Eléments pour adapter les systèmes de recherche d'information aux dyslexiques

Laurianne Sitbon, Patrice Bellot, Philippe Blache

Eléments pour adapter les systèmes de recherche d'information aux dyslexiques

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

[*] *Laboratoire d'Informatique d'Avignon - Université d'Avignon*
{laurianne.sitbon, patrice.bellot}@univ-avignon.fr
[**] *Laboratoire Parole et Langage - CNRS - Université de Provence*
blache@lpl.fr

RÉSUMÉ. *La prise en compte des troubles de la communication dans l'utilisation des systèmes de recherche d'information tels qu'on peut en trouver sur le web est généralement réalisée par des interfaces utilisant des modalités n'impliquant pas la lecture et l'écriture. Peu d'applications existent pour aider l'utilisateur en difficulté dans la modalité textuelle. Nous proposons la prise en compte de la conscience phonologique pour assister l'utilisateur en difficulté d'écriture de requêtes (dysorthographie) ou de lecture de documents (dyslexie). En premier lieu un système de réécriture et d'interprétation des requêtes entrées au clavier par l'utilisateur est proposé : en s'appuyant sur les causes de la dysorthographie et sur les exemples à notre disposition, il est apparu qu'un système combinant une approche éditoriale (type correcteur orthographique) et une approche orale (système de transcription automatique) était plus approprié. En second lieu une méthode d'apprentissage automatique utilise des critères spécifiques , tels que la cohésion grapho-phonémique, pour estimer la lisibilité d'une phrase, puis d'un texte.*

ABSTRACT. *Most applications intend to help disable users in information retrieval process by proposing non-textual modalities. This paper introduces specific parameters linked to phonological awareness in the textual modality. This will enhance the ability of systems to deal with orthographic issues and with the adaptation of results to the reader when for example the reader is dyslexic. We propose a phonology based sentence level rewriting system that combines spelling correction, speech synthesis and automatic speech recognition. This has been evaluated on a corpus of questions we get from dyslexic children. We propose a specific sentence readability measure that involves phonetic parameters such as grapho-phonemic cohesion. This has been learned on a corpus of reading time of sentences read by dyslexic children.*

MOTS-CLÉS : *Recherche d'information, dysorthographie, dyslexie, automates, lisibilité.*

KEYWORDS: *Information retrieval, dysorthography, dyslexia, finite state machines, readability.*

Introduction

La prise en compte des troubles de la communication dans le domaine de la recherche d'information peut généralement être réalisée en proposant des solutions techniques d'interface (clavier virtuel, interface vocale, utilisation d'outils de visualisation, ...). Cependant, pour certains troubles de la communication, le média écrit est maintenu mais dégradé. Les troubles concernés sont typiquement la dysorthographie et la dyslexie. Ces deux derniers sont des troubles déclarés en l'absence de tout autre trouble moteur, social ou cognitif. Cependant des cas similaires à de la dysorthographie ont été relevés chez des IMC (Boissière *et al.*, 2007). Les causes de la dyslexie et de la dysorthographie ne sont pas encore complètement déterminées, et plusieurs hypothèses se confrontent. Les travaux dans ce domaine sont toujours en cours et un état de l'art en est donné dans (collective, 2007). L'hypothèse la plus communément admise et la seule ayant à ce jour fait ses preuves est celle du déficit phonologique (dégradation de la conscience phonologique, qui sera détaillée dans la première section).

L'objet du travail que nous proposons ici est la prise en charge du déficit phonologique dans la communication écrite avec un système de recherche d'information. Cette prise en charge est nécessaire en vue de diminuer la fracture numérique inhérente à l'explosion des nouvelles technologies dans la vie de tous les jours. Elle est également une application de la loi du 11 février 2005 pour l'égalité des droits et des chances, la participation et la citoyenneté des personnes handicapées. Des applications dédiées aux dyslexiques existent. Elles font généralement abstraction des contraintes typographiques en proposant des interfaces audio[1], de type dictée vocale et lecteurs d'écran. Elles constituent une bonne compensation mais requièrent un matériel pas toujours disponible et accentuent le découragement des utilisateurs face à l'utilisation de la lecture l'écriture, ce qui ne favorise pas le travail de remédiation par ailleurs effectué. Nos propositions concernent la modalité textuelle. Elles se fondent sur l'analyse du déficit phonologique en général, et nous appuierons nos évaluations sur des cas de dyslexie-dysorthographie chez des enfants.

Après un survol des causes principales de la dyslexie et la dysorthographie dans la première section, nous établirons les problèmes rencontrés par des utilisateurs dyslexiques, et étudierons les solutions envisagées dans l'état de l'art. Nous proposerons dans la troisième section une technique pour pallier la dysorthographie pour des questions en langage naturel. Cette technique combine les atouts d'un correcteur orthographique et d'un modèle de langage dans le cadre formel des machines à états finis. La troisième section sera consacrée à l'établissement d'une mesure de lisibilité de phrases destinée aux domaines de la recherche documentaire et du résumé automatique, afin d'appliquer une sélection sur la lisibilité des documents. Cette mesure est établie collectivement et individuellement à partir d'enregistrements de temps de lecture de phrases et de mots, en utilisant des paramètres spécifiques à l'approche de la lecture dans des cas de déficit phonologique.

---

1. voir par exemple `http://www.01net.com/article/264021.html`

1. La dyslexie et la dysorthographie

1.1. *Impact sur la lecture et l'écriture*

La dyslexie est un trouble de la lecture. Un lecteur dyslexique rencontre des problèmes lors du décodage des mots, principalement si ceux-ci ont une graphie opaque. L'opacité d'un mot est reliée à la présence de graphèmes (lettres ou groupes de lettres) complexes, ou plus généralement au nombre et à la complexité des règles de lecture à employer pour le lire. Par exemple le mot *amphore* est très opaque (il contient deux graphèmes à deux lettres peu fréquents et un "e" muet) tandis que *atala* est un mot peu opaque (les règles de correspondance lettre/phonème suffisent pour le lire). De plus le lecteur dyslexique pourra être souvent perturbé par la présence d'informations autour de celle qu'il est en train de décoder à un instant donné. L'utilisation de caches sur le texte ou de repères visuels est souvent préconisée pour aider le lecteur.

La dysorthographie est un trouble de la production écrite. Elle est généralement associée à une dyslexie. Une phrase écrite par un dysorthographique aura généralement les caractéristiques suivantes :

– Segmentation en mots erronée (*Je vais à les colles* au lieu de *Je vais à l'école*)

– Erreurs de conversion graphème/phonème (*Ailaifen* au lieu de *Éléphant*)

– Confusions de phonèmes (*Monné* au lieu de *Monnaie*)

Par ailleurs, certaines dysorthographies entraînent des inversions de lettres (par exemple *itulisé* au lieu de *utilisé* lors de la production écrite. La dysorthographie est très différente de la dysgraphie, qui implique la confusion de lettres manuscrites à cause de leur graphie (*d* pour *p* ou *b*).

1.2. *La conscience phonologique*

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 ont montré qu'il s'agit en réalité d'un trouble phonologique (Snowling, 2000). Afin de comprendre ce qu'est la dyslexie, il faut remonter au fonctionnement cognitif du langage écrit, qui est un acquis généralement fondé sur le langage oral. L'acquisition du langage oral s'effectue par un apprentissage progressif fondé sur la répétition, qui permet de relier des séries de sons à des sens. Ainsi se créent en parallèle le lexique sémantique et le lexique phonologique. De plus, les règles de composition d'objets du lexique phonologique sont intégrées afin de permettre de former des phrases qui associent des éléments du lexique sémantique dans le but de transmettre des messages.

Plus tard, lors de l'apprentissage académique, la lecture et l'écriture sont enseignées. Dans des systèmes alphabétiques tel que le français, un certain nombre de règles régissent les correspondances entre les lettres et les sons (correspondances graphème-phonème), ainsi que les correspondances entre le rôle des mots dans la phrase et leur composition (règles d'accord et de conjugaison). La lecture s'effectue à

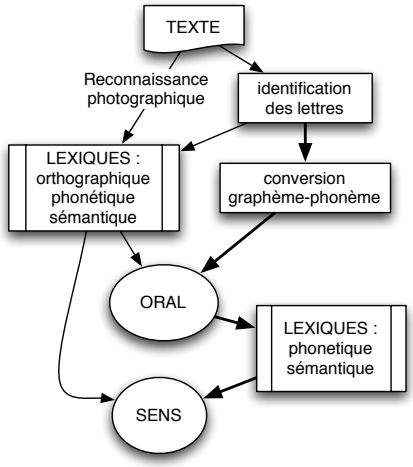

Figure 1. *Illustration du modèle de la lecture à double voie : les flèches en gras indiquent la procédure sub-lexicale et les autres flèches la procédure lexicale.*

la fois à l'aide de procédures dites lexicales (qui s'appuient directement sur des unités ayant du sens : les mots) et de procédures dites sub-lexicales (procédures de décodage phonologique). Ces procédures sont appliquées en parallèle, selon le modèle de lecture à double voie introduit par (Coltheart et Rastle, 1994) et dont la figure 1 illustre le fonctionnement. Lors de l'apprentissage, la procédure sub-lexicale est favorisée, jusqu'à ce que le cerveau ait rencontré un mot suffisamment souvent pour en mémoriser directement la composition. Ainsi les mots les plus fréquents sont atteints par la procédure lexicale. La production écrite (manuscrite ou typographiée) d'un énoncé identifié (dicté ou pensé) suppose pour sa part une première étape d'identification des mots qu'il contient (segmentation), une seconde étape de transcription de ces énoncés dont la phonologie est connue, et enfin une troisième étape de contrôle par la lecture. De la même manière que pour la lecture, l'étape de transcription fait appel à des procédures lexicales et sub-lexicales.

Chez les dyslexiques et les dysorthographiques, on constate une déficience des procédures sub-lexicales, identifiée principalement pour les systèmes alphabétiques (Ramus *et al.*, 2003) dits opaques (lorsque la cohérence entre les graphèmes et les phonèmes est très faible). La charge attentionnelle dédiée au décodage empêche le développement de la procédure lexicale, principalement pour les mots à orthographe opaque. D'un autre côté, la segmentation des énoncés ne se fait généralement pas correctement, à l'écrit notamment. Cependant, si les règles de conversion sont appliquées avec peine, elles sont connues, ce qui produit des énoncés transcrits phonème par phonème et non mot par mot.

2. Interactions avec un utilisateur dyslexique

Dans le cadre de la recherche d'information, c'est la machine qui devra pallier le déficit du module phonologique en le substituant par des composants équivalents. Il y aura autant de composants de compensation à réaliser qu'il y a d'utilisations cognitives de la conscience phonologique. Pour un dyslexique, les représentations mentales des liens entre les phonèmes et les graphèmes sont dégradées. Cette dégradation a un impact double lors du processus de recherche d'information : lors de l'écriture de requêtes et lors de la lecture de documents.

2.1. *Problèmes à résoudre*

Lors de l'écriture, la conscience phonologique permet de découper les unités de sens en mots et de découper les mots en phonèmes, qui seront associés à des graphèmes. Le déficit implique une reconnaissance incomplète des mots en tant qu'unités sémantiques. Une phrase écrite est vue comme un enchaînement de sons représentés par des graphèmes. Les règles de décodage de telles productions écrites ne sont pas les liens sémantico-lexicaux classiquement rencontrés, mais se limitent aux correspondances graphèmes/phonèmes au niveau de la phrase, à différents degrés. Un décodage acoustique permettrait alors de révéler le sens de la phrase écrite.

D'un point de vue de la lecture, les correspondances graphèmes-phonèmes les plus complexes (comme dans les mots *manteau* ou *amphore*) présentent une difficulté supplémentaire. Une haute fréquence de ces difficultés mobilise les 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. La dénomination employée dans ce cas sera la complexité mnésique, puisque ce sont les aspects qui influent sur la mémoire à court terme qui seront impliqués.

Une mesure de lisibilité adaptée à des lecteurs dyslexiques permettrait de sélectionner les réponses d'un système de recherche d'information non seulement en fonction de leur contenu informationnel, mais également en fonction de leur capacité à transmettre ce contenu à un lecteur dyslexique. En effet, la lisibilité a un impact sur la fluidité de la lecture, et de récentes études (collective, 2007) ont montré qu'en deçà d'un certain seuil de vitesse de lecture, la compréhension n'est plus possible. Notre hypothèse de départ est qu'en augmentant la lisibilité, on augmente la fluidité de la lecture, et par là même la compréhension.

La complexité des conversions se traduit en termes de ratio entre le nombre de phonèmes et le nombre de graphèmes présents dans un mot. Or des lexiques effectuant des correspondances entre la graphie et la prononciation des mots existent. L'identification de la difficulté est donc automatisable. Elle peut être utilisée en privilégiant la lecture de mots "simples", en privilégiant les textes courts, ou en mettant la difficulté en exergue de manière à éviter les confusions. Cependant chaque dyslexique aura une

sensibilité propre à chacun des paramètres affectant la lecture. Un modèle pourrait alors être réalisé individuellement pour chaque lecteur.

2.2. *Solutions envisagées*

2.2.1. *Comment réécrire pour les dysorthographiques ?*

Divers modèles pour la correction orthographique dédiée ont été proposés. (Loosemore, 1991) propose une modélisation globale des erreurs commises par des dyslexiques, arguant que la dyslexie implique des erreurs aggravées mais pas différentes par rapport à celles produites par des non dyslexiques. De la même manière, (Deorowicz et Ciura, 2005) propose des réseaux de confusions représentés par des automates, où les alternatives sont issues de modèles de confusion graphiques supposés modéliser les causes d'erreurs. (Spooner, 1998), toujours en partant de l'idée qu'une erreur commise par un dyslexique ne se différencie que par son niveau de gravité, propose des modèles spécifiques à chaque utilisateur. Le correcteur qu'il implémente à l'aide de ces modèles obtient des performances comparables à celles des correcteurs orthographiques grand public. Enfin, (Toutanova et Moore, 2002) propose une approche qui combine des modèles de lettres et de phonèmes sur les mots, en se basant sur les approches probabilistes de canal bruité introduites par (Brill et Moore, 2000). L'ensemble de ces systèmes permettent de corriger des mots hors d'un lexique mais ne tiennent pas compte des homophones. (Pedler, 2001) propose une détection de telles erreurs à l'aide de contextes syntaxiques et sémantiques, sur la base d'ensembles de confusion.

2.2.2. *Comment sélectionner des documents lisibles pour des dyslexiques ?*

La lisibilité peut être vue comme la facilité à s'approprier le contenu informationnel d'un texte. Cela peut être relié à la présentation générale, ou la composition des phrases. La présentation générale implique un découpage structurel du texte, la cohérence du texte, ainsi que plus simplement sa taille. La composition des phrases reflète la qualité des unités de traitement. Leur simplicité se rapporte à l'effort pour le lecteur de ce transfert d'information textuelle, et indirectement au temps moyen qu'il mettra pour lire le texte. La définition formelle de cette lisibilité proposée par (Chall et Dale, 1995) stipule que la lisibilité est "*la somme totale (incluant les interactions) de tous les éléments à l'intérieur d'un morceau de support imprimé qui affecte le succès qu'un groupe de lecteurs peut avoir avec. Le succès réfère à la manière dont ils le comprennent, le lisent à une vitesse optimale, et le trouvent intéressant*".

Les travaux effectués sur la "transparence de lecture" (*legibility*) dans le cadre de la WAI [2](*Web Accessibility Initiative*) imposent des normes de présentation de l'information qui permettent une adaptation des contrastes, des polices, des espacements et de l'alignement des phrases. Ces aspects ne seront pas considérés dans cet article,

2. www.w3.org/WAI/

étant donné qu'ils sont déjà pris en charge sur le plan théorique par les directives de la WAI et sur le plan pratique par des navigateurs Web adaptés.

La plupart des logiciels d'édition de texte grand public proposent dans leurs outils intégrés une estimation de la lisibilité d'un document en cours de rédaction. Elle se fait alors généralement à l'aide de la mesure établie par (Flesch, 1948) pour l'anglais et adaptée par (Kandel et Moles, 1958) pour le français :

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

$$pour\ le\ francais\ :\quad L = \quad 207 - (1,015 \times ASL) - (73,6 \times ASW) \quad [2]$$

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. Une autre approche du problème, proposée par (Dale et Chall, 1948), se base sur une liste de 3 000 mots connus comme étant non difficiles. Le nombre moyen de mots de cette liste dans chaque extrait de 100 mots du texte indique le niveau de lisibilité.

Les approches plus récentes utilisent des modèles de langage statistiques ainsi que divers algorithmes pour la classification : Expectation Maximization (Si et Callan, 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 Callan, 2005). Les données proviennent dans certains cas d'annotation manuelle par des enseignants sur des pages web (Petersen et Ostendorf, 2006) ou sur des livres entiers (Lennon et Burdick, 2004). Les principaux paramètres utilisées par ces méthodes sont la taille des phrases et des mots, et les caractéristiques syntaxiques et lexicales.

2.3. *Données d'observation et d'évaluation*

Le corpus de travail que nous avons recueilli est un corpus de questions tapées par des enfants dyslexiques (qui sont également dysorthographiques). Ce corpus a été réalisé lors de séances d'orthophonie de huit enfants (entre 9 ans et demi (classe de CE2) et 13 ans (classe de 4e)). Les enfants sont les plus prompts à interroger spontanément des systèmes de recherche d'information en formulant des questions complètes plutôt que des mots clés isolés.

Dans une perspective de pouvoir estimer l'impact de la réécriture sur les performances d'un système de questions réponses, le choix des questions a été guidé par les contraintes d'évaluation du système "Questions Réponses" en langage naturel du LIA, SQuALIA (Gillard *et al.*, 2005), ainsi que par le vocabulaire restreint des enfants. Nous avons sélectionné des questions factuelles de la campagne d'évaluation Technolangue EQUER (Ayache *et al.*, 2006) pour lesquelles SQuALIA a fourni une bonne réponse, soit environ 200 questions sur les 500 proposées. Parmi ces 200 questions, nous avons sélectionné celles dont tous les mots se trouvent dans le lexique de niveau

cours préparatoire de Manulex (Lété *et al.*, 2004), qui recense les fréquences des mots de manuels scolaires pour différentes classes d'âge, et donne ainsi un aperçu des mots écrits connus par les enfants. Les 5 questions finalement retenues portent sur le nom du maire de Bastia, l'âge de l'abbé Pierre, la capitale de Terre Neuve, le nom du frère de la princesse Leia (dans Star Wars), et la monnaie nationale en Hongrie.

Les questions ont été saisies au clavier par les 8 enfants de manière semi-spontanée, c'est à dire qu'elles ne leur ont pas été dictées. De plus il était supposé que les enfants connaissaient au moins oralement tous les mots constituant la question, en particulier les noms propres. Pour chaque question, les quatre étapes suivantes ont été suivies par l'orthophoniste : la réponse est dite à l'enfant dans une phrase (*Le maire de Bastia s'appelle X*) ; elle demande à l'enfant quelle question il poserait pour obtenir cette réponse (*que me demanderais-tu pour que je te réponde X ?*) ; l'enfant tape la question qu'il imagine ; l'enfant relit la question qu'il vient de taper et éventuellement corrige ce qu'il veut.

Le corpus ainsi obtenu, bien que de taille réduite (37 questions), est très représentatif car il permet beaucoup d'observations communes aux huit participants. En premier lieu il apparaît clairement que la plupart des observations faites classiquement sur les manuscrits d'enfants dyslexiques ne sont pas validées sur les écrits typographiés. 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, il n'est donc plus question de latéralisation), mais également à une plus grande motivation pour la frappe au clavier impliquant une plus grande attention au niveau de la production comme de la relecture (constatation rapportée par plusieurs orthophonistes et enseignants spécialisés). 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 phonème-graphème et des erreurs de segmentation des mots. Cela signifie à la fois une écriture phonétique mais pas nécessairement simpliste des mots (ainsi, *monnaie* s'écrit *monné, monais, moner, monnaie, moner, monaie* ou *monai*), et une segmentation en mots erronée (*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 même 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 d'un même utilisateur, elle devrait nécessairement être

dynamique car les utilisateurs sont généralement en cours d'apprentissage et les erreurs type peuvent évoluer.

Un second corpus, dit de *validation*, avec des questions plus variées d'un enfant à l'autre (portant sur le nombre d'habitants au Liban ou en France, sur le lieu des jeux olympiques une année donnée, sur le premier homme sur la lune, sur la date de la chute du mur de Berlin, sur l'âge de décès du plus gros homme au monde, ...), a été recueilli afin de valider nos propositions. Il a été réalisé par les auteurs auprès de 11 enfants d'une classe d'enseignement spécialisé pour des dyslexiques (CLasse d'Intégration Scolaire des Grands Cyprès à Avignon), entre 8 et 11 ans. Avant de faire saisir les questions, nous nous sommes cette fois assuré que les enfants avaient connaissance des éléments la constituant (qu'ils avaient déjà entendu parler de la ville de Berlin ou du Liban par exemple). Chaque enfant a tapé entre 3 et 6 phrases, le corpus est composé de 46 questions.

Pour étudier la lecture, nous disposons 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 (Lété et Ducrot, 2007). Ces 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). *Le chien de ma grand-mère aime beaucoup jouer avec mes chaussons* est un exemple où l'on peut observer qu'un mot composé compte pour le nombre de mots qui le composent.

3. Réécriture de questions dysorthographiées

Les applications de type correction orthographiques proposées à l'heure actuelle fonctionnent sur le postulat que les séquences de mots sont correctement identifiables, et que les erreurs sont isolées (pour les systèmes utilisant les informations syntaxiques et sémantiques notamment). Une étude menée par (James et Draffan, 2004) a permis de mesurer les performances très faibles des correcteurs orthographiques grand public pour des utilisateurs dysorthographiques. Comme l'a montré l'analyse de notre corpus de questions dysorthographiées, un traitement au niveau de la phrase s'impose. La majorité des erreurs étant de nature graphémique et non phonétique, cela suggère un traitement phonétique au niveau de la phrase entière. Cela lève à la fois le problème de la segmentation en mots et celui des homophones. Les outils de la reconnaissance automatique de la parole offrent des performances intéressantes dans ce domaine, notamment grâce à l'utilisation de modèles de langage.

3.1. *Combinaison d'alternatives graphémiques et phonologiques pour l'interprétation*

Une fois oralisées, la plupart des phrases de notre corpus deviennent compréhensibles et interprétables par des êtres humains, et la question de la segmentation en mots ne se pose plus. A partir de ce constat, nous avons émis l'idée d'un système automatique fonctionnant sur ce principe, réalisant séquentiellement une synthèse vocale de la phrase et une transcription du signal généré par la synthèse. La figure 2 illustre en détails la séquence de processus impliqués.

Les systèmes de synthèse vocale permettent de transformer une séquence de lettres en une séquence de sons parlés, appelés phonèmes. Pour cela, ils s'appuient sur un lexique de correspondances connues pour un ensemble de mots (le lexique phonétique) ainsi que sur un ensemble de règles de conversion des graphèmes vers les phonèmes, plus ou moins spécifique. De plus une identification des catégories morpho-syntaxiques est nécessaire, étant donné que la prononciation de certains graphèmes peut dépendre de leur rôle. Ainsi, dans l'exemple bien connu *les poules du couvent couvent*, le *ent* ne se prononce pas dans le verbe en tant que terminaison flexionnelle, mais il se prononce dans le nom. Au mieux, les systèmes de synthèse peuvent proposer plusieurs hypothèses sur la suite de phonèmes correspondant à une phrase écrite, dans les cas où une ambiguité de prononciation n'est pas levée.

Les systèmes de transcription automatique utilisent des modèles de langage associés aux lexiques phonétiques. A partir d'une séquence sonore, ils extraient les phonèmes correspondant (modulo une certaine probabilité de reconnaissance) 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. Un modèle de langage est constitué des probabilités d'occurrences de séquences de 1, 2, ou n mots, apprises sur un corpus d'exemples textuels. Ces séquences sont appelées *n-grammes*.

En pratique, nous nous sommes rendu compte qu'on ne peut pas se contenter d'une unique séquence de phonèmes correspondant à la phrase, car les confusions phonétiques (entre les voyelles ouvertes et fermées notamment) n'étaient pas prises en compte. Il faut donc construire un graphe de phonèmes et non pas une séquence de phonèmes. D'autre part, les omissions et les inversions de lettres ne peuvent pas être traitées si l'on s'en tient aux règles de conversion, et générer toutes les possibilités de ce type (en générant des arcs supplémentaires dans le graphe de phonèmes) risquerait d'apporter trop de confusions. Une solution à cela est de générer un graphe d'hypothèses de mots qui représente la séquence écrite, puis de phonétiser toutes les phrases issues de ce graphe de mots afin de générer un graphe de phonèmes plus complet. Les hypothèses de mots peuvent être obtenues à l'aide d'un correcteur orthographique, la plupart se basant sur des distances d'édition.

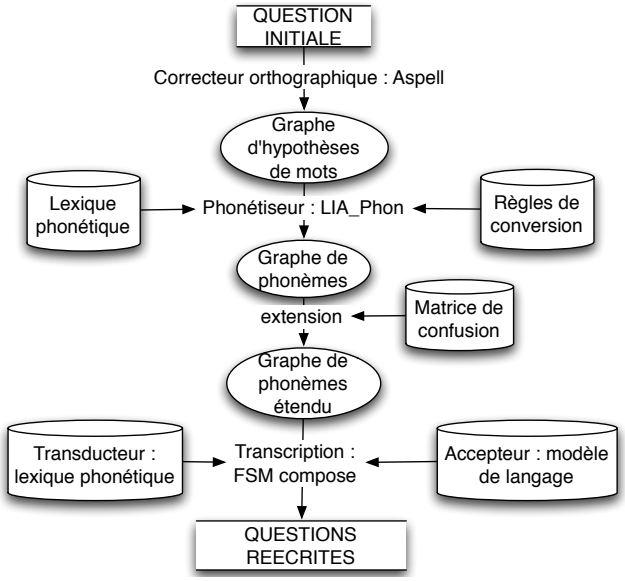

Figure 2. *Etapes de la réécriture d'une question : état initial et final, représentations intermédiaires, outils pour les transitions entre les représentations, données utilisées par ces outils.*

Dans les graphes représentant la phrase aux étapes intermédiaires, les arcs portent les phonèmes ou les mots de la phrase et les coûts de transition qui leur correspondent. Les nœuds sont les étapes intermédiaires, entre deux phonèmes ou entre deux mots écrits. Ainsi les différents chemins correspondant à des hypothèses graphiques ou phonétiques ont un coût associé correspondant à la somme des coûts de transition du chemin emprunté. Le chemin correspondant exactement à ce qui a été écrit doit avoir un coût nul, et plus on s'en écarte plus le coût doit être important. A la fin les propositions les plus vraisemblables sont obtenues par une recherche des plus courts chemins dans le graphe complet.

Un poids Wg est attribué aux mots alternatifs H (hypothèses graphémiques) proposés par un correcteur orthographique :

$$Wg(H) = f(d(H, I)) \quad [3]$$

où f est une fonction de normalisation de la distance $d(H, I)$ entre l'hypothèse proposée et le mot initialement écrit. Cette distance peut être fournie par le correcteur, ou calculée *a posteriori* (distance d'édition par exemple).

Un poids Wp est attribué aux alternatives phonétiques H (hypothèses phonétiques) obtenues à l'aide d'une matrice de confusion :

$$Wp(H) = g(m(H, I)) \quad [4]$$

où g est une fonction de normalisation d'une distance $m(H, I)$ entre le phonème alternatif et le phonème initial. Cette distance peut faire partie intégrante de la matrice de confusion.

3.2. *Mise en œuvre*

Le cadre théorique et applicatif proposé par les machines à états finis (FSM) (Mohri *et al.*, 2002) pour la reconnaissance automatique de la parole correspond à nos besoins de représentations intermédiaires en graphes, à travers leur implémentation dans le AT&T FSM Toolkit (Mohri *et al.*, 1997). En effet l'implémentation de modèles de langage dans le formalisme des automates telle que proposée par (Allauzen et Mohri, 2005) avec la bibliothèque *grm* permet de les utiliser pour décoder un automate construit à partir du graphe de phonèmes étendu, à condition de le coupler avec un transducteur permettant de faire correspondre des séquences de phonèmes à des mots écrits. Il s'agit alors de rechercher les meilleurs chemins en fonction des coûts de transition associés dans le graphe composé des hypothèses phonétiques, du modèle de langage et du transducteur déduit du lexique phonétique. Le modèle de langage est ici appris sur un mois d'articles du journal *Le Monde* disponibles dans le corpus de la campagne EQUER (Ayache *et al.*, 2006).

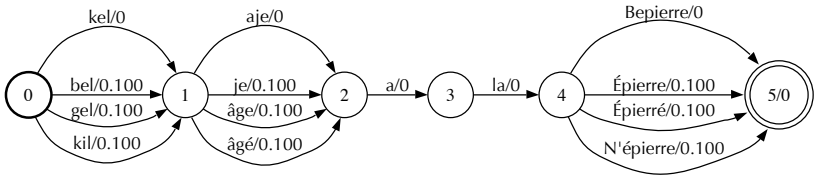

Figure 3. *Graphe de mots pour la phrase "kel aje a la Bepierre".*

Les hypothèses graphémiques permettant de générer le graphe de mots sont obtenues en retenant pour chaque mot (groupe de lettres accolées sans espace), s'il y a lieu, les trois premières hypothèses proposées par correcteur orthographique libre GNU ASPELL [3] qui utilise à la fois des distances d'édition de Levenshtein 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 [4]. La figure 3 montre un exemple de graphe

3. `http://aspell.sourceforge.net`
4. `http://aspell.net/test/`

de mots ainsi construit. La phonétisation est effectuée à l'aide de l'outil LIA_phon (Bechet, 2001), qui dispose à la fois d'un lexique phonétique de 80 000 mots et d'un système de 1996 règles de conversions ordonnées des plus générales aux plus exceptionnelles. 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 illustre le graphe de phonèmes étendus correspondant au graphe de mots de la figure 3.

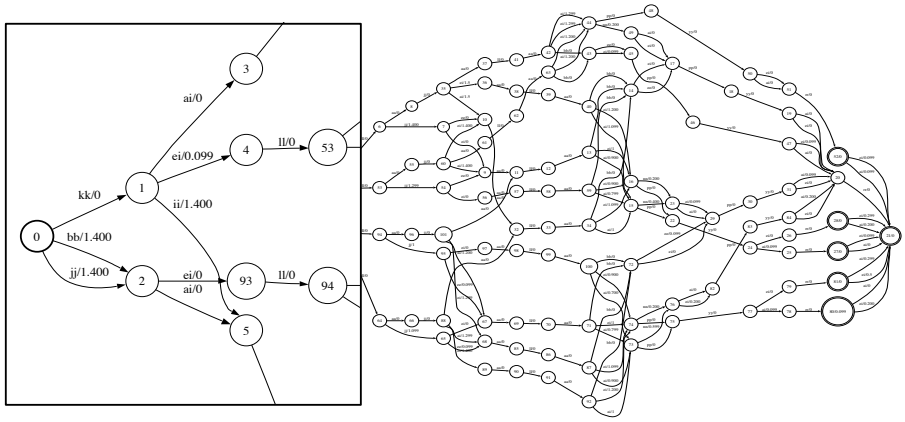

Figure 4. *Graphe de phonèmes étendu de la phrase "kel aje a la Bepierre" (phonème/poids, noeuds intermédiaires).*

Phrase	Coût du chemin
quel âge a l' abbé pierre	45,7884293
quel âge à l' abbé pierres	46,495369
quel âge alla et pierre	48,4406662

Tableau 1. *Réécritures les moins coûteuses de "kel aje a la Bepierre".*

Les fonctions de coût normalisé associées aux alternatives graphémiques ou phonétiques des équations (1) et (2) ont été établies de manière empirique avec :

$$f(d(H,I)) = \begin{cases} 0 \; if \; H = I \\ 0.1 \; if \; H \neq I \end{cases} \quad [5] \quad g(m(H,I)) = \begin{cases} 0 \; if \; H = I \\ 0.1 \; if \; H \neq I \end{cases} \quad [6]$$

Le graphe final est composé à partir du graphe d'hypothèses phonétiques, de l'accepteur du modèle de langage et du transducteur du lexique phonétique. Il prend en compte les coûts de transition issus de chacun des automates, et permet d'affecter des coûts à tous les chemins possibles. Le tableau 1 représente les trois hypothèses les plus probables selon un calcul du plus court chemin dans le graphe de la figure 4 et les

coûts (somme des coûts des transitions) associés à chacune de ces hypothèses. Dans ce cas, l'hypothèse attendue est la première.

Le problème qui se pose alors est l'utilisation des meilleures hypothèses en entrée d'un système de questions réponses, et plus généralement d'un système de recherche d'information. Une première solution consiste à poser les questions correspondant aux trois (ou N) premières réécritures, puis de décider de la réponse la plus pertinente à l'aide de scores de confiance affectés à l'ensemble des réponses à chaque question réécrite. Des méthodes adaptées aux questions réponses pour le calcul de scores de confiance ont notamment été proposées dans (Sitbon *et al.*, 2006). Une autre manière d'exploiter le graphe est d'extraire des probabilités *a posteriori* pour chacun des mots contenus dans le graphe. Pour cela, on calcule pour chaque mot le ratio entre la somme des poids des chemins qui passent par ce mot et la somme totale des chemins. Il en résulte alors un score pour chaque mot, et il devient possible de créer un vecteur pondéré des mots clés de la requête. L'avantage de cette technique par rapport à la précédente est qu'elle permet de considérer de manière unique un ensemble d'hypothèses. Il se pourrait en effet que l'intention exprimée se trouve répartie sur deux hypothèses, comme dans l'exemple suivant (qui n'est qu'académique). La phrase *Qui est le maire de Bastia* mal entrée pourrait donner après interprétation les hypothèses : *qui est la mère de Bastien ?*, *comment est la mer à Bastia ?*, *Où est le maire du Bastion ?*. Dans ce cas, chaque hypothèse prise séparément est fausse, mais les mots clés importants se trouvent tous dans la réunion des trois hypothèses.

3.3. *Evaluation*

Dans le cadre de notre étude sur la communication avec un système de recherche d'information, plusieurs hypothèses de réécriture de la requête initiale sont envisageables. Les systèmes utilisant généralement les lemmes de mots de la requête au lieu des formes, une hypothèse contenant des fautes d'accord sera acceptable. Il s'agit bien de réécriture en vue d'un traitement automatique et non de correction orthographique.

3.3.1. *Evaluation des hypothèses de réécriture*

Nous avons constitué une référence de la même manière que l'on transcrit manuellement les textes audio pour tester la reconnaissance : 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. La plate forme d'évaluation des outils de reconnaissance de la parole SCTK [5] 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 la phrase correspondante qui peut contenir plusieurs hypothèses (représentées par un graphe de mots), en prenant en compte les insertions, omissions et les substitutions.

5. `http://www.nist.gov/speech/tools`

Dans le cadre d'une réécriture en entrée d'un système de recherche d'information, il n'est pas nécessaire que tous les mots de la question soient corrects ni qu'ils soient bien accordés. En effet la plupart des systèmes effectuent en premier lieu une lemmatisation et un filtrage des requêtes, c'est à dire que les mots outils sont retirés et les mots fléchis sont ramenés à leur lemme. Par exemple la phrase *Comment s'appellent le maires des Bastia* sera traitée à l'identique de *Comment s'appelle le maire de Bastia* via la phrase lemmatisée *Comment se appeler maire Bastia*. Ainsi nous proposons pour l'évaluation de comparer les versions lemmatisées des phrases de référence et des phrases réécrites. De plus en accord avec un modèle étendu pour les hypothèses de la requête en entrée des systèmes, l'évaluation peut prendre en compte par exemple les trois premières hypothèses du système de réécriture, ou uniquement la première hypothèse. Les hypothèses fournies par le système de combinaison étant indépendantes, dans le cas de l'évaluation des trois premières hypothèses c'est le score de la phrase proposée la plus proche de la référence qui sera retourné.

Afin de comparer les performances de la combinaison phonétique et graphémique avec l'utilisation d'un correcteur orthographique pour la réécriture, nous avons évalué un système de réécriture basé uniquement sur les hypothèses fournies par Aspell, qui correspondent en réalité aux graphes de mots tels que celui de la figure 3. Dans ce cas, l'évaluation des trois premières hypothèses correspond à l'évaluation du chemin le plus proche de la référence. L'apport spécifique du correcteur orthographique et de la matrice de confusion à l'intérieur du système de combinaison n'a pas été quantifié précisément. Pourtant il est apparu clairement sur un certain nombre d'exemples que leurs traitements s'avéraient nécessaire. Le correcteur orthographique par exemple intervenu notamment pour les cas d'inversion et d'omission de lettres.

Le tableau 2 contient les résultats de l'évaluation par SCLITE des phrases d'origine (Initial), des premières hypothèses du système graphémiques (Asp 1) et du système de combinaison (FSM 1), ainsi que des trois premières hypothèses fournies par ces systèmes (Asp 3 et FSM 3). L'évaluation est effectuée selon deux critères différents : le taux de mots corrects et le taux de questions correctes (pourcentage de réécritures totalement correctes), qui constitue un aperçu des cas où l'on est certain que le système aura la possibilité de répondre, étant donné qu'il contient les même informations que la phrase de référence (il contient également des informations bruitées dans le cas où l'on considère plusieurs hypothèses de réécriture). Les résultats pour les

Mesure	Initial	Asp 1	Asp 3	FSM 1	FSM 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

Tableau 2. *Taux de lemmes corrects et pourcentage de phrases identiques à la référence après lemmatisation et filtrage, sur les phrases tapées initialement ou réécrites à l'aide de Aspell (Asp) ou de notre système (FSM), pour 1 ou 3 hypothèses.*

trois premières hypothèses montrent que si l'amélioration en terme de taux de mots corrects est déjà importante dans l'absolu (de 30% rapport à l'initial), elle l'est aussi

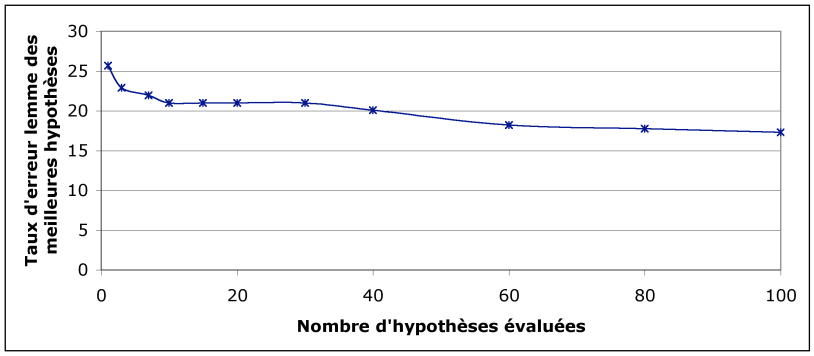

Figure 5. *Taux d'erreurs lemmes des meilleures hypothèses parmi un nombre variable de propositions pour chaque question entrée.*

par rapport à un correcteur orthographique performant (dont le taux est de 11% plus bas). Les résultats en terme de taux de mots corrects sont également probants si l'on ne considère que la première hypothèse du système par combinaison, ce qui laisse à penser que l'ajout de bruit qu'apporteraient des hypothèses multiples sera peut être plus néfaste que la perte engendrée par la conservation d'une seule hypothèse. Cela est confirmé par les résultats au niveau des phrases. En effet, on atteint 43,2 % de phrases identiques à la référence après filtrage et lemmatisation de la première hypothèse FSM, alors qu'il n'y en avait que 5,4 % à l'origine et qu'on atteint moins de 20% avec Aspell. La différence avec l'évaluation des trois premières hypothèses FSM est significative mais faible.

La pertinence de l'utilisation de plus de trois hypothèses peut être remise en cause, étant donnée l'évolution faible de performance entre FSM1 et FSM3. Le graphe de la figure 5 reporte l'évolution du taux d'erreurs lemmes (taux d'erreur que l'on cherche à minimiser, portant sur les mots et prenant en compte les insertions, substitutions et suppressions) pour un nombre variable d'hypothèses issues de la combinaison FSM prises en compte. L'évolution très lente au delà des trois premières hypothèses suggère que l'apport en performance de plus d'hypothèses risque de ne pas compenser la perte d'efficacité due au bruit que cela implique.

Mesure	Initial	Asp 1	Asp 3	FSM 1	FSM 3
Taux de lemmes corrects	**37**	44,4	48,8	**59,7**	61,2
Taux de phrases correctes	0	4,3	4,3	7,1	9,5

Tableau 3. *Taux de lemmes corrects et taux de phrases correctes (après filtrage et lemmatisation) pour les différentes méthodes appliquées au corpus de validation (recueilli à l'école).*

Les résultats obtenus par les premières hypothèses du système de réécriture par combinaison sont très bons d'autant qu'il n'y a pas de dégradation des parties de phrases déjà correctes. Nous avons appliqué la même méthode sur le corpus de validation, composé des 46 questions recueillies à l'école primaire. Les résultats obtenus sur ce second corpus, présentés dans le tableau 3 sont beaucoup moins bons. Parmi les explications, on peut déjà noter que le taux de mots corrects au départ était beaucoup plus bas. Par ailleurs, la pression intellectuelle exercée par l'orthophoniste est certainement plus importante, étant donné que dans le premier cas le corpus a été réalisé lors de séances de rééducation. De plus, parmi les enfants de la classe où a été menée l'étude, il y a deux enfants dysphasiques. Ces enfants ont des difficultés de production du langage qui s'ajoutent aux difficultés d'écriture, ce qui a généralement ôté de la cohérence aux questions tapées. Un des enfants a produit ses phrases sans aucun espace entre les mots. Pour des cas comme celui-ci, il faudrait pouvoir générer plusieurs hypothèses de phonétisation sur un même mot, ce qui jusque là n'était pas effectué afin d'éviter une explosion des hypothèses phonétiques. Globalement, la baisse de performances sur les données recueillies à l'école suggère que les résultats sont variables selon les individus.

Les performances du système pour les questions de chaque enfant sont consignées dans le tableau 4. Elles montrent que si les variations au sein du groupe du corpus initial existent entre les individus, à part pour les enfants 1 et 4, elles ne sont pas très significatives étant donné qu'elles s'appliquent sur cinq exemples au maximum. La répartition des résultats par thème a par ailleurs fait ressortir une nette différence, et l'on remarque notamment que les thèmes de questions 1 et 2 maintiennent des taux d'erreurs importants et que le système ne parvient à une phrase lemmatisée identique à la référence dans aucun cas. La raison de ces erreurs est que les noms propres associés à ces questions ne se trouvent ni dans le lexique phonétique ni dans le modèle de langage et sont par conséquent impossibles à proposer dans les hypothèses. Cela suggère que les performances du système par combinaison de processus graphémiques et phonétiques pourront encore être améliorés par un enrichissement dynamique des ressources, ou par un enrichissement statique se basant sur l'ensemble du corpus sur lesquelles les questions sont posées. En effet le modèle de langage a ici été appris sur un sous ensemble du corpus EQUER, et l'on peut imaginer y ajouter les phrases contenant des mots inconnus du lexique et du modèle initial.

3.3.2. *Evaluation des vecteurs de mots pondérés*

L'évaluation de vecteurs de mots pondérés s'avère plus complexe que la comparaison de phrases étant donné qu'il faut prendre à la fois les mots corrects, les mots incorrects, et les mots manquants. Le tableau 5 propose l'un des vecteurs correspondant à la question sur l'âge de l'abbé Pierre. Cependant, contrairement aux cas d'expansion de requête, où les mots ajoutés peuvent avoir un certain degré de pertinence, il est peu probable dans notre cas qu'un mot "alternatif" soit pertinent pour la requête, si l'on considère les lemmes et non les formes. Par ailleurs nous cherchons ici à obtenir un indice de la performance de notre système de génération de requêtes pondérées et non la performance du système dans sa globalité, ce qui écarte encore une fois l'idée de

Premier corpus

enfant	Initial	FSM3
0	63	89
1	37	65
2	51	89
3	59	78
4	31	92
5	70	81
6	54	83
7	40	70

Corpus de validation

enfant	Initial	FSM3
1	0	48
2	43	58
3	28	46
4	42	58
5	41	52
6	25	40
7	47	64
8	37	48
9	73	79
10	39	82
11	38	73

Tableau 4. *Taux de lemmes corrects pour chaque enfant, calculé pour les phrases tapées initialement et pour les trois premières hypothèses du système s'appuyant sur les FSM, pour le corpus initial et le corpus de validation.*

quel	âge	à	et	la	abbé	pierre	pierres
0,98	0,98	0,98	0,06	0,06	0,94	0,64	0,29

Tableau 5. *Vecteur d'hypothèses de la phrase "quelle age a l'abée pierre".*

l'évaluation en contexte. Nous proposons en premier lieu une mesure de la densité des lemmes non vides corrects $Q_{correct}$ dans la requête pondérée Q_{pond}, qui se rapproche de la notion de précision en recherche d'information. Le poids d'un mot m dans la requête pondérée $p(m, Q_{pond})$ est défini par sa probabilité *a posteriori*, qui est nulle si le mot n'apparaît pas dans la requête. La densité se calcule en fonction du poids des mots corrects dans la requête par rapport au poids total des mots proposés dans la requête (équation 7). Une seconde mesure, l'indice de présence, permet de mesurer la quantité de mots clés proposés dans le vecteur par rapport à ceux qui apparaissent pour la version corrigée de la question. Il est possible de comparer cette mesure aux mesures de rappel en recherche d'information. L'indice de présence $pres(Q_{pond}, Q_{correct})$ des mots clés initiaux dans la requête pondérée est calculé selon l'équation 8.

$$dens(Q_{pond}, Q_{correct}) = \frac{\sum_{m \in Q_{correct}} p(m, Q_{pond})}{\sum_{m \in Q_{pond}} p(m, Q_{pond})} \quad [7]$$

$$pres(Q_{pond}, Q_{correct}) = \frac{\sum_{m \in Q_{correct}} p(m, Q_{pond})}{|m \in Q_{correct}|} \quad [8]$$

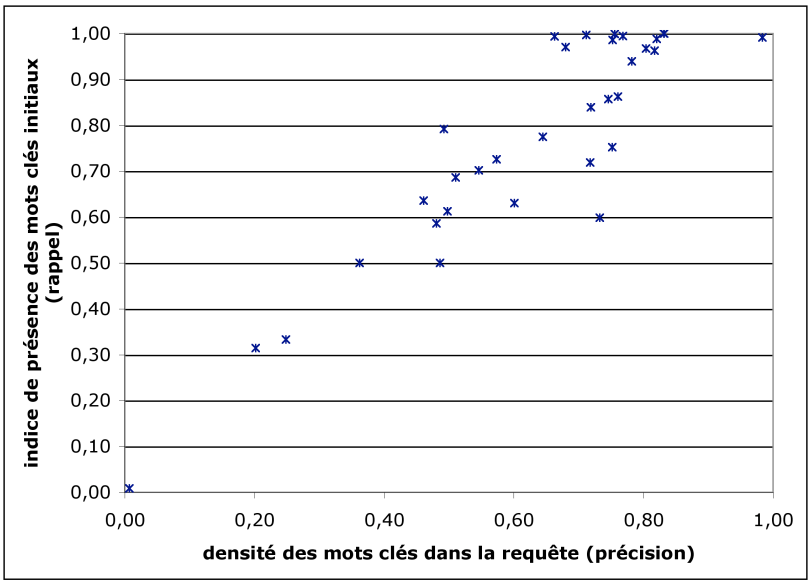

Figure 6. *Positionnement des vecteurs de requêtes pondérés en fonction de leur indice de présence et de leur densité.*

La figure 6 montre les densités des requêtes pondérées correspondant à chacune des questions tapées dans le corpus initial, ainsi que l'indice de présence des mots clés corrects correspondant. Cette représentation individualisée montre que, à part de rares cas d'échec total, les densités des requêtes proposées varient de 0,5 à 1, avec plus d'un tiers des requêtes au-dessus de 0,8. L'indice de présence est généralement situé au-dessus de 0,6, et proche de 1 pour un tiers des questions. Cette seconde valeur prenant en compte le poids des mots clés initiaux dans le vecteur de mots pondérés, un score au dessus de 0,6 est acceptable pour une densité au delà de 0,5, en estimant que les mots corrects auront tendance à se retrouver dans des documents, et lever l'ambiguïté générée par le bruit. En effet le poids d'un document contenant tous les mots corrects, même avec un poids de 0,5 dans le vecteur pondéré, sera plus important que celui d'un document contenant un unique mot erroné ayant un poids plus important.

3.4. *Conclusions sur la réécriture*

Les performances obtenues par un système combinant des aspects graphémiques et phonétiques au niveau de la phrase entière permettent de proposer des réécritures qui multiplient par 8 le nombre de questions correctement orthographiées une fois lemmatisées. L'évaluation des phrases filtrées et lemmatisées montre que l'on peut faire descendre le taux d'erreurs de 51% à 23% en considérant uniquement la première

hypothèse, alors qu'un correcteur orthographique performant ne permet de descendre qu'à 35,7%.

Une observation détaillée des résultats a montré que ce n'était pas le processus qui était en cause dans les erreurs qui restent, mais la sur-abondance 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 les hypothèses. Cependant la multiplication des hypothèses risque à son tour de générer du bruit. Le degré de grammaticalité pourrait alors être une piste de la pondération d'hypothèses, même de parties d'hypothèses.

Une évaluation des performances sur des phrases entières a également été réalisée. D'une manière générale, le gain en performance est de 10% par rapport à Aspell, et de 20% par rapport aux phrases d'origine, lorsque l'on considère 3 hypothèses. Une des explications est que les mots outils sont souvent initialement correctement orthographiés. De plus, l'interprétation par les phonèmes génère des fautes d'accords. L'utilisation de ce système pour de la correction orthographique apporterait l'avantage de meilleures performances, mais soulève en même temps la question difficile de l'interface en l'absence d'alignement mot à mot entre les erreurs et les propositions.

4. Estimation de la lisibilité d'une phrase pour un lecteur dyslexique

Si l'on considère 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. En ce sens, les techniques d'apprentissage par régression sont adaptées pour fournir une mesure continue de la lisibilité d'une phrase. En premier lieu, l'expérimentation a porté sur 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 régression linéaire est choisie pour sa capacité à fournir une mesure transparente, combinant linéairement des paramètres les plus discriminants.

4.1. *Apprentissage automatique pour l'établissement d'une mesure de lisibilité*

Les paramètres utilisés sont les mêmes que ceux impliqués dans les approches fondées sur l'apprentissage évoquées dans la section 3.1, 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. La complexité lexicale de la phrase est le critère utilisé pour les normo-lecteurs. Elle est reflétée par la fréquence d'apparition lexicale des mots qui

la compose, ainsi que la longueur moyenne de ces mots. Dans le cas des enfants, les fréquences sont disponibles dans la base de données Manulex (Lété *et al.*, 2004). La figure 7 illustre les données utilisées pour refléter les paramètres d'une phrase.

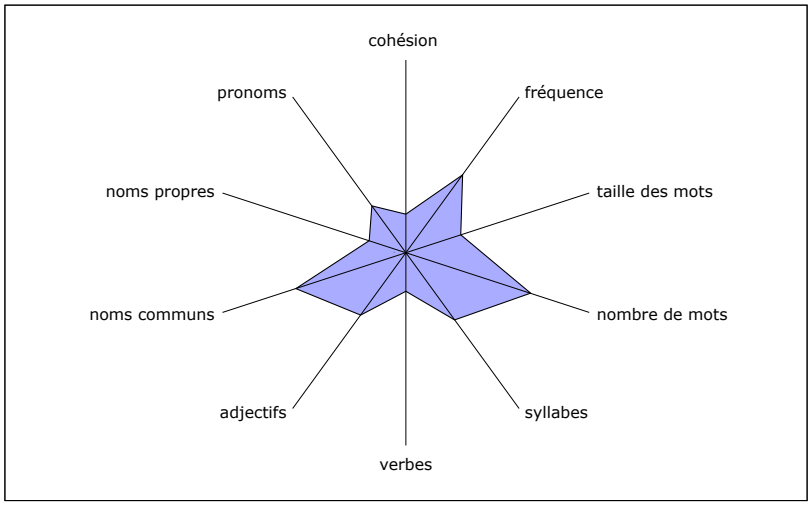

Figure 7. *Paramètres de la lisibilité d'une phrase pour un lecteur dyslexique.*

Les expériences ont été réalisées à l'aide de la plate-forme WEKA [6] (Ian H. Witten, 1999) afin de trouver la meilleure mesure de la lisibilité d'une phrase. Les temps de lecture des mots sont normalisés pour chaque utilisateur sur une échelle allant de 0 à 100. Les temps de lecture normalisés des phrases sont les moyennes des temps de lecture normalisés des mots les constituant. Aucune normalisation n'est effectuée par rapport à la taille des phrases ni par rapport à la taille des mots.

L'ensemble des paramètres décrivant une phrase peut être visuellement représenté par des dimensions paramétriques telles que proposées par la figure 7. Formellement, les paramètres sont basés sur les propriétés de chacun des mots m_i composant la phrase : sa longueur en nombre de lettres ($S(m_i)$), sa fréquence, sa catégorie morpho-syntaxique $Cat(m_i)$, et son taux de cohésion grapho-phonémique ($C(m_i)$). Les catégories morpho-syntaxiques sont obtenues à l'aide du TreeTagger (Schmid, 1994). Le taux de cohésion grapho-phonémique est estimé à l'aide de l'équation 9. Il s'agit d'une estimation qui part du principe qu'à un phonème correspond au moins une lettre et qu'à une lettre seule correspond au plus un phonème.

$$C(m_i) \quad = \frac{nombre\ de\ phonemes\ dans\ m_i}{S(m_i)} \qquad [9]$$

Les paramètres pour une phrase S, explicités dans l'équation 11, sont calculés en fonction d'une moyenne ou d'un comptage des paramètres des mots la composant. Les

6. `http://www.cs.waikato.ac.nz/~ml/`

paramètres calculés sont la longueur moyenne des mots $S(S)$, leur fréquence moyenne $F(S)$ et leur cohésion moyenne $C(S)$. Pour les catégories morphosyntaxiques, on déterminera autant de paramètres T que de catégories Cat_j choisies, et chacun sera égal au nombre de mots étiquetés avec la catégorie morphosyntaxique correspondante.

$$T_{Cat_j}(S) = |Cat(m_i) = Cat_j|, \quad Cat_j \in \{verbe, nom, adjectif, adverbe, ...\} \quad [10]$$

4.2. *Résultats*

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 6 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 (ZeroR affecte la valeur moyenne des données d'entraînement à toutes les données de test), et un classifieur aléatoire (qui affecte des valeurs entre 0 et 100 en suivant une distribution gaussienne). 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 étant donné qu'elle fournit une mesure transparente pour des résultats équivalents aux SVM. La mesure de lisibilité ainsi obtenue est définie par :

$$L = 1.12 * ADV - 0.69 * CON + 6.48 * cohesion + 15.58 \quad [11]$$

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	ZeroR	Aléatoire
mots	**9,38**	9,74	10,1	37,97
phrases	5,01	**5,00**	5,07	35,69

Tableau 6. *Taux d'erreurs (validation croisée 10 plis) des classifieurs SVM et régression linéaire, d'un classifieur basé sur la moyenne des données disponibles (ZeroR) et d'un classifieur aléatoire, en prédiction des temps de lecture de mots ou phrases.*

Par ailleurs nous avons évalué la pertinence de mesures adaptées à chaque utilisateur. Pour cela nous avons établi et testé indivuellement des prédictions de temps de lecture de mots (les données de phrases lues étant trop peu nombreuses). Les résultats présentés dans le tableau 7 suggèrent que certains lecteurs ont un comportement très régulier pour lequel une mesure de lisibilité serait favorisée par un modèle individuellement adapté. Cependant l'utilisation de modèles sur les mots n'est pas suffisante pour estimer la lisibilité d'une phrase car elle ne prend pas en compte les relations syntaxiques entre les mots.

	Algorithme	
Enfant	SMOReg	LinReg
1	7,35	7,86
2	9,15	9,85
3	6,93	7,25
4	7,50	7,97
5	8,28	9,43
6	7,28	7,27
7	7,50	8,17
8	11,20	12,29
9	12,59	13,39

Tableau 7. *Taux d'erreurs (obtenus par validation croisée 10 plis) des classifieurs testés (SVM et régression linéaire) lors de l'établissement de modèles propres à chaque utilisateur pour prédire le temps de lecture de mots.*

4.3. *Perspectives sur la lisibilité*

Il est possible de réécrire le score de pertinence d'un document, d'un passage, ou d'une phrase (d'un résumé) comme une combinaison linéaire de sa pertinence originelle et de sa lisibilité. Les expériences sont en cours sur les données des campagnes d'évaluation CLEF[7] et TREC[8] en recherche d'information *ad-hoc*, et DUC[9] en résumé automatique. Les premiers résultats suggèrent que la lisibilité moyenne peut être augmentée significativement tout en conservant de bonnes performances sur le plan informationnel lorsque le score de pertinence est composé d'environ 30% de score de lisibilité et 70% de score de similarité.

Les phrases dont nous disposons pour l'apprentissage sont toutes de 12 mots, et un nouveau recueil de temps de lecture est en cours pour des phrases de longueur et de syntaxe plus variables. Par ailleurs, la difficulté de lecture de mots pourra être utilisée afin d'étudier la répartition de mots difficiles dans un texte. En effet un mot difficile sera mieux lu dans un contexte compris (avec des mots faciles).

Une autre solution consiste à réduire la quantité de texte à faire lire à l'utilisateur, sans trop réduire pour autant la quantité d'informations. 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 pourront également être intégrées à ces tâches de sélection de phrases ou de passages.

7. `www.clef-campaign.org`
8. `http://trec.nist.gov`
9. `http://duc.nist.gov/`

5. Conclusions et perspectives

Nous avons proposé deux solutions pour prendre en charge respectivement les difficultés de lecture et d'écriture liées à un handicap. Les solutions s'appuient sur les spécificités de la nature du déficit, à savoir une dégradation de la conscience phonologique. Elles sont ainsi plus adaptées que les solutions classiques. Les évaluations ont montré des résultats encourageants. Les techniques que nous proposons pourraient par ailleurs être utilisées en vue d'un diagnostic. L'évaluation du système de réécriture permettrait de déterminer si la cause principale de la dysorthographie est bien le déficit phonologique, ou si elle rejoint l'hypothèse temporelle qui induit des inversions de phonèmes dans la transcription, ou encore si elle correspond à l'hypothèse auditive de confusion de sons.

6. Bibliographie

Allauzen C., Mohri M., « The design principles and algorithms of a weighted grammar library », *International Journal of Foundations of Computer Science*, vol. 16, n° 3, p. 403-421, 2005.

Ayache C., Grau B., Vilnat A., « EQueR : the French Evaluation campaign of Question Answering system EQueR/EVALDA », *LREC*, Genoa, Italy, p. 1157-1160, mai, 2006.

Bechet F., « LIA_PHON - Un système complet de phonétisation de textes », *revue T.A.L.*, vol. 42, n° 1, p. 47-67, 2001.

Boissière P., Bouraoui J.-L., Vella F., Lagarrigue A., Mojahid M., Vigouroux N., Nespoulous J.-L., « Méthodologie d'annotation des erreurs en production écrite. Principes et résultats préliminaires », *Actes de TALN*, vol. 2, Toulouse, France, p. 529-538, Juin, 2007.

Brill E., Moore R. C., « An improved error model for noisy channel spelling correction », *Proceedings of the 38th Annual Meeting of the ACL*, p. 286-293, 2000.

Chall J. S., Dale E., *Readability Revisited : The New Dale-Chall Readability Formula*, Cambridge : Brookline books, 1995.

collective E., *Dyslexie, dysorthographie, dyscalculie - Bilan des données scientifiques*, Les éditions Inserm, 2007.

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, novembre, 2005.

Coltheart M., Rastle K., « Serial processing in reading aloud : Evidence for dual-route models in reading », *Journal of experimental psychology*, vol. 20, p. 1197-1211, 1994.

Dale E., Chall J., « A formula for predicting readability », *Educational research bulletin*, vol. 27, p. 11-20, 1948.

Deorowicz S., Ciura M. G., « Correcting spelling errors by modelling their causes », *International journal of applied mathematics and computer science*, vol. 15, n° 2, p. 275-285, 2005.

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

Gillard L., Sitbon L., Bellot P., El-Beze M., « Dernieres évolutions de SQuALIA, le système de Questions/Réponses du LIA », *revue T.A.L.*, vol. 46, n° 3, p. 41-70, 2005.

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

James A., Draffan E., « The accuracy of electronic spell checkers for dyslexic learners », *PATOSS bulletin*, Août, 2004.

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 », *Actes de 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, Avril, 2004.

Lété B., Ducrot S., *La perception du mot écrit chez l'apprenti lecteur et l'enfant dyslexique : Évaluation en fovéa et en parafovéa*, vol. Acquisition du langage : approche intégrée, SOLAL, Marseille, p. 125-172, 2007.

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.

Loosemore R. P. W., « A neural net model of normal and dyslexic spelling », *International Joint Conference on Neural Networks*, vol. 2, Seattle, USA, p. 231-236, 1991.

Mohri M., Pereira F. C. N., Riley M., « Weighted Finite-State Transducers in Speech Recognition », *Computer Speech and Language*, vol. 16, n° 1, p. 69-88, 2002.

Mohri M., Pereira F. C. N., Riley M. D., « AT&T FSM LibraryTM – Finite-State Machine Library », , http ://www.research.att.com/ fsmtools/fsm/, 1997.

Pedler J., « The detection and correction of real-word spelling errors in dyslexic text », *Proceedings of the 4th Annual CLUK Colloquium*, p. 115-119, 2001.

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

Ramus F., Rosen S., Dakin S., Day B., J.M. C., White S., U. F., « Theories of developmental dyslexia : insights from a multiple case study of dyslexics adults », *Brain*, vol. 126, p. 841-865, 2003.

Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *International Conference on New Methods in Language Processing*, Manchester, UK, p. 44-49, 1994.

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

Sitbon L., Gillard L., Grivolla J., Bellot P., Blache P., « Vers une prédiction automatique de la difficulté d'une question en langue naturelle », *actes de TALN*, Louvain, Belgique, p. 337-346, 10-13 Avril, 2006.

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

Spooner R., A spelling checker for dyslexic users : user modelling for error recovery, PhD thesis, HCI Group, University of York, Heslington, York, Septembre, 1998.

Toutanova K., Moore R. C., « Pronunciation Modeling for Improved Spelling Correction », *Proceedings of the 40th annual meeting of ACL*, Philadelphia, p. 144-151, Juillet, 2002.

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.