special publication SP 500-251.

(Jackiewicz, 2006) A. Jackiewicz, 2006. Relations intersubjectives dans les discours rapportés. *Traitement Automatique des Langues (TAL) Discours et document : traitements automatiques*, 65–87.

(Jacquemart et Zweigenbaum, 2003) P. Jacquemart et P. Zweigenbaum, 2003. Towards a medical question-answering system : a feasibility study. Dans P. L. B. e. R. Baud (Ed.), *Actes Medical Informatics Europe*. Amsterdam : IOS Press.

(Jacquemin, 1997) C. Jacquemin, 1997. *Variation terminologique : reconnaissance et acquisition automatique de termes et de leurs variantes en corpus*. Habilitation à diriger des recherches, Université de Nantes, France.

(Jakubowicz, 2007) C. Jakubowicz, 2007. Grammaire universelle et trouble spécifique du langage. Dans J. Bricmont et J. Franck (Eds.), *Chomsky (Les Cahiers de l'Herne)*. Paris : Editions de l'Herne.

(James et Draffan, 2004) A. James et E. Draffan, 2004. The accuracy of electronic spell checkers for dyslexic learners. *PATOSS bulletin*.

(Jansen et Kroner, 2003) B. Jansen et G. Kroner, 2003. The impact of automated assistance on the information retrieval process. Dans les actes de *ACM-CHI 2003*, Ft. Lauderdale, USA, 1004–1005. ACM Press.

(Jenner et al., 2003) B. Jenner, J. Kobler, P. McKenzie, et J. Toran, 2003. Completeness results for graph isomorphism. *Journal of Computer and System Sciences 66*, 549–566.

(Johnson et al., 2000) S. Johnson, P. Jourlin, K. Sparck-Jones, et P. Woodland, 2000. Spoken document retrieval for trec-8 at cambridge university. Dans les actes de *Eighth Text REtrieval Conference (TREC-8)*, Gaithersburg (USA), 197–206. NIST Special Publication 500-246.

(Jolion, 2001) J. Jolion, 2001. Feature similarity. Dans M. S. Lew (Ed.), *Principles of Visual Information Retrieval*. London, UK : Springer-Verlag.

(Jones et al., 1996) G. J. F. Jones, J. T. Foote, K. Sp, r. Jones, et S. J. Young, 1996. Retrieving spoken documents by combining multiple index sources. Dans les actes de *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, Zurich, Switzerland, 30–38. ACM. 243208 30-38.

(Jonsson et al., 2004) A. Jonsson, F. Anden, L. Degerstedt, A. Flycht-Eriksson, M. Merkel, et S. Norberg, 2004. Experiences from combining dialogue system development with information extraction techniques. Dans M. T. Maybury (Ed.), *New directions in question answering*, 153–164. AAAI Press / The MIT Press.

(Jumel, 2005) B. Jumel, 2005. *Comprendre et aider l'enfant dysexique*. Paris : Dunod.

(Kan et al., 1998) M. Y. Kan, J. Klavans, et K. McKeown, 1998. Linear segmentation and segment significance. Dans les actes de *6th International Workshop of Very Large Corpora*.

(Kandel et Moles, 1958) L. Kandel et A. Moles, 1958. Application de l'indice de flesch à la langue française. *The Journal of Educationnal Research 21*, 283–287.

(Kaszkiel et Zobel, 1997) M. Kaszkiel et J. Zobel, 1997. Passage retrieval revisited. Dans les actes de *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, Philadelphia, Pennsylvania, United States, 178–185. ACM Press. 258561.

(Kelly et Lin, 2007) D. Kelly et J. Lin, 2007. Overview of the trec 2006 ciqa task. *ACM SIGIR Forum 41*(1), 107–116.

(Kempson et al., 2000) R. Kempson, W. Meyer-Viol, et D. Gabbay, 2000. *Dynamic Syntax : The Flow of Language Understanding*. Wiley-Blackwel.

(Kobler et al., 1993) J. Kobler, U. Schoning, et J. Toran, 1993. *The graph isomorphism problem : its structural complexity*. Basel : Birkhauser.

(Kouylekov et al., 2006) M. Kouylekov, M. Negri, B. Magnini, et B. Coppola, 2006. Towards entailment-based question-answering. Dans les actes de *Working notes of the CLEF 2005 workshop*.

(Kuhn et De Mori, 1995) R. Kuhn et R. De Mori, 1995. The application of semantic classification trees to natural language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence 17*(5), 449–460.

(Kumaran et Allan, 2004) G. Kumaran et J. Allan, 2004. Text classification and named entities for new event detection. Dans les actes de *ACM SIGIR'04*, ACM Press, 297–304.

(Kupiec et al., 1995a) J. Kupiec, J. Pedersen, et F. Chen, 1995a. A trainable document summarizer. Dans les actes de *SIGIR'95*, 68–73s. ACM Press.

(Kupiec et al., 1995b) J. Kupiec, J. Pederson, et F. Chen, 1995b. A trainable document summarizer. Dans les actes de *SIGIR'95*, 68–73. ACM Press.

(Kwok, 1989) K. L. Kwok, 1989. A neural network for probabilistic information retrieval. Dans les actes de *12th annual international ACM SIGIR conference on Research and development in information retrieval*, 21–30. 75338.

(Kwok, 1995) K. L. Kwok, 1995. A network approach to probabilistic information retrieval. *ACM Trans. Inf. Syst. 13*(3), 324–353. 203067.

(Kwok, 2005) K. L. Kwok, 2005. An attempt to identify weakest and strongest queries. Dans les actes de *SIGIR'06*, Salvador. ACM Press.

(Labadié et al., 2005) A. Labadié, Y. Romero, et L. Sitbon, 2005. Segmentation et classification : deux politiques complémentaires. Dans les actes de *atelier DEFT à TALN'05*, Dourdan, France, 189–192.

(Lalmas et Tombros, 2007) M. Lalmas et A. Tombros, 2007. Evaluating xml retrieval effectiveness at inex. *SIGIR Forum (ACM Press) 41*(1), 40–57.

(Lavenus et al., 2004a) K. Lavenus, J. Grivolla, L. Gillard, et P. Bellot, 2004a. Deux pistes complémentaires pour améliorer l'appariement question réponse. Dans les actes de *11è conférence TALN*, Fez, Maroc, 403–412.

(Lavenus et al., 2004b) K. Lavenus, J. Grivolla, L. Gillard, et P. Bellot, 2004b. Question-answer matching : two complementary methods. Dans les actes de *7è conférence RIAO*, Avignon, France, 244–259.

(Lavenus et Lapalme, 2002) K. Lavenus et G. Lapalme, 2002. Evaluation des systèmes de question réponse. aspects méthodologiques. *Traitement Automatique des Langues (TAL) 43*(3), 181–208.

(Lee et al., 2001) G. G. Lee, J. Seo, S. Lee, H. Jung, B. H. Cho, C. Lee, B. K. Kwak, J. Cha, D. Kim, et J. An, 2001. Siteq : Engineering high performance qa system using lexico-semantic pattern matching and shallow nlp. Dans les actes de *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*, 442–451. NIST Special Publication 500-250.

(Lee, 1997) J. H. Lee, 1997. Analyses of multiple evidence combination. Dans les actes de *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, Philadelphia, Pennsylvania, United States, 267–276. ACM. 258587.

(Leek et al., 2002) T. Leek, R. Schwartz, et S. Sista, 2002. Probabilistic approaches to topic detection and tracking. Dans J. Allan (Ed.), *Topic Detection and Tracking – Event-based Information Organization*.

(Lehnert, 1978) W. Lehnert, 1978. *The process of question answering : A computer simulation of cognition.* Lawrence Erlbaum Associates.

(Lespinasse et al., 1999) K. Lespinasse, P. Kremer, D. Schibler, et L. Schmitt, 1999. Evaluation des outils d'accès à l'information textuelle, les expériences américaine (trec) et française (amaryllis). *Langues, John Libbey 2*(2), 100–109.

(Li et Roth, 2002) X. Li et D. Roth, 2002. Learning question classifiers. Dans les actes de *COLING'02*.

(Lillis et al., 2006) D. Lillis, F. Toolan, R. Collier, et J. Dunnion, 2006. Probfuse : a probabilistic approach to data fusion. Dans les actes de *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, Seattle, Washington, USA, 139–146. ACM. 1148197.

(Lin, 2005) J. Lin, 2005. Evaluation of resources for question answering evaluation. Dans les actes de *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, Salvador, Brazil, 392–399. ACM Press. 1076102.

(Lin, 2007a) J. Lin, 2007a. An exploration of the principles underlying redundancy-based factoid question answering. *ACM Trans. Inf. Syst. 25*(2), 4–53. 1229180.

(Lin, 2007b) J. Lin, 2007b. Is question answering better than information retrieval ? a task-based evaluation framework for question series. *Proceedings of the 2007 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2007)*, 212–219.

(Lin et al., 2003) J. Lin, D. Quan, V. Sinha, K. Bakshi, D. Huynh, B. Katz, et D. R. Karger, 2003. What makes a good answer ? the role of context in question answering. Dans les actes de *Human-Computer Interaction (INTERACT 2003)*, Zurich, Switzerland.

(Lin et Smucker, 2008) J. Lin et M. D. Smucker, 2008. How do users find things with pubmed ? : towards automatic utility evaluation with user simulations. Dans les actes de *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, Singapore, Singapore. ACM. 1390340 19-26.

(Litkowski, 2002) K. C. Litkowski, 2002. Cl research experiments in trec-10 question answering. Dans les actes de *The Tenth Text Retrieval Conference (TREC 2001)*, 500–250. NIST Special Publication 500-250.

(Liu et Croft, 2002) X. Liu et W. B. Croft, 2002. Passage retrieval based on language models. Dans les actes de *Proceedings of the eleventh international conference on Information and knowledge management*, McLean, Virginia, USA, 375–382. ACM. 584854 375-382.

(Loosemore, 1991) R. P. W. Loosemore, 1991. A neural net model of normal and dyslexic spelling. Dans les actes de *International Joint Conference on Neural Networks*, Seattle, USA, 231–236.

(Lucas et Nadine, 2004) E. G. Lucas et Nadine, 2004. La détection automatique des citations et des locuteurs dans les textes informatifs. Dans les actes de *Le discours rapporté dans tous ses états : question de frontières*, 410–418. L'Harmattan, Paris.

(Lytinen et Tomuro, 2002) S. Lytinen et N. Tomuro, 2002. The use of question types to match questions in faqfinder. Dans les actes de *AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases*, 46–53.

(Lété et Ducrot, 2007) B. Lété et S. Ducrot, 2007. La perception du mot écrit chez l'apprenti lecteur et l'enfant dyslexique : Evaluation en fovea et en parafovea. Dans E. Demont, J.-E. Gombert, et M. N. Metz-Lutz (Eds.), *Acquisition du langage : approche intégrée*, 125–172. SOLAL.

(Lété et al., 2004) B. Lété, L. Sprenger-Charolles, et P. Colé, 2004. Manulex : A grade-level lexical database from french elementary-school readers. *Behavior Research Methods, Instruments and Computers 36*, 156–166.

(Malaisé et al., 2005) V. Malaisé, T. Delbecque, et P. Zweigenbaum, 2005. Recherche en corpus de réponses à des questions définitoires. Dans les actes de *TALN*, Dourdan, France.

(Manmatha et al., 2001) R. Manmatha, T. Rath, et F. Feng, 2001. Modeling score distributions for combining the outputs of search engines. Dans les actes de *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, New Orleans, Louisiana, United States, 267–275. ACM. 384005.

(Marchand et Friedman, 2005) Y. Marchand et R. Friedman, 2005. Impaired oral reading in two atipical dyslexics : A comparison with a computational lexical-analogy model. *Brain and Language 93*, 255–266.

(Marteau et al., 1999) P.-F. Marteau, C. de Loupy, P. Bellot, et M. El-Bèze, 1999. Le traitement automatique du langage naturel, outil d'assistance à la fonction d'intelligence economique. *Systèmes et Sécurité 5*(4), 8–41.

(Marton, 2006) G. Marton, 2006. Nuggeteer : Automatic nugget-based evaluation using descriptions and judgements. Rapport technique.

(Maybury, 2004) M. T. Maybury, 2004. *New Directions in Question Answering*. The MIT Press.

(Mehler et Dupoux, 1992) J. Mehler et E. Dupoux, 1992. *Naître humain*. Paris : Odile Jacob.

(Meir et Ratsch, 2003) R. Meir et G. Ratsch, 2003. An introduction to boosting and leveraging. Dans les actes de *Advanced Lectures on Machine Learning*, 119–184. Springer.

(Messerschmitt et Flohic, 2002) Messerschmitt et Flohic, 2002. *Les troubles d'acquisitions du langage - la dyslexie*.

(Miller, 1995) G. Miller, 1995. Wordnet : A lexical database for english. *Communications of the ACM 38*(11), 39–42.

(Miller et al., 1990) G. Miller, R. Beckwith, C. Fellbaum, D. Gross, et K. Miller, 1990. Introduction to wordnet : An online lexical database. *International Journal of Lexicography 3*(4), 235–244.

(Minel, 2004) J.-L. Minel, 2004. L'évaluation des systèmes de résumé automatique. Dans S. Chaudiron (Ed.), *Evaluation des systèmes de traitement de l'information*, 171–186. Paris : Hermès.

(Minsky et Papert, 1969) M. Minsky et S. Papert, 1969. *Perceptrons. An Introduction to Computational Geometry*, Volume 165 de *Science*. Cambridge, Mass. : MIT Press.

(Mizzaro, 1997) S. Mizzaro, 1997. Relevance : the whole history. *Journal of the American Society for Information Science 48*(9), 810–832.

(Moldovan et al., 2003) D. Moldovan, C. Clark, S. Harabagiu, et S. Maiorano, 2003. Cogex : A logic prover for question answering. Dans les actes de *North American Chapter of the Association for Computational Linguistics, Human Language Technology Conference (HLT-NAACL)*, Edmonton, Canada.

(Moldovan et al., 2002) D. Moldovan, S. Harabagiu, S. Girju, P. Morarescu, F. Lacutusu, A. Novischi, A. Badulescu, et O. Bolohan, 2002. Lcc tools for question answering. Dans les actes de *11th Text Retrieval Conference (TREC-2002)*. NIST Special Publication 500-251.

(Moldovan et al., 2003) D. Moldovan, M. Pasca, S. Harabagiu, et M. Surdeanu, 2003. Performance issues and error analysis in an open-domain question answering system. *ACM Transactions on Information Systems (TOIS) 21*(2), 133–154.

(Molla et al., 2000) D. Molla, R. Schwitter, M. Hess, et R. Fournier, 2000. Extrans, an answer extraction system. *Traitement Automatique des Langues (TAL) 41*(2), 496–522.

(Monaghan et Shillcock, 2008) P. Monaghan et R. Shillcock, 2008. Hemispheric dissociation and dyslexia in a computational model of reading. *Brain and Language In Press, Corrected Proof*.

(Mondary et al., 2007) T. Mondary, A. Bouffier, et A. Nazarenko, 2007. Between browsing and search, a new model for navigating through large documents. Dans les actes de *Second European Cognitive Science Conference*, Delhi, India.

(Montague et Aslam, 2001) M. Montague et J. A. Aslam, 2001. Metasearch consistency. Dans les actes de *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, New Orleans, Louisiana, United States, 386–387. ACM. 384030.

(Monz, 2003) C. Monz, 2003. *From Document Retrieval to Question Answering*. Thèse de Doctorat, Institute for Logic, Language and Computation.

(Moreau et al., 2007) F. Moreau, V. Claveau, et P. Sébillot, 2007. Intégrer plus de connaissances linguistiques en recherche d'information peut-il augmenter les performances des systèmes ? Dans les actes de *4ème Conférence en recherche d'informations et applications, (CORIA'07)*, Saint-Etienne (France).

(Moriceau, 2007) V. Moriceau, 2007. *Intégration de données dans un système question-réponse sur le Web*. Thèse de doctorat, Université Paul Sabatier.

(Morrison et Ellis, 1995) C. Morrison et A. Ellis, 1995. Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology : Learning, Memory and Cognition 21*(1), 116–133.

(Morton, 1969) J. Morton, 1969. Interaction of information in word recognition. *Psychological Review 76*, 165–178.

(Mothe et Tanguy, 2005) J. Mothe et L. Tanguy, 2005. Linguistic features to predict query difficulty - a case study on previous trec campaigns. Dans les actes de *SIGIR'05*, Salvador, 7–10. ACM Press.

(Mourad et Desclès, 2004) G. Mourad et J.-P. Desclès, 2004. Identification et extraction automatique des informations citationnelles dans un texte. Dans les actes de *Le discours rapporté dans tous ses états : question de frontières*. L'Harmattan, Paris.

(Munos, 1997) R. Munos, 1997. *L'apprentissage par renforcement, étude du cas continu*. Thèse de doctorat, Ecole des Hautes Etudes en Sciences Sociales.

(Munteanu et al., 2006) C. Munteanu, G. Penn, R. Baecker, et Y. Zhang, 2006. Automatic speech recognition for webcasts : how good is good enough and what to do when it isn't. Dans les actes de *Proceedings of the 8th international conference on Multimodal interfaces*, Banff, Alberta, Canada, 39–42. ACM. 1181005 39-42.

(Murray et al., 2005) G. Murray, S. Renals, et J. Carletta, 2005. Extractive summarization of meeting recordings. Dans les actes de *9th European Conference on Speech Communication and Technology (Eurospeech'05)*, 593–596.

(New et al., 2006) B. New, L. Ferrand, C. Pallier, et M. Brysbaert, 2006. Reexamining the word length effect in visual word recognition : New evidence from the english lexicon project. *Psychonomic Bulletin and Review 13*(1), 45–52.

(New et al., 2001) B. New, C. Pallier, L. Ferrand, et R. Matos, 2001. Une base de données lexicales du français contemporain sur internet : Lexique. *L'Année Psychologique 101*, 447–462.

(Nietzsche, 1888) F. Nietzsche, 1888. *Le Crépuscule des idoles*.

(Nocéra et al., 2004) P. Nocéra, C. Fredouille, G. Linares, D. Matrouf, S. Meignier, J.-F. Bonastre, D. Massonié, et F. Béchet, 2004. The lia's french broadcast news transcription system. Dans les actes de *SWIM : Lectures by Masters in Speech Processing*, Hawaii (USA).

(Nyberg, 2003) E. Nyberg, 2003. Piqasso 2002. Dans les actes de *The Eleventh Text Retrieval Conference (TREC 2002)*, Volume NIST Special Publication SP 500-251.

(Ogilvie et Callan, 2003) P. Ogilvie et J. Callan, 2003. Combining document representations for known item search. Dans les actes de *26th ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

(Oh et al., 2000) H.-J. Oh, S. Myaeng, et M. Lee, 2000. A practical hypertext categorization method using links and incrementally available class information. Dans les actes de *ACM SIGIR 2000*, 264–271.

(O'Regan, 1990) J. O'Regan, 1990. Eye movements and cognitive processes. Dans E. Kowler (Ed.), *Eye movements and their role in visual and cognitive processes*. North-Holland : Elsevier.

(Pasca et Harabagiu, 2001) M. Pasca et S. Harabagiu, 2001. The informative role of wordnet in open-domain question anwering. Dans les actes de *NAACL 2001*.

(Paulesu et al., 2001) E. Paulesu, J. Demonet, F. Fazio, E. Mc Crory, V. Chanoine, N. Brunswick, S. Cappa, M. Habib, C. Frith, et U. Frith, 2001. Dyslexia : cultural diversity and biological unity. *Science 291*(5511), 2165–2167.

(Pedler, 2001) J. Pedler, 2001. The detection and correction of real-word spelling errors in dyslexic text. Dans les actes de *4th Annual CLUK Colloquium*, 115–119.

(Peereman et al., 2007) R. Peereman, B. Lété, et L. Sprenger-Charolles, 2007. Manulex-infra : Distributional characteristics of grapheme-phoneme mappings, infra-lexical and lexical units in child-directed written material. *Behavior Research Methods 39*(3), 579–589.

(Perea et Rosa, 2002) M. Perea et E. Rosa, 2002. Does the 'whole-word shape' play a role in visual word recognition ? *Perception and Psychophysics 64*(5), 785–794.

(Pereira et Ziviani, 2003) A. R. Pereira et N. Ziviani, 2003. Syntactic similarity of web documents. Rapport technique, Department of Computer Science, Federal University of Minas Gerais Belo Horizonte, Brazil.

(Perry et al., 2007) C. Perry, J. C. Ziegler, et M. Zorzi, 2007. Nested incremental modelling in the development of computational theories : The cdp+ model of reading aloud. *Psychological Review 114*(2), 273–315.

(Pevzner et Hearst, 2002) L. Pevzner et M. Hearst, 2002. A critique and improvement of an evaluation metric for text segmentatin. *Computational Linguistics*, 19–36.

(Pexman et al., 2004) P. Pexman, Y. Hino, et S. Lupker, 2004. Semantic ambiguity and the process of generating meaning from print. *Journal of Experimental Psychology : Learning, Memory and Cognition 30*, 1252–1270.

(Piaget et Chomsky, 1979) J. Piaget et N. Chomsky, 1979. *Théories du langage, théories de l'apprentissage - Le débat entre Jean Piaget et Noam Chomsky*. Essais, Points. Paris : Seuil.

(Piwowarski et Gallinari, 2003) B. Piwowarski et P. Gallinari, 2003. A machine learning model for information retrieval with structured documents. Dans P. Petner (Ed.), *Machine Learning and Data Mining in Pattern Recognition (MLDM'03)*, Leipzig, 425–438. Springer-Verlag.

(Plamondon et al., 2003) L. Plamondon, G. Lapalme, et L. Kosseim, 2003. The quantum question answering system at trec 11. Dans les actes de *The Eleventh Text Retrieval Conference (TREC 2002)*, Volume NIST Special Publication SP 500-251.

(Poibeau, 2003) T. Poibeau, 2003. *Extraction automatique d'information. Du texte brut au web sémantique*. Paris : Hermes.

(Poibeau et Vilnat, 2008) T. Poibeau et A. Vilnat, 2008. Traitement automatique des langues et système de question-réponse. Dans B. Grau et J.-P. Chevallet (Eds.), *La recherche d'informations précises : apprentissage, traitement automatique de la langue et connaissances pour les systèmes de question-réponse*, 105–132. Paris : Hermès.

(Polanyi, 1987) L. Polanyi, 1987. Keeping it all straight : Interpreting narrative time in real discourse. *WCCFL* (6), 229–245.

(Pollock, 2007) J.-Y. Pollock, 2007. La grammaire générative et le programme minimaliste. Dans J. Bricmont et J. Franck (Eds.), *Chomsky (Les Cahiers de l'Herne)*. Paris : Editions de l'Herne.

(Ponte et Croft, 1998) J. M. Ponte et W. B. Croft, 1998. A language modeling approach to information retrieval. Dans les actes de *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, Melbourne, Australia, 275–281. ACM Press.

(Poulard et al., 2008) F. Poulard, T. Waszak, N. Hernandez, et P. Bellot, 2008. Repérage de citations, classification des styles de discours et identification des constituants citationnels en écrits journalistiques. Dans les actes de *TALN 2008*, Avignon, France.

(Prager et al., 1999) J. Prager, E. Brown, D. R. Radev, et K. Czuba, 1999. One search engine or two for question-answering. Dans les actes de *Proceedings of the TREC-9 Conference*, 235–240. NIST Special Publication 500-249.

(Quinlan, 1992) J. Quinlan, 1992. *C4.5 : Programs for Machine Learning*. Morgan Kaufmann.

(Radev et al., 2002) D. Radev, W. Fan, H. Qi, H. Wu, et A. Grewal, 2002. Probabilistic question answering on the web. Dans les actes de *Proceedings International WWW Conference*, Honolulu, Hawaii, USA.

(Radev et al., 2005) D. Radev, W. Fan, H. Qi, H. Wu, et A. Grewal, 2005. Probabilistic question answering on the web. *Journal of the American Society for Information Science and Technology 56*(6), 571–583.

(Radev et al., 2001) D. R. Radev, H. Qi, Z. Zheng, S. Blair-Goldensohn, Z. Zhang, W. Fan, et J. Prager, 2001. Mining the web for answers to natural language questions. Dans les actes de *Proceedings of the tenth international conference on Information and knowledge management*, Atlanta, Georgia, USA, 143–150. ACM Press. 502610 143-150.

(Ranjan et al., 2006) A. Ranjan, R. Balakrishnan, et M. Chignell, 2006. Searching in audio : the utility of transcripts, dichotic presentation, and time-compression. Dans les actes de *Proceedings of the SIGCHI conference on Human Factors in computing systems*, Montr&#233 ;al, Qu&#233 ;bec, Canada, 721–730. ACM. 1124879 721-730.

(Rasolofo et Savoy, 2003) Y. Rasolofo et J. Savoy, 2003. Term proximity scoring for keyword-based retrieval systems. Dans les actes de *Proceedings 25th European Conference on IR Research (ECIR 2003)*, 207–218.

(Rastle et Coltheart, 1998) K. Rastle et M. Coltheart, 1998. Whammy and double whammy : Length effects in nonword naming. *Psychonomic Bulletin and Review 5*, 277–282.

(Raymond et al., 2002) C. Raymond, P. Bellot, et M. El-Bèze, 2002. Enrichissement de requêtes pour la recherche documentaire selon une classification non supervisée. Dans les actes de *13è Congrès Francophone AFRIF-RFIA de Reconnaissance des Formes et d'Intelligence Artificielle*, 625–632.

(Rey et al., 1998) A. Rey, A. Jacobs, F. Schmidt-Weigand, et J. C. Ziegler, 1998. A phoneme effect in visual word recognition. *Cognition 68*, B71–B80.

(Rey et al., 2000) A. Rey, J. C. Ziegler, et A. Jacobs, 2000. Graphemes are perceptual reading units. *Cognition 74*, 1–12.

(Rey et al., 2001) V. Rey, C. Sabater, et C. de Cormis, 2001. Un déficit de la conscience morphologique comme prédicteur de la dysorthographie chez l'enfant présentant une dyslexie phonologique. *Glossa* (78), 4–20.

(Reynar, 2000) J. Reynar, 2000. *Topic segmentation : Algorithms and applications*. Phd thesis, University of Pennsylvania, Seattle, WA, USA.

(Rinaldi et al., 2003) F. Rinaldi, J. Dowdall, M. Hess, K. Kaljurand, et M. Karlsson, 2003. The role of technical terminology in question answering. Dans les actes de *Terminologie et Intelligence Artificielle, TIA 2005*, Strasbourg, France.

(Rinaldi et al., 2004) F. Rinaldi, M. Hess, J. Dowdall, D. Molla, et R. Schwitter, 2004. Question answering in terminology-rich domains. Dans M. T. Maybury (Ed.), *New Directions in Question Answering*, 71–82. AAAI Press.

(Rizzi, 2007) L. Rizzi, 2007. L'acquisition de la langue et la faculté de langage. Dans J. Bricmont et J. Franck (Eds.), *Chomsky (Les Cahiers de l'Herne)*, 147–156. Paris : Editions de l'Herne.

(Rizzolatti et Sinigaglia, 2008) G. Rizzolatti et C. Sinigaglia, 2008. *Les neurones miroires*. Paris : Odile Jacob (pour la trad. française).

(Robertson et Sparck-Jones, 1976) S. Robertson et K. Sparck-Jones, 1976. Relevance weighting of search terms. *Journal of the American Society for Information Science 27*(3), 129–146.

(Robertson et al., 1994) S. Robertson, S. Walker, M. Hancock-Beaulieu, et M. Gatford, 1994. Okapi at trec-3. Dans les actes de *Text REtrieval Conference TREC-3*, Volume NIST special publication n° 500-225, 109–126.

(Robertson et al., 2004) S. Robertson, H. Zaragoza, et M. Taylor, 2004. Simple bm25 extension to multiple weighted fields. Dans les actes de *Proceedings of the thirteenth ACM international conference on Information and knowledge management 1-58113-874-1*, Washington, D.C., USA, 42–49. ACM. http ://doi.acm.org/10.1145/1031171.1031181.

(Rocchio, 1971) J. Rocchio, 1971. Relevance feedback in information retrieval. Dans G. Salton (Ed.), *The SMART Retrieval Storage and Retrieval System*, 313–323. Englewood Cliffs, N.J. : Pentice Hall Inc.

(Rossignol, 2001) C. Rossignol, 2001. *Inadaptation, Handicap, Invalidation ? Histoire et étude critique des notions, de la terminologie et des pratiques dans le champ professionnel de l'Éducation spéciale.* Doctorat d'état : Université Louis Pasteur - Strasbourg I - Presses Universitaires du Septentrion.

(Rubino et Lavalley, 2008) R. Rubino et R. Lavalley, 2008. Mise en oeuvre de méthodes de tal afin d'aider les enseignants dans l'élaboration d'exercices d'apprentissage de la lecture pour enfants dyslexiques. Dans les actes de *Majecstic'08*, Marseille.

(Rumelhart et al., 1985) D. Rumelhart, G. Hinton, et R. Williams, 1985. Learning internal representations by error propagation. Rapport technique, University of California.

(Régnier, 2007) A. Régnier, 2007. *Analyse et représentation formelle du discours pour la classification automatique des textes*. Thèse de doctorat, Université de Provence Aix-Marseille I.

(Sabbah, 1996) G. Sabbah, 1996. Le « carnet d'esquisses » : une mémoire interprétative dynamique. Dans les actes de *RFIA*, 1096–1105.

(Sabbah, 2006a) G. Sabbah, 2006a. Compréhension automatique des langues : où va-t-on ? où pourrait-on aller ? Dans G. Sabbah (Ed.), *Compréhension des langues et interaction*. Paris : Hermès.

(Sabbah, 2006b) G. Sabbah, 2006b. *Compréhension des langues et interaction*. Paris : Hermès.

(Sabbah et Popescu-Belis, 1999) G. Sabbah et A. Popescu-Belis, 1999. Experiments in language acquisition by artificial systems. Dans les actes de *MIND-4*, Dublin, 1–10.

(Salton, 1975) G. Salton, 1975. *Dynamic information and library processing*. Englewood Cliffs, USA.

(Salton et al., 1993) G. Salton, J. Allan, et C. Buckley, 1993. Approaches to passage retrieval in full text information systems. Dans les actes de *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, Pittsburgh, Pennsylvania, United States, 49–58. ACM. 160693.

(Salton et al., 1983) G. Salton, E. Fox, et H. Wu, 1983. Extended boolean information retrieval. *Communications of the ACM 31*(2), 1002–1036.

(Samuelson, 1994) P. Samuelson, 1994. Self-plagiarism or fair use ? *Communications of the ACM 37*(8), 21–25.

(Saracevic, 1995) T. Saracevic, 1995. Evaluation of evaluation in information retrieval. Dans les actes de *18è ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle WA, USA, 138–145.

(Savoy, 2003) J. Savoy, 2003. Modèles en recherche d'information. Dans E. Gaussier et M.-H. Stéfanini (Eds.), *Assistance intelligente à la recherche d'informations*, 31–70. Paris : Hermès.

(Savoy, 2006) J. Savoy, 2006. Un regard statistique sur l'évaluation de performance : l'exemple de clef 2005. Dans les actes de *3è conférence en Recherche d'Informations et Applications (CORIA 2006)*, Lyon, 73–84.

(Savoy et al., 1997) J. Savoy, A. Calvé, et D. Vrajitoru, 1997. Report on the trec-5 experiment. Dans les actes de *TREC-5*, 489–502. NIST Special Publication.

(Schleimer et al., 2003) S. Schleimer, D. S. Wilkerson, et A. Aiken, 2003. Winnowing : Local algorithms for document fingerprinting. Dans les actes de *2003 ACM SIGMOD international conference on Management of data*, 76–85.

(Schultz et Liberman, 2002) J. Schultz et M. Liberman, 2002. Towards a 'universal dictionary' for multi-language information retrieval applications. Dans J. Allan (Ed.), *Topic Detection and Tracking – Event-based Information Organization*. Kluwer.

(Scott et Galan, 1998) N. Scott et J. Galan, 1998. The total access system. Dans les actes de *1998 CSUN Conference*.

(Seidenberg et McClelland, 1989) M. Seidenberg et J. McClelland, 1989. A distributed developmental model of word recognition and naming. *Psychological Review 96*, 523–568.

(Seidenberg et al., 1984) M. Seidenberg, G. Waters, M. Barnes, et M. Tanenhaus, 1984. When does irregular spelling or pronunciation influence word recognition ? *Journal of Verbal Learning and Verbal Behavior 23*, 383–404.

(Sejnowski et Rosenberg, 1987) T. J. Sejnowski et C. R. Rosenberg, 1987. Parallel networks that learn to pronounce english text. *Complex Systems 1*(1), 145–168.

(Sekine et al., 2002) S. Sekine, K. Sudo, et C. Nobata, 2002. Extended named entity hierarchy. *Proceedings of the LREC-2002 Conference*, 1818–1824.

(Senator, 2005) T. Senator, 2005. Link mining applications : Progress and challenges. *ACM SIGKDD Explorations 7*(2), 76–83.

(Singhal et al., 2000) A. Singhal, S. Abney, M. Bacchiani, M. Collins, D. Hindle, et F. Pereira, 2000. At&t at trec-8.

(Singhal et al., 1996) A. Singhal, G. Salton, M. Mitra, et C. Buckley, 1996. Document length normalization. *Information Processing and Management 32*(5), 619–633.

(Sitbon, 2007) L. Sitbon, 2007. *Robustesse en recherche d'information -Application à l'accessibilité aux personnes handicapées*. Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse.

(Sitbon et Bellot, 2004a) L. Sitbon et P. Bellot, 2004a. Adapting and comparing linear segmentation methods for french. Dans les actes de *7è conférence en Recherche d Information Assistée par Ordinateur (RIAO)*, Avignon, France, 623–637.

(Sitbon et Bellot, 2004b) L. Sitbon et P. Bellot, 2004b. Evaluation de méthodes de segmentation thématique linéaire non supervisées après adaptation au français. Dans les actes de *TALN*, Fez, Maroc, 441–450.

(Sitbon et Bellot, 2005a) L. Sitbon et P. Bellot, 2005a. Liaseg, segmentation thématique par chaînes lexicales pondérées. Dans les actes de *TALN 2005*, Dourdan, France.

(Sitbon et Bellot, 2005b) L. Sitbon et P. Bellot, 2005b. Segmentation thématique par chaînes lexicales pondérées. Dans les actes de *TALN 2005*, Dourdan (France).

(Sitbon et Bellot, 2006) L. Sitbon et P. Bellot, 2006. Tools and methods for topic segmentation of texts and contextual evaluation. Dans les actes de *Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genova, Italie.

(Sitbon et Bellot, 2007) L. Sitbon et P. Bellot, 2007. Topic segmentation using weighted lexical links (wll). Dans les actes de *ACM SIGIR 07*, Amsterdam, Pays-Bas.

(Sitbon et Bellot, 2008a) L. Sitbon et P. Bellot, 2008a. How to cope with questions typed by dyslexic users. Dans les actes de *Proceedings of the second workshop on Analytics for noisy unstructured text data (AND at SIGIR 2008)*, Singapore. ACM. 1390752 1-8.

(Sitbon et Bellot, 2008b) L. Sitbon et P. Bellot, 2008b. A readability measure for an information retrieval process adapted to dyslexics. Dans les actes de *Second international workshop on Adaptive Information Retrieval (AIR 2008) (in conjunction with IIiX 2008)*, Londres.

(Sitbon et al., 2007a) L. Sitbon, P. Bellot, et P. Blache, 2007a. Phonetic based sentence level rewriting of questions typed by dyslexic spellers in an information retrieval context. Dans les actes de *Interspeech 2007*, Antwerpen (Belgique).

(Sitbon et al., 2007b) L. Sitbon, P. Bellot, et P. Blache, 2007b. Traitements phrastiques phonétiques pour la réécriture de phrases dysorthographiées. Dans les actes de *TALN 2007*, Toulouse.

(Sitbon et al., 2008a) L. Sitbon, P. Bellot, et P. Blache, 2008a. Evaluating robustness of question answering system through a corpus of real-life questions. Dans les actes de *6th edition of the Language Resources and Evaluation Conference (LREC 2008)*, Marrakech (Maroc).

(Sitbon et al., 2008b) L. Sitbon, P. Bellot, et P. Blache, 2008b. Lisibilité et recherche d'information : vers une meilleure accessibilité. Dans les actes de *5è Conférence en Recherche d Informations et Applications (CORIA)*, Trégastel, France.

(Sitbon et al., 2008c) L. Sitbon, P. Bellot, et P. Blache, 2008c. Éléments pour adapter les systèmes de recherche d'information aux dyslexiques. *Traitement Automatique des Langues (TAL) 48*(2), 123–147.

(Sitbon et al., 2007) L. Sitbon, J. Grivolla, L. Gillard, P. Bellot, et P. Blache, 2007. Vers une prédiction automatique de la difficulté d'une question en langue naturelle. Dans les actes de *13ième conférence Traitement Automatique des Langues Naturelles (TALN)*, Louvain (Belgique), 337–346.

(Somasundaran et al., 2007) S. Somasundaran, T. Wilson, J. Wiebe, et V. Stoyanov, 2007. Qa with attitude : Exploiting opinion type analysis for improving question answering in on-line discussions and the news. Dans les actes de *International Conference on Weblogs and Social Media (ICWSM'07)*.

(Southwood et Chatterjee, 2000) M. H. Southwood et A. Chatterjee, 2000. The interaction of multiple routes in oral reading : Evidence from dissociations in naming and oral reading in phonological dyslexia. *Brain and Language 72*(1), 14–39.

(Sparck-Jones, 1988) K. Sparck-Jones, 1988. A look back and a look forward. Dans les actes de *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, Grenoble, France, 13–29. ACM. 62438.

(Spooner, 1998) R. Spooner, 1998. *A spelling checker for dyslexic users : user modelling for error recovery*. Thèse de Doctorat, University of York.

(Sprenger-Charolles et Colé, 2003) L. Sprenger-Charolles et P. Colé, 2003. *Lecture et dyslexie - Approche cognitive*. Paris : Dunod.

(Spriet et El-Bèze, 1998) T. Spriet et M. El-Bèze, 1998. Introduction of rules into a stochastic approach for language modelling. Dans K. Ponting (Ed.), *NATO ASI Series F*, Volume 196, 350–355.

(Staab et Maedche, 2001) S. Staab et A. Maedche, 2001. Ontology learning for the semantic web. *IEEE Intelligent Systems, Special Issue on the Semantic Web 16*(2).

(Stoyanov et Cardie, 2006) V. Stoyanov et C. Cardie, 2006. Toward opinion summarization : Linking the sources. Dans les actes de *COLING-ACL 2006 Workshop on Sentiment and Subjectivity in Text*.

(Stroop, 1935) J. Stroop, 1935. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology 18*, 643–662.

(Sutcliffe, 2003) R. Sutcliffe, 2003. Question answering using the dlt system at trec 2002. Dans les actes de *The Eleventh Text Retrieval Conference (TREC 2002)*, Volume NIST Special Publication SP 500-251.

(Tellex et al., 2003) S. Tellex, B. Katz, J. Lin, A. Fernandes, et G. Marton, 2003. Quantitative evaluation of passage retrieval algorithms for question answering. Dans les actes de *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, Toronto, Canada, 41–47. ACM Press. 860445.

(Temple et al., 2003) E. Temple, G. Deutsch, R. Poldrack, S. Miller, P. Tallal, M. Merzenich, et J. Gabrieli, 2003. Neural deficits in children with dyslexia ameliorated by behavorial remediation : evidence from functional mri. *National Academia of Science, USA 100*(5), 13907–13912.

(Tomuro, 2003) N. Tomuro, 2003. Interrogative reformulation patterns and acquisition of question paraphrases. Dans les actes de *The Second International Workshop on Paraphrasing : Paraphrase Acquisition and Applications*, 33–40.

(Tomuro et Lytinen, 2004) N. Tomuro et S. Lytinen, 2004. Retrieval models and q and a learning with faq files. Dans M. T. Maybury (Ed.), *New Directions in Question Answering*, 183–194. The MIT Press.

(Torres-Moreno et al., 2005) J.-M. Torres-Moreno, P. Velazquez-Morales, et J. Meunier, 2005. Cortex : un algorithme pour la condensation automatique de textes. *ARCo 2*.

(Toutanova et Moore, 2002) K. Toutanova et R. C. Moore, 2002. Pronunciation modeling for improved spelling correction. Dans les actes de *40th annual meeting of ACL*, Philadelphia, USA, 144–151.

(Usunier et al., 2008a) N. Usunier, M. Amini, et P. Gallinari, 2008a. Apprentissage et systèmes de question-réponse. Dans B. Grau et J.-P. Chevallet (Eds.), *La recherche d'informations précises*, 69–104. Hermes-Lavoisier.

(Usunier et al., 2008b) N. Usunier, M. Amini, et P. Gallinari, 2008b. Apprentissage et systèmes de question-réponse. Dans B. Grau et J.-P. Chevallet (Eds.), *La recherche d'informations précises*, 69–104. Paris : Hermès.

(Vicedo et al., 2001) J. L. Vicedo, A. Ferrandez, et F. Llopis, 2001. University of alicante at trec-10. Dans les actes de *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*, 510–518. NIST Special Publication 500-250.

(Voorhees, 2000) E. M. Voorhees, 2000. Overview of the trec-9 question answering track. Dans les actes de *The Ninth Text Retrieval Conference (TREC-9)*, 71–80. NIST Special Publication 500-249.

(Voorhees, 2001) E. M. Voorhees, 2001. Overview of the trec 2001 question answering track. Dans les actes de *The Tenth Text Retrieval Conference (TREC 2001)*, 42–50. NIST Special Publication 500-251.

(Voorhees, 2002) E. M. Voorhees, 2002. Overview of trec 2002. Dans les actes de *The Eleventh Text Retrieval Conference (TREC 2002)*. NIST Special Publication.

(Voorhees, 2003) E. M. Voorhees, 2003. Evaluating the evaluation : a case study using the trec 2002 question answering track. Dans les actes de *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, Edmonton, Canada, 181–188. Association for Computational Linguistics. 1073479.

(Voorhees et Harman, 2005) E. M. Voorhees et D. K. Harman, 2005. *TREC - Experiment and Evaluation in Information Retrieval*. The MIT Press.

(Voorhees Ellen, 2003) M. Voorhees Ellen, 2003. Overview of the trec 2003 robust retrieval track. Dans les actes de *TREC 12*, Gaithersburg, USA, 69–77s.

(Voorhees Ellen et Harman, 1999) M. Voorhees Ellen et D. K. Harman, 1999. Overview of the eighth text retrieval conference (trec-8). Dans les actes de *The Eighth Text REtrieval Conference (TREC 8)*, 1–24. NIST Special Publication 500-246.

(Véronis et al., 1998) J. Véronis, P. Di Cristo, F. Courtois, et C. Chaumette, 1998. A stochastic model of intonation for text-to-speech synthesis. *Speech Communication 26*(4), 233–244.

(W3C, 2001) W3C, 2001. How people with disabilities use the web.

(Wang et al., 1994) J. Wang, K. Zhang, K. Jeong, et D. Shasha, 1994. Atbe : a system for approximate tree matching. *IEEE Transactions on Knowledge and Data Engineering 6*(4), 559–571.

(Wang et Si, 2008) M. Wang et L. Si, 2008. Discriminative probabilistic models for passage based retrieval. Dans les actes de *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, Singapore, Singapore, 419–426. ACM. 1390407.

(Waszak, 2007) T. Waszak, 2007. *Détection des citations pour l'identification des plagiats*. Mémoire de master recherche, Université d'Avignon et des Pays de Vaucluse, France.

(Werbos, 1994) P. Werbos, 1994. *The Roots of Backpropagation : From Ordered Derivatives to Neural Networks and Political Forecasting*. Wiley-IEEE.

(Whittaker et al., 2002) S. Whittaker, J. Hirschberg, B. Amento, L. Stark, M. Bacchiani, P. Isenhour, L. Stead, G. Zamchick, et A. Rosenberg, 2002. Scanmail : a voicemail interface that makes speech browsable, readable and searchable. Dans les actes de *Proceedings of the SIGCHI conference on Human factors in computing systems : Changing our world, changing ourselves*, Minneapolis, Minnesota, USA, 275–282. ACM. 503426 275-282.

(Whittaker et al., 1999) S. Whittaker, J. Hirschberg, J. Choi, D. Hindle, F. Pereira, et A. Singhal, 1999. Scan : designing and evaluating user interfaces to support retrieval from speech archives. Dans les actes de *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, Berkeley, California, United States, 26–33. ACM. 312639 26-33.

(Wilkinson, 1994) R. Wilkinson, 1994. Effective retrieval of structured documents. Dans les actes de *SIGIR'94*. ACM Press.

(Wise, 1996) M. Wise, 1996. Yap3 : Improved detection of similarities in computer programs and other texts. Dans les actes de *SIGCSE*, 130–134.

(Witten et Frank, 2005) I. Witten et E. Frank, 2005. *Data Mining : Practical machine learning tools and techniques*. Morgan Kaufmann.

(Wu et al., 2003) L. Wu, X. Huang, Y. Zhou, Y. Du, et L. You, 2003. Fduqa on trec 2003 qa task. Dans les actes de *TREC-12*.

(Xu et Croft, 1996) J. Xu et W. B. Croft, 1996. Query expansion using local and global document analysis. Dans les actes de *ACM-SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Suisse, 4–11. ACM.

(Yang et al., 2000) Y. Yang, T. Ault, et C. Lattimer, 2000. Improving text categorization methods for event tracking. Dans les actes de *ACM SIGIR 2000*, 65–72.

(Yates et al., 2004) M. Yates, L. Locker, et G. Simpson, 2004. The influence of phonological neighbourhood on visual word perception. *Psychonomic Bulletin and Review 11*(3), 452–457.

(Yu et al., 1983) C. Yu, C. Buckley, K. Lam, et G. Salton, 1983. A generalized term dependence model in information retrieval. *Information Technology : Research Development 2*(4), 129–154.

(Yvon et al., 1998) F. Yvon, P. Boula de Mareil, C. d'Alessandro, V. Aubergé, M. Bagein, G. Bailly, F. Béchet, S. Foukia, J. F. Goldman, E. Keller, D. O'Shaughnessy, V. Pagel, F. Sannier, J. Véronis, et B. Zellner, 1998. Objective evaluation of grapheme to phoneme conversion for text-to-speech synthesis in french. *Computer Speech and Language 12*(4), 393–410.

(Zhang et Lee, 2003) D. Zhang et W. Lee, 2003. A language modeling approach to passage question answering. Dans les actes de *TREC*.

(Zhang, 1996) K. Zhang, 1996. A constrained edit distance between unordered labeled trees. *Algorithmica 15*, 205–222.

(Zhang et al., 1992) K. Zhang, R. Statman, et D. Shasha, 1992. On the editing distance between unordered labeled trees. *Inform. Process. Lett. 42*, 133–139.

(Zhang et Callan, 2004) Y. Zhang et J. P. Callan, 2004. Cmu dir supervised tracking report. Dans les actes de *TDT 2004*.

(Ziegler, 2006) J. C. Ziegler, 2006. Do differences in brain activation challenge universal theories of dyslexia ? *Brain and Language 98*(3), 341–343.

(Ziegler et al., 1996) J. C. Ziegler, A. Jacobs, et G. Stone, 1996. Statistical analysis of the bidirectionnal inconsistency of spelling and sound in french. *Behavior Research Methods, Instruments and Computers 28*(4), 504–515.

(Zoccolotti et al., 2005) P. Zoccolotti, M. De Luca, E. Di Pace, F. Gasperini, A. Judica, et D. Spinelli, 2005. Word length effect in early reading and in developmental dyslexia. *Brain and Language 93*(3), 369–373.

(Zweigenbaum, 2003) P. Zweigenbaum, 2003. Question answering in biomedecine. Dans les actes de *EACL 2003 Workshop on Natural Language Processing for Question Answering*, Budapest, Hongrie.

(Zweigenbaum, 2005) P. Zweigenbaum, 2005. Question-answering for biomedicine methods and state of the art. Dans les actes de *MIE 2005 Workshop "Terminologies and Ontologies in Biomedicine : can text mining help ?"*.

(Zweigenbaum et al., 2008) P. Zweigenbaum, B. Grau, M.-L. Ligozat, I. Robba, S. Rosset, X. Tannier, A. Vilnat, et P. Bellot, 2008. Apports de la linguistique dans les systèmes de recherche d'informations précises. *Revue Française de Linguistique Appliquée (RFLA) XIII*(I), 41–62.