# Lisibilité et recherche d'information : vers une meilleure accessibilité

## Intégration de la lisibilité au calcul de la pertinence

**Laurianne Sitbon**[*,**] — **Patrice Bellot**[*] — **Philippe Blache**[**]

[*] *Laboratoire d'Informatique d'Avignon*
*Université d'Avignon*
*339, chemin des Meinajaries*
*Agroparc BP 1228*
*84911 AVIGNON Cedex 9*
*{patrice.bellot, laurianne.sitbon}@univ-avignon.fr*

[**] *Laboratoire Parole et Langage*
*Université de Provence*
*29 av. Robert Schuman*
*13621 Aix en Provence Cedex 1*
*blache@lpl-aix.fr*

RÉSUMÉ. *Dans cet article, nous proposons en premier lieu une mesure de la lisibilité adaptée à des lecteurs dyslexiques en utilisant des caractéristiques issues d'une analyse fine des causes des difficultés de lectures rencontrées. Nous proposons ensuite un cadre pour la prise en compte de la lisibilité dans la mesure de pertinence accordée par les systèmes de recherche d'informations, qui est généralement calculée sur la seule base de la similarité. Ce cadre part de l'hypothèse que les données thématiquement pertinentes existent en nombre suffisant pour qu'on choisisse les plus lisibles. On atteint un taux optimal de prise en compte de la lisibilité de 30% en observant l'évolution des performances dans le cadre de campagnes d'évaluation en recherche documentaire (CLEF) et en résumé (DUC).*

ABSTRACT. *This paper introduces readability constraints in relevance measures for document retrieval and summarisation. The readability constraints are specifically estimated for dyslexic readers. The optimal integration rate is estimated around 30% from the observation of performances on CLEF and DUC evaluation campaigns.*

MOTS-CLÉS : *Recherche documentaire, résumé automatique, lisibilité, dyslexie*

KEYWORDS: *document retrieval, summarisation, readability, dyslexia*

## 1. Introduction

Si la prise en compte de l'utilisateur dans les systèmes de recherche d'information est une amélioration intéressante, c'est une nécessité dans le cas où l'utilisateur est handicapé. En particulier, les difficultés de lecture induites par la dyslexie (Snowling, 2000) créent un fossé informationnel pour les personnes souffrant de ce trouble.

Nous proposons dans cet article de faire évoluer les systèmes de recherche d'information en y intégrant une contrainte de lisibilité, celle-ci étant spécifiquement évaluée pour des lecteurs dyslexiques. La décomposition du besoin de l'utilisateur en un besoin thématique et un besoin orthogonal (tel que le niveau d'expertise, la langue, le type de document) est fréquemment envisagée dans la littérature. Dans ces cas le besoin orthogonal est généralement une contrainte qui ne s'exprime pas de manière continue. Cette contrainte peut être satisfaite par un filtrage des documents retournés. Pour intégrer la lisibilité, nous envisageons soit une solution de réordonnancement des documents de manière à retourner en priorité les plus lisibles, soit une solution de réduction de la quantité de texte à lire pour obtenir l'information.

D'un point de vue expérimental, il est difficile d'obtenir des données en grande quantité sur les facultés de lecture de dyslexiques, étant donné le temps nécessaire et la difficulté de la tâche. Aucune donnée concernant les retours de tels utilisateurs sur la lisibilité de documents n'étant à ce jour disponible, nous avons choisi d'estimer empiriquement le taux optimal de prise en compte de la lisibilité (évaluée spécifiquement pour des lecteurs dyslexiques) en regard de la baisse de pertinence en recherche documentaire (selon les références des données issues de campagnes d'évaluation), à des fins de comparaison. Une telle étude est réalisée pour des normo-lecteurs ainsi que pour des dyslexiques.

La seconde solution consiste à réduire la quantité d'informations à faire lire à l'utilisateur. Cela est réalisable soit en sélectionnant les parties de document les plus pertinentes, soit en réalisant un résumé de tous les documents en fonction de la requête posée par l'utilisateur. Les contraintes de lisibilité peuvent également être intégrées à ces tâches de sélection de phrases ou de passages.

La première section introduit les mesures de lisibilité existantes, ainsi qu'une nouvelle mesure adaptée aux utilisateurs dyslexiques. La seconde section s'intéresse à l'intégration des critères de lisibilité dans un processus de recherche documentaire dans le cadre de la campagne d'évaluation CLEF. La troisième section traite de la réduction de la granularité des résultats d'une recherche documentaire, à l'aide d'une segmentation thématique des documents ou d'un résumé automatique intégrant les critères de lisibilité.

## 2. Critères de lisibilité

### 2.1. *Mesures usuelles de la lisibilité*

L'évaluation de la lisibilité d'un texte au sens de son niveau de compréhension (et non purement visuel) se fait dans la plupart des logiciels d'édition de texte grand public à l'aide de la mesure établie par (Flesch, 1948) pour l'anglais et adaptée par (Kandel *et al.,* 1958) pour le français :

$$\text{pour l'anglais : } L_{Flesh} = 206,8 - (1,015 \times ASL) - (84,6 \times ASW) \quad [1]$$

$$\text{pour le français : } L_{Kandel} = 207 - (1,015 \times ASL) - (73,6 \times ASW) \quad [2]$$

où ASL est la longueur moyenne des phrases exprimée en nombre de mots et ASW est le nombre moyen de syllabes par mot contenu dans le texte. Cette mesure établit une échelle de lisibilité de 0 à 100, sur laquelle un score de 30 situe un document très difficile à lire, et un score de 70 un document correctement lisible par des adultes. Cette mesure est toujours utilisée comme option des éditeurs de texte grand public [1].

Les approches plus récentes pour estimer la lisibilité d'un document utilisent des modèles de langage statistiques ainsi que divers algorithmes pour la classification : *Expectation Maximization* (Si *et al.,* 2001), les arbres de décision (Kane *et al.,* 2006), l'analyse sémantique latente (LSA) (Wolfe *et al.,* 1998) ou des modèles de catégorisation (Collins-Thompson *et al.,* 2005). Les données sur lesquelles s'appuient ces approches proviennent dans certains cas d'annotation manuelle par des enseignants sur des pages web (Petersen *et al.,* 2006) ou sur des livres entiers (Lennon *et al.,* 2004). Les principaux paramètres utilisés par ces méthodes de catégorisation automatique sont la taille des phrases et des mots, et les caractéristiques syntaxiques et lexicales des mots.

### 2.2. *Mesure de lisibilité adaptée à la dyslexie*

Jusque dans les années 70 la dyslexie était considérée comme un trouble visuel associé à la confusion de lettres ou de syllabes. Les recherches en psycholinguistique (Snowling, 2000) ont montré qu'il s'agit en réalité d'un dysfonctionnement des représentations phonologiques qui est à l'origine de la dyslexie. Une des conséquences de cela est que les représentations mentales des liens entre les phonèmes (les sons parlés) et les graphèmes (les lettres ou groupes de lettres correspondants) sont dégradées. Du point de vue de la lisibilité, les correspondances graphèmes-phonèmes les plus complexes (comme dans les mots *manteau* ou *amphore*) vont présenter une difficulté supplémentaire pour le lecteur. Une haute fréquence de ces difficultés mobilise les

1. Dans MS Word, il faut activer l'option "afficher la lisibilité" dans l'onglet grammaire et orthographe des préférences pour voir les statistiques de lisibilité s'afficher à la fin de la vérification du document.

ressources attentionnelles du lecteur dyslexique qui perd des capacités de mémorisation à court terme, ce qui rend la compréhension de la phrase et du texte plus difficile. C'est ainsi que la complexité de la phrase devient également un facteur important pour l'évaluation de sa lisibilité. Il s'agit dans ce cas de complexité mnésique, puisque ce sont les aspects qui influent sur la mémoire à court terme qui sont impliqués.

La lisibilité est donc une caractéristique essentielle des réponses qu'un système d'information doit fournir à l'utilisateur s'il est dyslexique. Cette mesure de la lisibilité peut cependant s'inspirer des travaux déjà effectués dans ce domaine pour les normo-lecteurs, qui catégorisent généralement les textes en niveau de lecture en référence aux niveaux d'expertise.

#### 2.2.1. *Techniques d'apprentissage pour prédire la difficulté de lecture*

Pour l'établissement d'une mesure de lisibilité, nous avons élaboré (en partenariat avec des chercheurs en psycholinguistique de notre laboratoire) une base de données des temps de lecture de 20 phrases de 12 mots lues par 9 enfants dyslexiques dans le cadre d'expérimentations sur le diagnostic de la dyslexie par l'empan perceptif (où il s'agit d'établir un lien entre la taille et le positionnement de la fenêtre de lecture et la dyslexie). 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 à ce niveau, ainsi qu'au niveau de chaque phrase. 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 parmi deux dessins).

En faisant l'hypothèse que le temps de lecture d'un mot ou d'une phrase est relié à sa difficulté, alors mesurer la lisibilité d'une phrase peut se ramener à prédire son temps de lecture. La première approche choisie pour cette évaluation sont les SVM (*Support Vector Machines*), pour leur capacité à travailler sur des faibles volumes de données. Les SVM projettent les données initiales dans un espace de plus grande dimension jusqu'à trouver un hyperplan séparateur. La seconde approche est la régression linéaire, choisie pour sa capacité à fournir une mesure transparente, une combinaison linéaire des paramètres les plus discriminants.

Les paramètres utilisés sont ceux utilisés pour l'établissement de la lisibilité pour des normo-lecteurs dans le cadre d'autres expériences basées sur l'apprentissage, ainsi que ceux qui sont spécifiques à la lecture de documents par des dyslexiques. La complexité des correspondances graphèmes-phonèmes se mesure à l'aide de la cohésion grapho-phonologique : c'est le ratio entre le nombre de phonèmes et le nombre de lettres. La complexité mnésique de la phrase peut être évaluée selon les axes syntaxiques et lexicaux. La complexité syntaxique de la phrase peut être évaluée en fonction des éléments syntaxiques qui la composent (nombre de noms, pronoms, adjectifs, noms propres, verbes, adverbes, conjonctions, ...). La complexité lexicale de la phrase est le critère utilisé pour les normo-lecteurs. Elle est reflétée par la fréquence d'apparition lexicale sur le corpus des mots qui la composent, ainsi que la longueur moyenne de ces mots. Dans le cas des enfants francophones, de telles fréquences sont disponibles dans la base de données Manulex (Lété *et al.,* 2004). Le graphe de la figure 1

illustre l'ensemble des données utilisées pour refléter les différents paramètres d'une phrase, avec des valeurs attribués à chacun des paramètres.

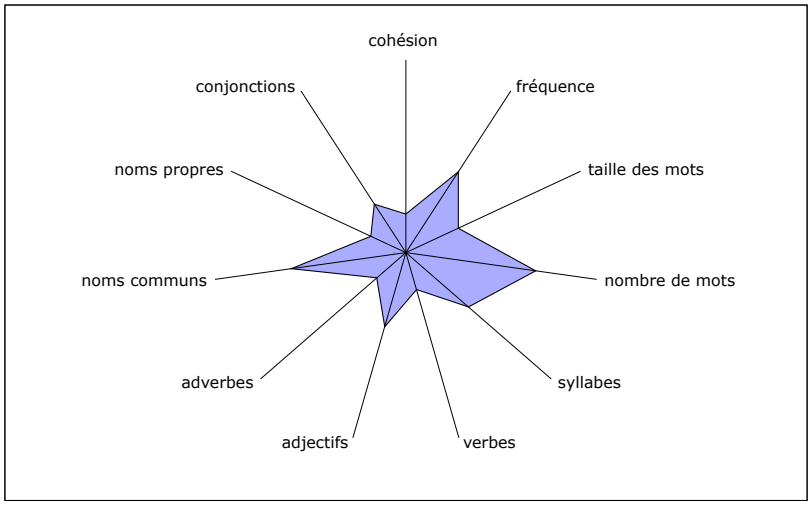


**Figure 1.** *Dimensions paramétriques d'une phrase pour évaluer sa lisibilité.*

2.2.2. *Résultats de la prédiction de temps de lecture*

Pour trouver la meilleure manière de déterminer la lisibilité d'une phrase les expériences ont été réalisées à l'aide de l'environnement WEKA [2] (Witten *et al.,* 1999). Les temps de lecture des mots ont été normalisés pour chaque utilisateur sur une échelle allant de 0 à 100 (0 étant le temps de lecture normalisé du mot le plus vite lu et 100 celui du mot lu le plus longuement). A partir de là, les temps de lecture normalisés des phrases sont les moyennes des temps de lecture normalisés des mots les constituant. Aucune normalisation n'a été effectuée par rapport à la taille des phrases car elles comportent toutes 12 mots, ni par rapport à la taille des mots.

Des modèles sur la base de données commune à tous les utilisateurs ont été réalisés et évalués à l'aide d'une validation croisée. Le tableau 1 contient l'écart moyen entre les temps prédits par les classifieurs testés (SVM et régression linéaire) et les temps réels. Une comparaison est effectuée avec un classifieur naïf (l'algorithme ZeroR affecte la valeur moyenne des données d'entraînement à toutes les données de test), et un classifieur aléatoire (qui affecte des valeurs aléatoires entre 0 et 100). Si l'on considère qu'une phrase est lue en approximativement 20 secondes, un écart de 2 point est de l'ordre du dixième de seconde. Les résultats du classifieur naïf montrent que les données utilisées sont très homogènes et centrées autour de la moyenne. Les résultats similaires pour la prédiction des temps de lecture normalisés des phrases avec les deux classifieurs testés suggèrent l'utilisation prioritaire de la régression linéaire

2. http://www.cs.waikato.ac.nz/~ml/

étant donné qu'elle fournit une mesure transparente pour des résultats équivalents aux SVM. La mesure de lisibilité ainsi obtenue est définie par :

$$L_{lin} = 1,12 * ADV - 0,69 * CON + 6,48 * cohesion + 15,58 \quad [3]$$

où $ADV$ et $CON$ sont le nombre d'adverbes et de conjonctions dans la phrase, et $cohesion$ est le nombre de phonèmes divisé par le nombre de lettres de la phrase.

| | SVM | Reg, linéaire | Naif | Aléatoire |
|---|---|---|---|---|
| mots | **9,38** | 9,74 | 10,1 | 37,97 |
| phrases | 5,01 | **5,00** | 5,07 | 35,69 |

**Tableau 1.** *Taux d'erreurs (obtenus par validation croisée 10 plis) des classifieurs utilisés (SVM et régression linéaire), d'un classifieur basé sur la moyenne des données disponibles (Naif) et d'un classifieur aléatoire, pour une prédiction des temps de lecture de mots ou de phrases.*

Cependant les bons résultats du classifieur naïf montrent que les données utilisées sont très homogènes et centrées autour de la moyenne. D'autre part les critères utilisés dans les mesures de lisibilité pour les normo-lecteurs (nombre de syllabes, taille de la phrase) restent valables pour les dyslexiques. Notre corpus d'apprentissage ne prenant pas en compte la taille des phrases (elles comportent toutes 12 mots), la mesure finalement proposée dans l'équation 4 est la moyenne arithmétique de celle issue de l'apprentissage et celle proposée par Kandel pour le français. Elle retourne une valeur entre 0 (pour un document idéalement lisible) et 100 (pour un document idéalement illisible). Des expérimentations en cours avec des phrases de taille et de composition variables permettront de valider ou d'affiner ce choix.

$$L = 0,5 \times L_{Kandel} + 0,5 \times L_{lin} \quad [4]$$

## 3. Prise en compte de la lisibilité en recherche documentaire

### 3.1. *La pertinence en recherche documentaire*

L'objectif d'un système de recherche d'information est de fournir les documents pertinents pour l'utilisateur par rapport au besoin exprimé (requête). La notion de pertinence a été largement débattue pour préciser ce qu'elle doit prendre en compte. (Mizzaro, 1997) propose un cadre de définition de la pertinence qui permet d'englober toutes les dimensions jusqu'alors évoquées.

La pertinence peut ainsi être définie selon quatre dimensions principales :

– le *besoin d'information*, décomposé en besoin réel, besoin perçu par l'utilisateur, besoin exprimé, et besoin formalisé par un langage de requête,

– les *composants* : l'information elle-même, la tâche et le contexte,

– le temps mis pour retrouver l'information,

– la granularité de l'information recherchée : document complet, sujet du document, ou information précise à l'intérieur de ce document.

Les modèles de base mettent en relation les mots de la requête avec ceux des documents, qu'ils soient explicités ou non. Le besoin sous-jacent de l'utilisateur peut s'exprimer soit à travers la sélection d'une tâche précise de recherche d'information (recherche documentaire, questions-réponses, ...), soit par son opinion sur le résultat de précédentes recherches (retour de pertinence), soit par un profil utilisateur déclaré ou déduit. A l'heure actuelle, les modèles (vectoriel, probabiliste, ...) ne prennent en compte ce besoin qu'*a posteriori*.

L'intégration du critère de lisibilité dans un système de recherche documentaire nécessite de reformuler le modèle définissant ce qu'est un document pertinent. Au sein des modèles classiques de recherche documentaire, la pertinence d'un document est évaluée en fonction de sa corrélation thématique estimée avec la requête posée par l'utilisateur. Les mots de la requête sont représentés dans un espace sémantico-lexical plus ou moins vaste (augmenté dans les cas d'expansion de requête, ou réduit dans des représentations réalisées à l'aide de l'analyse sémantique latente) et les documents les plus similaires au sens de cet espace sont retournés par ordre décroissant de score de similarité. Quelques systèmes prennent en compte des caractéristiques thématiques de l'utilisateur, en ajoutant à la requête un historique des requêtes déjà effectuées et des retours de pertinence des documents consultés.

Pour prendre en compte les capacités de lecture de l'utilisateur, il faut considérer la lisibilité comme une donnée continue que l'on cherche à maximiser tout en maintenant une forte similarité. Dans ce cadre, la similarité peut être estimée par le score de similarité d'un système de recherche documentaire, et on peut s'inspirer des travaux réalisés par (Vogt *et al.,* 1999) sur les métamoteurs de recherche pour intégrer linéairement la lisibilité.

### 3.2. *Intégration de la lisibilité dans la pertinence*

Nous proposons d'utiliser un score de pertinence $Pert(d, q)$ entre un document $d$ et une requête $q$ calculé à l'aide de la combinaison linéaire de leur similarité $Sim(d, q)$ et de la lisibilité du document $L(d)$, normalisées entre 0 et 1 :

$$Pert(d, q) = (1 - \lambda).Sim(d, q) + \lambda.L(d) \quad [5]$$
$$avec\ 0 < \lambda < 1$$

La lisibilité $L(d)$ du document $d$ est calculée à partir de la moyenne arithmétique des lisibilités de toutes les phrases le contenant.

Si la lisibilité peut se ramener dans tous les cas à une valeur continue normalisée entre 0 et 1, la similarité sur laquelle se basent les classements des moteurs de

recherche n'est pas toujours disponible. Ainsi, on peut calculer la similarité soit en normalisant le score s'il est disponible, soit en normalisant le rang $Rk(d)$ du document $d$ parmi les $N$ premiers documents retournés. Les deux manières de calculer la pertinence sont :

$$Pert(d,q) = (1-\lambda).Sim(d,q) + \lambda.\frac{L(d)}{100} \quad [6]$$

$$Pert(d,q) = (1-\lambda).(1-\frac{Rk(d)}{N}) + \lambda.\frac{L(d)}{100} \quad [7]$$

Etant donné qu'aucune donnée n'est disponible à ce jour concernant à la fois la pertinence de documents en fonction de leur similarité et de leur lisibilité, les expériences suivantes tentent de déterminer une valeur optimale du paramètre $\lambda$ qui permet d'augmenter la lisibilité des résultats fournis tout en maintenant une forte pertinence du point de vue thématique.

### 3.3. *Estimation sur les données d'une campagne d'évaluation*

La campagne d'évaluation CLEF (Cross Language Evaluation Forum)[3] fournit une référence en recherche documentaire francophone pour la tâche *ad hoc* monolingue. Cette tâche consiste à retrouver les documents pertinents pour 60 requêtes dans une collection d'environ 130 000 documents. La référence est construite par des validations manuelles des résultats de plusieurs moteurs de recherche. Il y a en moyenne 16 documents pertinents par requête posée.

L'évaluation se fait généralement à l'aide des mesures de rappel et précision sur les mille premiers documents retournés. Dans l'optique où l'utilisateur est en difficulté de lecture, l'évaluation est pertinente si elle concerne les 20, voire les 10 premiers documents retournés. Il est en effet connu que la plupart des utilisateurs du moteur de recherche Google ne dépassent que rarement les 2 premières pages de résultats.

L'estimation de la valeur optimale du paramètre réglant l'importance de la similarité par rapport à la lisibilité dans la pertinence a été réalisée en testant différentes valeurs de ce paramètre pour réorganiser les résultats d'un système uniquement basé sur la similarité. Le moteur de recherche Lucene [4] se base sur un mélange du modèle vectoriel (Salton *et al.,* 1975) et du modèle booléen pour estimer la similarité des documents avec la requête. L'expérience est réalisée à partir des scores de similarité fournis par Lucene (avec ses paramètres par défaut) pour les données françaises de la campagne d'évaluation CLEF 2003, en appliquant un score de lisibilité calculée à l'aide de la formule 3 précédemment établie pour des lecteurs dyslexiques.

3. http://clef-campaign.org
4. http://lucene.apache.org

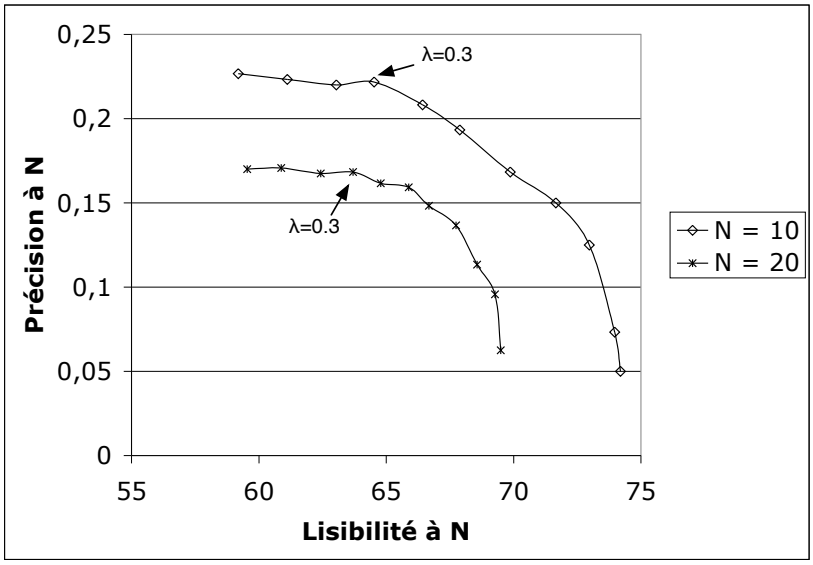


**Figure 2.** *Précision au rang N (10 ou 20) corrélée avec la lisibilité moyenne des N premiers documents, pour des résultats obtenus avec différentes valeurs de* $lambda$ *(Equation 6, en utilisant Lucene pour la similarité).*

Les résultats de l'application des deux formules d'intégration du paramètre de lisibilité (Equations 6 et 7) montrent sur les figures 2 et 3 que le calcul de la pertinence en fonction de la similarité permet d'augmenter la lisibilité sans dégrader la précision, jusqu'à un taux d'intégration de 30% de la lisibilité. En revanche, le calcul de la pertinence basé sur le rang initial des documents retournés par Lucene fait très rapidement chuter la précision des 10 premiers documents dès lors qu'on prend en compte la lisibilité. L'augmentation de la lisibilité pour $\lambda = 0,3$ dans le calcul à partir des scores de similarité est assez faible mais significative. Etant donné que la pertinence n'est pas dégradée, on peut conclure que si elle est contrôlée, l'intégration de la lisibilité apporte une amélioration notable des résultats.

La même expérience a été réalisée sur les données de la piste *ad hoc* de la campagne TREC 8. Dans ce cas, on dénombre pour chacune des 50 requêtes en moyenne 95 documents pertinents parmi la collection de 530 000 documents du corpus. Les résultats obtenus sont similaires à ceux obtenus sur CLEF 2003, et la valeur optimale du facteur $\lambda$ est obtenue pour 0,3 pour le calcul basé sur les similarités, et de 0,2 pour le calcul basé sur les rangs.

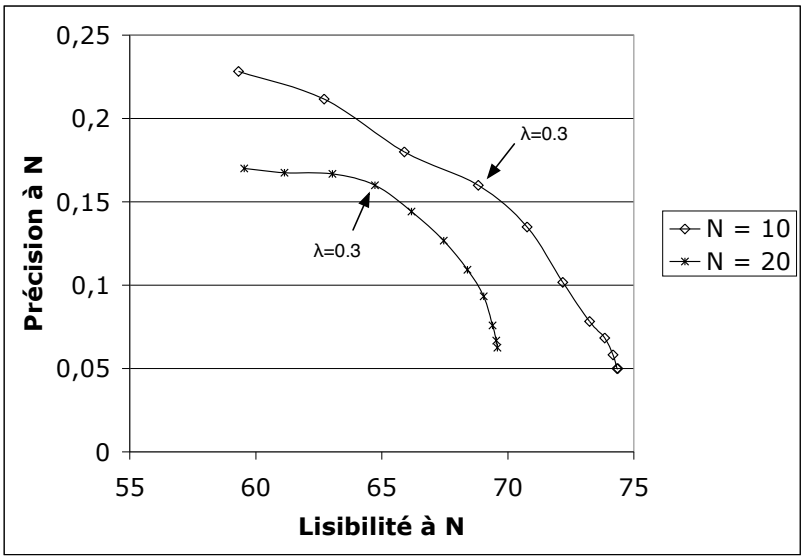


**Figure 3.** *Précision au rang N (10 ou 20) corrélée avec la lisibilité moyenne des N premiers documents, pour des résultats obtenus avec différentes valeurs de* $lambda$ *(Equation 7, en utilisant Lucene pour les rangs initiaux).*

## 4. Amélioration de la lisibilité par réduction de la granularité des résultats

Une autre manière d'aider le lecteur en difficulté est de retourner l'information sous forme plus condensée, en proposant uniquement une sélection de paragraphes ou en réalisant un résumé le plus lisible possible de ce que proposent les documents en regard de la requête.

### 4.1. *Utilisation de la segmentation thématique pour réduire la taille des documents*

En ciblant l'information recherchée à l'intérieur des documents, on peut réduire l'effort de lecture de l'utilisateur. De plus (J.Callan, 1994) suggère que la réduction des unités de traitement textuelles améliore la qualité des informations retrouvées.

De nombreux algorithmes ont été proposés pour segmenter un texte en segments cohérents de plus petite taille lorsque les démarcations en paragraphes de l'auteur ne sont pas disponibles, ou ne correspondent pas à la taille recherchée (Sitbon *et al.,* 2004). Cependant étant donnée l'efficacité relative de ces algorithmes, et vu que les limites de paragraphes sont disponibles dans les données de la campagne CLEF, nous nous sommes appuyés sur les paragraphes existants pour les expériences.

La première approche proposée est d'indexer les paragraphes comme s'ils étaient des documents à part entière. Les résultats de cette approche dans le tableau 2 montrent que les résultats sont fortement dégradés lorsqu'on réduit de cette façon les unités documentaires indexées. En effet la précision moyenne (MAP) ainsi que la précision des 10 ou 20 premiers documents subissent des dégradations significatives.

| Unité de texte | MAP | P10 | P20 |
|---|---|---|---|
| Document | 0,31 | 0,23 | 0,17 |
| Segment | 0,19 | 0,17 | 0,15 |

**Tableau 2.** *Mean average precision (MAP) et précision au 10ème rang (P10) et au 20ème rang (P20) sur les requêtes courtes, avec une indexation par Lucene des documents complets ou des segments uniquement.*

Ces résultats suggèrent que l'index des documents est plus performant et donc qu'il serait plus raisonnable de le conserver. Cela n'empêche pas néanmoins de remplacer les documents par leur segment le plus pertinent (du moins celui ayant le score le plus élevé) dans les cas où l'index des segments en valide la pertinence par rapport à la requête. En effet on considère que si un segment d'un document est aussi pertinent que le document en entier, c'est que la majorité de l'information pertinente du document est contenue dans ce segment. Cette approche peut être combinée avec une sélection des documents les plus lisibles, en considérant qu'un segment pris au lieu d'un document a une lisibilité maximale.

Cette seconde approche a été également évaluée sur les données de la campagne CLEF 2003, sans prise en compte de la lisibilité des documents complets ($\lambda = 0$) ou avec une prise en compte équivalente à l'intégration optimale calculée pour la recherche documentaire ($\lambda = 0,3$). Le tableau 3 contient la précision des 10 ou 20 premiers éléments retournés, leur lisibilité moyenne, ainsi que le nombre d'éléments étant des segments remplaçant des documents. Si un élément est un segment, le document dont il est issu est utilisé pour l'évaluation. De plus, on lui confère une lisibilité maximale, qui reflète ici le gain considérable apporté par la réduction de la quantité de texte à lire et non la lisibilité moyenne des phrases.

| $\lambda$ | N | Précision à N | Lisibilité | Segments |
|---|---|---|---|---|
| 0 | 10 | 0,23 | 85,68 | 6,55 |
| 0 | 20 | 0,17 | 86,92 | 13,6 |
| 0,3 | 10 | 0,20 | 93,01 | 8,2 |
| 0,3 | 20 | 0,16 | 93,80 | 16,65 |

**Tableau 3.** *Précision, lisibilité moyenne et nombre de segments retournés parmi les N premiers éléments retrouvés, avec différentes valeurs du paramètre $\lambda$ pour la prise en compte de la lisibilité dans le classement des résultats.*

Les résultats montrent qu'en remplaçant les documents contenant un segment de similarité supérieure par ce segment, et sans estimer la pertinence en fonction de la lisibilité, on aboutit à une faible perte de précision en regard de celle subie lors de l'indexation pure des segments introduite dans la table 2. Dans tous les cas, 50 % des documents contiennent suffisamment d'information pertinente dans un seul segment.

Dans une configuration où l'on prend en compte la lisibilité des documents complets (à hauteur de 30%), on aboutit à 80% de documents qui peuvent être référés par un segment aussi pertinent. La précision obtenue est à mi-chemin entre celle qui est obtenue en indexant uniquement les segments et celle qui est obtenue en ne prenant pas en compte la lisibilité.

### 4.2. *Sélections des phrases les plus lisibles pour la génération du résumé*

La campagne d'évaluation DUC[5] (*Document Understanding Conference*) est dédiée au résumé automatique. Elle comporte une tâche de résumé multi-documents orienté requête, ce qui correspond à une synthèse de l'information disponible et distillée dans le corpus. Si cette tâche permet de faire ressortir des informations noyées dans des documents plus vastes, elle présente aussi l'avantage de réduire la quantité de texte à lire pour obtenir des informations. Les méthodes les plus efficaces dans ce domaine effectuent une extraction des phrases les plus pertinentes.

#### 4.2.1. *Génération de résumés orientés requête par MMR-LSA*

Le système de résumé sur lequel nous nous appuyons (Favre *et al.,* 2006) sélectionne une par une les phrases selon un critère qui maximise à la fois leur similarité à la requête et leur dissimilarité au résumé constitué des phrases précédemment sélectionnées, ceci afin d'éviter la redondance.

La sélection des phrases du résumé se fait en autant d'étapes que nécessaire pour parvenir au nombre de mots souhaités dans le résumé. La méthode MMR (*Maximum Marginal Relevance*) proposée par (Carbonell *et al.,* 1998). A chaque étape, un algorithme glouton sélectionne la phrase qui maximise sa similarité avec la requête tout en minimisant sa similarité avec la moyenne des phrases déjà sélectionnées.

La notion de similarité suggère de placer les phrases dans un espace vectoriel à l'intérieur duquel il sera possible de calculer des distances. L'approche proposée ajoute à la représentation vectorielle une projection dans un espace réduit. Les phrases sont alors représentées dans une projection de l'espace des mots vers l'espace sémantique correspondant réduit à l'aide de l'analyse sémantique latente (Deerwester *et al.,* 1990). Celle-ci permet de créer des classes de mots en fonction de leurs cooccurrences, et s'appuie sur une décomposition en valeurs singulières de la matrice de cooccurrences du corpus. Ainsi la similarité cosine entre une phrase $p$ et une requête

5. http://duc.nist.gov/

$q$, ou entre une phrase $p$ et un résumé $q$, est exprimée en fonction des poids de leurs termes communs $p_i$ et $q_i$ par :

$$cosine(p,q) \quad = \frac{p.q}{|p|.|q|} = \frac{\sum_i p_i q_i}{\sqrt{\sum_i p_i^2}\sqrt{\sum_i q_i^2}} \qquad [8]$$

Les phrases sont alors ordonnées au sein du résumé (les premières sont les plus pertinentes), mais l'on ne dispose pas de score de pertinence associé. En effet, la mesure utilisée à chaque étape par l'algorithme pour sélectionner la phrase la plus pertinente à cette étape fournit un score de la phrase uniquement par rapport au résumé existant à cet instant.

#### 4.2.2. *Evaluation*

La campagne DUC en 2006 propose l'évaluation de 50 résumés d'une taille maximale de 250 mots. Chaque résumé est réalisé à l'aide d'un *titre* et d'une *description*. Par exemple le résumé *D0629B* a pour titre *"Les virus informatiques"* et pour descriptif *"Identifiez les virus informatiques ayant eu une propagation mondiale. Détaillez de quelle façon ils se répandent, les systèmes d'exploitation affectés, leurs pays d'origine, et leurs créateurs quand cela est possible."* Par ailleurs, les résumés peuvent s'appuyer sur une liste des 25 documents pertinents pour chaque requête, liste qui est fournie aux participants

Les résumés sont évalués à l'aide d'une mesure de comparaison entre les $n$-grammes des résumés de référence (4 par requête, rédigés manuellement) et ceux produits automatiquement. La mesure ROUGE-2 proposée par (Lin, 2004) se base sur la comparaison des bigrammes et est réputée comme étant la plus fiable parmi les mesures d'évaluation automatique. Lors de la campagne DUC, les résumés sont également évalués manuellement en termes de qualité linguistique. Elle prend en compte des critères de cohérence et de style, mais n'est pas corrélée à la lisibilité en termes de confort de lecture. Le système précédemment présenté s'est très bien positionné par rapport aux systèmes concurrents.

Etant donné que les données disponibles sont rédigées en anglais, nous avons choisi d'appliquer une mesure de lisibilité établie pour l'anglais, même si elle n'est pas spécifique aux dyslexiques. La mesure de Flesch, détaillée dans l'équation 1, est la plus largement utilisée encore de nos jours. Etant donné que le système de résumé de propose pas de score de pertinence pour chacune des phrases sélectionnées, nous avons utilisé une intégration de la lisibilité par rapport au calcul du rang, telle que proposée par l'équation 7.

Le graphe de la figure 4 montre la corrélation entre les taux de lisibilité (calculés selon la mesure de Flesch) et les valeurs de la mesure ROUGE-2 pour les résumés produits en intégrant la lisibilité selon l'équation 7 avec différentes valeurs de $\lambda$. Les lignes de référence sont les valeurs de ROUGE-2 pour le meilleur système de DUC et les valeurs pour un système "naïf" (résumé obtenu en reproduisant le document le plus récent de la collection), ainsi que la valeur de Flesch pour les résumés de référence,

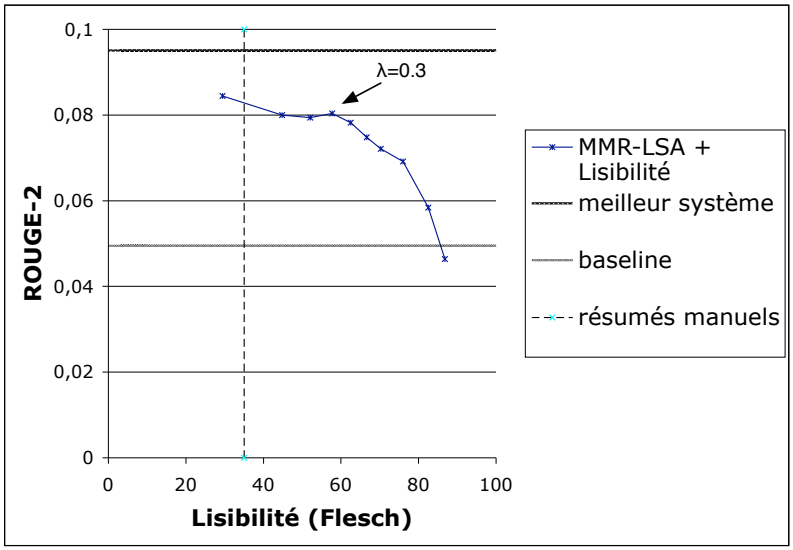


**Figure 4.** *Lisibilité et pertinence selon la mesure ROUGE-2 de résumés produits avec différentes valeurs de lambda entre 0 et 1. Les lignes de référence sont les valeurs de ROUGE-2 pour le meilleur système de DUC et pour le système naïf, ainsi que la valeur de Flesch pour les résumés produits manuellement.*

produits manuellement. La courbe présente un point d'optimalité pour une valeur de $\lambda$ à 0,3, qui correspond également à la valeur optimale de prise en compte de la lisibilité pour la recherche documentaire. Le gain de lisibilité est très important, puisqu'il est de près d'un tiers de l'échelle d'évaluation de Flesch. De plus, la lisibilité des résumés obtenus dépasse celles des résumés manuels. La figure 5 montre un exemple de résumé sur le thème des *virus informatiques* produit en prenant la lisibilité en compte avec cette valeur optimale (à droite), et le résumé produit pour le même thème sans prendre en compte la lisibilité.

## 5. Conclusion

La raison pour laquelle il est possible de réorganiser les données afin d'optimiser un critère orthogonal au besoin thématique est qu'il existe dans les cas étudiés suffisamment d'informations thématiquement pertinentes pour pouvoir sélectionner uniquement les plus lisibles. D'après les expériences menées sur la recherche documentaire et le résumé automatique, il est possible de prendre en compte la lisibilité pour 30% du score de pertinence sans pour autant fortement dégrader les performances.

Les expériences ont été réalisées sur des données en français et en basant l'évaluation de la lisibilité sur la mesure élaborée pour des enfants dyslexiques. Des ex-

| $\alpha$ = 0.0 - 10 sentences - 240 words - R = 26,9 | $\alpha$ = 0.3 - 19 sentences - 229 words - R = 58,5 |
|---|---|
| The Melissa macro or W 97 M Melissa virus spreads via infected e mail and attacks computers loaded with Microsoft's widely used Word 97 or Word 2000 programs, according to CERT or Computer Emergency Response Team Carnegie Mellon's Department of Defense funded computer security team. Disguised as a list of pornographic Internet sites and allegedly named after a stripper David Smith once knew, Melissa is a macro virus, a document with a malignant computer program built in. When the software was downloaded, computer users infected other files on their hard drive. Zip virus, which enters machines in almost the same way as the recent Melissa virus by disguising itself as a friendly piece of e mail. Melissa, as the new viral vixen was named by its creator as unknown combines elements of both a computer virus and a worm program. Melissa typically enters a computer hidden in a Microsoft Word 97 or Word 2000 file attached to an electronic mail message. Computer experts used unique identification numbers embedded in Microsoft Word documents to trace Melissa back to a well known virus writer who calls himself VicodinES. No matter how it arrives, Melissa can infect any computer that uses Microsoft Word, including Macintoshes. It generally gets into your computer by way of an attachment to an e mail. Unlike the recent Melissa scare, which automatically propagated via e mail, this virus doesn't spread as quickly because it requires a person to launch an infected program file to contaminate a computer. | Zip began to spread. The new virus, named W 32 /Kriz. Zip on his computer. As the virus spreads, the file certainly will change. Chen did not come up with an anti virus program. If an infected program was sent in an e mail, the virus was passed on to the recipient. Since both Word and Outlook are so widely used, the Melissa virus spread with shocking speed. Gets in via e mail, floppies or downloaded software. Many were caught off guard by the amount of damage and said it was much worse than the Melissa virus. Here are some recent viruses; all of them can be blocked by anti virus software. It generally gets into your computer by way of an attachment to an e mail. New viruses are being created all the time. It is clear that the virus caused much damage. Bc CIR computer virus list NYT. Computer experts said Chen might not be charged because he did not intend to spread the virus. On the screens of infected computers when a user tries to open an MS Word file. The disk from the helpline would detect and remove more than 9400 other computer viruses. Zip is the third major bug to sweep across the Internet since March, when the Melissa virus overwhelmed systems with floods of e mail. A third virus, called Mad Cow Joke has appeared and works like Melissa, sending itself to 20 people in the victim's e mail address book. |

**Figure 5.** *Exemples de deux résumés sur le thème des virus informatiques, produits respectivement sans prise en compte de la lisibilité, et avec une lisibilité prise en compte avec un facteur de 0,3.*

périences similaires ont également été réalisées sur des données en anglais, avec la mesure de Flesch pour la lisibilité, en utilisant un outil de segmentation thématique. Les résultats obtenus sont tout à fait similaires, ce qui tend à valider leur généralisation. De même des expériences similaires ont été réalisées sur le résumé automatique en évaluant la lisibilité à l'aire de la mesure établie sur le français pour les dyslexiques, ce qui amène des résultats identiques. On peut imaginer une généralisation des résultats à n'importe quelle mesure sur les textes qui ne dépende pas de leur sens.

## 6. Bibliographie

Carbonell J., Goldstein J., « The use of mmr, diversity-based reranking for reordering documents and producing summaries », *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, p. 335-336, August, 1998.

Collins-Thompson K., Callan J., « Predicting reading difficulty with statistical language models », *Journal of the American Society for Information Science and Technology*, vol. 56, n° 13, p. 1448-1462, November, 2005.

Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R., « Indexing by Latent Semantic Analysis », *Journal of the American Society for Information Science*, vol. 41, n° 6, p. 391-407, 1990.

Favre B., Béchet F., Bellot P., Boudin F., El-Bèze M., Gillard L., Lapalme G., Torres-Moreno J.-M., « The LIA-Thales summarization system at DUC-2006 », *Proceedings of Document Understanding Conference (DUC-2006)*, New York, USA, June, 2006.

Flesch R., « A new readability yardstick », *Journal of applied psychology*, vol. 32, p. 221-233, 1948.

J.Callan, « Passage-Level Evidence in Document Retrieval », *Proccedings of the ACM/SIGIR Conference of Research and Development in Information Retrieval*, p. 302-310, 1994.

Kandel L., Moles A., « Application de l'indice de flesch à la langue française », *The journal of educationnal research*, vol. 21, p. 283-287, 1958.

Kane L., Carthy J., Dunnion J., « Readability Applied to Information Retrieval », *Proceedings of the European Conference on Information Retrieval (ECIR)*, London, England, p. 523-526, 2006.

Lennon C., Burdick H., « The Lexile Framework as an Approach for Reading Measurement and Success », electronic publication on www.lexile.com, April, 2004.

Lété B., Sprenger-Charolles L., Colé P., « MANULEX : A grade-level lexical database from French elementary-school readers », *Behavior Research Methods, Instruments, and Computers*, vol. 36, p. 156-166, 2004.

Lin C.-Y., « ROUGE : a Package for Automatic Evaluation of Summaries », *Proceedings of WAS*, 2004.

Mizzaro S., « Relevance : the whole history », *Journal of the American Society for Information Science*, vol. 48, n° 9, p. 810-832, 1997.

Petersen S. E., Ostendorf M., « Assessing the Reading Level of Web Pages », *Proceedings of Interspeech 2006 - ICSLP*, Pittsburgh, Pennsylvania, p. 833-836, September, 2006.

Salton G., Wong A., Yang C. S., « A Vector Space Model for Automatic Indexing », *Communications of the ACM,*, vol. 18, n° 11, p. 613-620, 1975.

Si L., Callan J., « A statistical model for scientific readability », *Proceedings of CIKM'01*, Atlanta, USA, p. 574-576, 2001.

Sitbon L., Bellot P., « Adapting and comparing linear segmentation methods for french. », *Proceedings RIAO'04*, Avignon, France, 2004.

Snowling M. J., *Dyslexia*, Blackwell, 2000.

Vogt C. C., Cottrell G. W., « Fusion Via a Linear Combination of Scores », *Information Retrieval*, vol. 1, n° 3, p. 151-173, 1999.

Witten I. H., Frank E., *Data Mining : Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco, 1999.

Wolfe M., Schreiner M., Rehder B., Laham D., Kinstch W., Landauer T., « learning from text : matching readers and texts by latent semantic analysis », *Discourse Processes*, vol. 25, p. 309-336, 1998.