

## Traitements phrastiques phonétiques pour la réécriture de phrases dysorthographiées

Laurianne Sitbon<sup>1,2</sup> Patrice Bellot<sup>1</sup> Philippe Blache<sup>2</sup>

(1) Laboratoire d'Informatique d'Avignon - Université d'Avignon

(2) Laboratoire Parole et Langage - Université de Provence

{laurianne.sitbon, patrice.bellot}@univ-avignon.fr, blache@lpl.univ-aix.fr

**Résumé** Cette article décrit une méthode qui combine des hypothèses graphémiques et phonétiques au niveau de la phrase, à l'aide d'une représentation en automates à états finis et d'un modèle de langage, pour la réécriture de phrases tapées au clavier par des dysorthographiques. La particularité des écrits dysorthographiés qui empêche les correcteurs orthographiques d'être efficaces pour cette tâche est une segmentation en mots parfois incorrecte. La réécriture diffère de la correction en ce sens que les phrases réécrites ne sont pas à destination de l'utilisateur mais d'un système automatique, tel qu'un moteur de recherche. De ce fait l'évaluation est conduite sur des versions filtrées et lemmatisées des phrases. Le taux d'erreurs mots moyen passe de 60% à 22% avec notre méthode, et est de 0% sur 43% des phrases testées.

**Abstract** This paper introduces a sentence level method combining written correction and phonetic interpretation in order to automatically rewrite sentences typed by dyslexic spellers. The method uses a finite state automata framework and a language model. Dysorthographics refers to incorrect word segmentation which usually causes classical spelling correctors fail. Our approach differs from spelling correction in that we aim to use several rewritings as an expression of the user need in an information retrieval context. Our system is evaluated on questions collected with the help of an orthophonist. The word error rate on lemmatised sentences falls from 60% to 22% (falls to 0% on 43% of sentences).

**Mots-clefs :** Réécriture de phrases, dyslexie, automates, correction orthographique

**Keywords:** Sentence level rewriting, dyslexia, FSM, spell checking

### 1 Introduction

*koman sapel le mer de bastya ?* Si votre système de conversion graphème-phonème et votre conscience phonologique fonctionnent parfaitement, vous devriez être avoir une intuition de l'intention de cette question. Peut-être même pouvez-vous y répondre ? En revanche, les systèmes automatiques pourtant très élaborés de questions réponses en resteront coi, dans le meilleur des cas ils répondront *la Méditerranée*, s'ils s'aident d'un correcteur orthographique. La raison principale est que ces systèmes ne sont pas conçus pour prendre en compte un profil linguistique de l'utilisateur, et dans le cas d'utilisateurs avec un handicap de langage ces systèmes ne sont généralement pas assez robustes. Les performances des correcteurs orthographiques sont par exemple faibles dans le cas où l'auteur est dysorthographique. Les correcteurs orthographiques

grand public supposent une segmentation en mots correcte, ce qui les rend très peu efficaces dans des cas d'écriture dysorthographique, comme le montre une étude menée par (James & Draffan, 2004). Or la dysorthographie provient d'un trouble de la conscience phonologique, qui implique généralement une écriture essentiellement phonétique et une segmentation en mots souvent erronée. La conscience phonologique permet de découper une séquence de phonèmes en unités sémantiques dans une phrase (Gillon, 2004). Si les unités sémantiques sont identifiées correctement, le passage à l'orthographe se fait ensuite par des voies de conversion phonème-graphème couplées avec des connaissances sur les exceptions orthographiques liées au mot.

Avant d'envisager un système de réécriture, la première section situe les besoins par rapport aux systèmes auxquels les réécritures s'adressent ainsi qu'aux utilisateurs à travers la constitution d'un corpus d'évaluation. La seconde section pose les bases d'un système dédié aux dysorthographiques qui s'appuie sur une combinaison d'hypothèses graphémiques et phonétiques au niveau de la phrase, ainsi que sa mise en application. La troisième section propose une évaluation de ce système sur le corpus constitué, en comparant ses performances à celles d'un correcteur orthographique.

## **2 Enjeux de la réécriture**

### **2.1 La réécriture vue comme l'expansion de la requête**

Dans le cadre de la recherche d'information, on pourra admettre plusieurs hypothèses de réécriture de la requête initiale à l'aide d'un modèle robuste capable de pondérer les différentes hypothèses, ce qui est pratiqué dans le cadre de l'expansion de requêtes. De plus, les systèmes utilisant généralement les lemmes de mots de la requête à la place des formes, une hypothèse contenant des fautes d'accord reste acceptable. Ainsi, il s'agit bien de réécriture en vue d'un traitement automatique et non pas de correction orthographique.

### **2.2 Constitution d'un corpus d'exemples de questions dysorthographiées**

Parmi les différents types de système de recherche d'information, nous nous sommes focalisés sur les systèmes de questions réponses car ils exigent une requête formulée en une phrase cohérente, soulèvent moins le problème du choix des termes employés que les systèmes admettant des mots clés comme requête.

Le corpus que nous avons recueilli est donc 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 (CE2) et 13 ans (4e)).

Le choix des questions a été guidé par les contraintes d'évaluation du système de questions réponses (SQR) SQuALIA (Gillard *et al.*, 2006) dont nous disposons ainsi que par le vocabulaire restreint des enfants. Nous avons sélectionné des questions factuelles de la campagne d'évaluation Technolanguage EQUER (Ayache *et al.*, 2006) pour lesquelles SQuALIA a fourni une bonne réponse (Gillard *et al.*, 2005), 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 se trouvent dans le tableau 1.

Qui est le maire de Bastia ? Quel âge a l'abbé Pierre ? Quelle est la capitale de Terre Neuve ? Qui est le frère de la princesse Leia ? Quelle est la monnaie nationale en Hongrie ?
--

TAB. 1 – Questions de la campagne EQUER retenues pour constituer le corpus de questions dysorthographiées.

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. 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 vient de formuler ;
- 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 phrases), 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 du 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 essentiellement des erreurs de conversion phonème-graphème au niveau de la phrase. 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, labpier, la Bepierre, labepier, labée pierre, l abepier, l'abée pierre* ou *labpier*). On rencontre 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*). Ces régularités pour un même utilisateur suggèrent la possibilité de définir des modèles individuels d'erreurs modélisant les transitions des orthographes erronées vers les orthographes correctes. Cependant la définition d'un modèle générique pour tous les utilisateurs se révèle impossible, étant donné que comme ces exemples le confirment, il existe autant de dyslexies que de dyslexiques. D'autre part la définition de modèles individuels devrait nécessairement être dynamique car les utilisateurs sont généralement en cours d'apprentissage et les erreurs type peuvent évoluer.

### 3 Un système de réécriture dédié

La réécriture peut se faire à l'aide d'un correcteur orthographique étant donné qu'ils proposent généralement plusieurs alternatives pour chaque mot rencontré hors de leur lexique. Cependant comme le démontre l'étude publiée dans (James & Draffan, 2004) les correcteurs grand public ne répondent généralement pas aux besoins spécifiques des dyslexiques, qui ont tendance à produire un mauvais découpage en mots ainsi qu'à la substitution d'homophones (lesquels homophones se trouvent généralement dans le lexique et ne sont donc pas repérés).

Des applications dédiées aux dyslexiques ont été réalisées, qui font généralement abstraction des contraintes typographiques en proposant des interfaces audio <sup>1</sup>. Elles constituent une bonne compensation mais requièrent un matériel pas toujours disponible et accentuent le découragement des utilisateurs face à l'utilisation de l'écriture, ce qui ne favorise pas le travail de remédiation par ailleurs effectué.

Des modèles pour la correction orthographique dédiée ont également été proposés. Ainsi (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 & 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. On se rend bien compte avec un corpus tel que celui que nous avons recueilli que ces modèles génériques peuvent rapidement prendre des proportions considérables. (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 & 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 & 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.

Cependant toutes ces applications 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). Cependant comme le montre l'analyse de notre corpus, 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 en se fondant sur des 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. A partir de ce constat, nous avons émis l'idée d'un système

<sup>1</sup>voir par exemple <http://www.01net.com/article/264021.html>

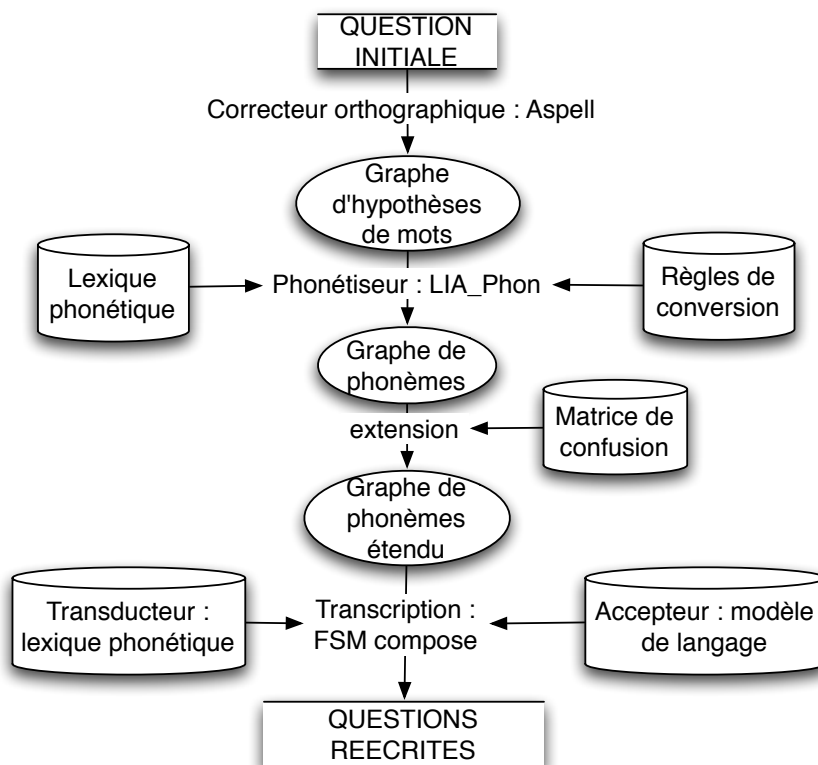


FIG. 1 – Etapes de la réécriture d’une question : état initial et final, représentations intermédiaires, outils pour les transitions entre les représentations, données utilisées par ces outils

automatique fonctionnant sur ce principe, en enchaînant une phonétisation de la phrase et une transcription du signal ainsi généré. En pratique le passage par un signal audio n’est pas nécessaire, on peut se contenter d’un passage par une séquence de phonèmes en sortie du phonétiseur et en entrée du système de reconnaissance. Le phonétiseur s’appuie sur un ensemble de règles de conversion graphème-phonèmes ainsi que sur un lexique phonétisé de la langue. Le système de reconnaissance s’appuie sur un modèle de langage ainsi qu’un lexique phonétique. pour produire plusieurs hypothèses de transcription, dont le contenu verbal se limite au lexique utilisé.

En pratique, nous nous sommes rendu compte qu’en se contentant d’une unique séquence de phonèmes correspondant à la phrase, nous perdions trop d’informations. Ainsi 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 graphes 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. Le graphe de la figure 1 illustre ce processus.

Dans les graphes représentant la phrase aux étapes intermédiaires, les arcs portent les coûts de transition associés aux phonèmes ou aux mots qu’ils portent également, et les noeuds sont les étapes qui séquent la phrase. Ainsi les différentes phrases 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.

On attribue un poids  $Wg$  aux mots alternatifs  $H$  (hypothèses graphémiques) proposés par le correcteur orthographique :

$$Wg(H) = f(d(H, I)), \quad (1)$$

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).

On attribue un poids  $Wp$  aux alternatives phonétiques  $H$  (hypothèses phonétiques) obtenues à l'aide d'une matrice de confusion :

$$Wp(H) = g(m(H, I)), \quad (2)$$

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 application

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 langages dans le formalisme des automates telle que proposée par (Allauzen & 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 appris sur un mois d'articles du journal *Le Monde* disponibles dans le corpus de la campagne EQUER.

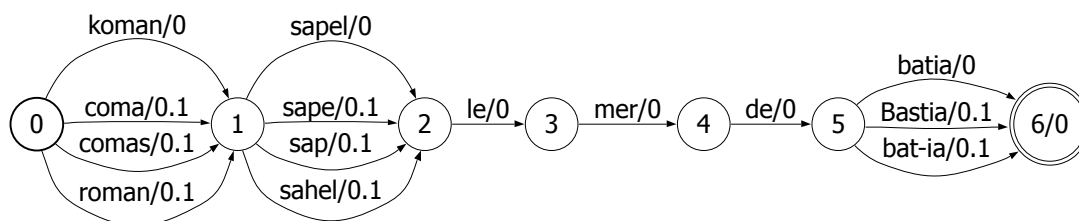


FIG. 2 – Graphe de mots pour la phrase *koman sapel le mer de batia*

Les hypothèses graphémiques permettant de générer le graphe de mots sont obtenues à l'aide du correcteur orthographique libre GNU ASPELL<sup>2</sup> qui, en mode *badspellers*, 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

<sup>2</sup><http://aspell.sourceforge.net>

rapport aux autres outils commerciaux et libres grand public <sup>3</sup>. La figure 2 montre un exemple de graphe 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. Par la suite elle pourra être étendue à l'aide de modèles de confusion phonétiques appris sur un grand corpus. La figure 3 illustre le graphe de phonèmes étendus correspondant au graphe de mots de la figure 2.

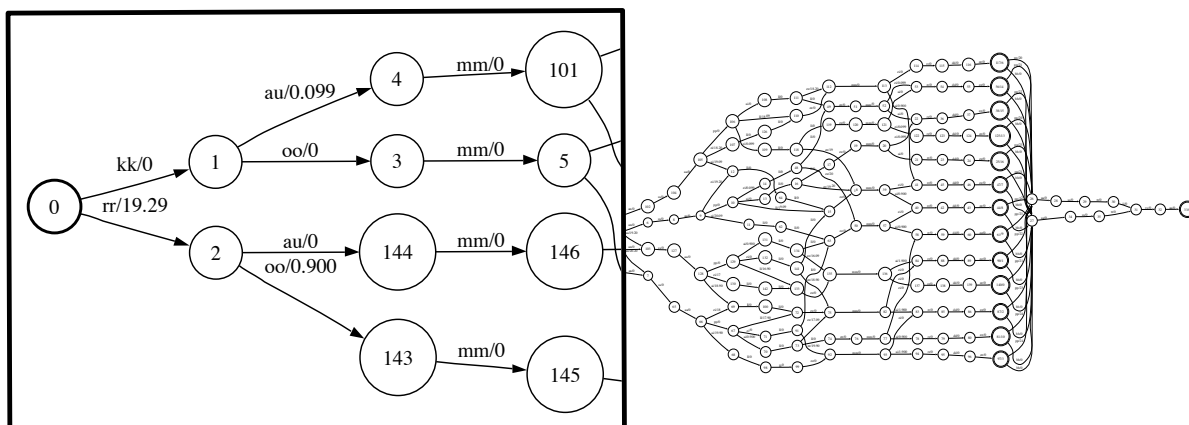


FIG. 3 – Graphe de phonèmes étendu de la phrase *koman sapel le mer de batia*

Phrase	Score
comment s'appelle le maire de bail à	52,2848701
comment s'appelle le maire de bahia	54,6559029
comment s'appelle le maire de bastia	54,8422737

TAB. 2 – Réécritures les plus probables de *koman sapel le mer de batia*

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

$$f(d(H, I)) = \begin{cases} 0 & \text{if } H = I \\ 0.1 & \text{if } H \neq I \end{cases} \quad (3)$$

$$g(m(H, I)) = \begin{cases} 0 & \text{if } H = I \\ 0.1 & \text{if } H \neq I \end{cases} \quad (4)$$

Ainsi la recherche du meilleur chemin dans le graphe final composé du graphe d'hypothèses phonétiques, de l'accepteur du modèle de langage et du transducteur du lexique phonétique prend en compte les coûts de transition issus de chacun des automates, et permet d'affecter des scores à tous les chemins. Le tableau 2 représente les trois hypothèses les plus probables issues du graphe de la figure 3 et les coûts associés à chacune de ces hypothèses. Dans ce cas, l'hypothèse attendue est la troisième et son score est proche des deux premières. Les phrases improbables se voient affecter des scores au delà de 100.

<sup>3</sup><http://aspell.net/test/>

## 4 Evaluation

Pour l'évaluation, 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 SCKT<sup>4</sup> 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.

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 forme de base. Par exemple la phrase *Comment s'appellent les maires de Bastia* sera traitée à l'identique de *Comment s'appelle le maire de Bastia* via la phrase lemmatisée *Comment se appeller maire Bastia*. Ainsi nous proposons pour l'évaluation de comparer les versions lemmatisées des phrases de références 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 indépendante, 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 tel que celui de la figure 2. Dans ce cas, l'évaluation des trois premières hypothèses correspond à l'évaluation du chemin le plus proche de la référence.

Le tableau 3 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 d'erreurs mots prend en compte à la fois les insertions, substitutions et omissions de mots, il faut le minimiser ; le pourcentage de phrases correctes constitue un aperçu des cas où l'on est certain que le système aura la possibilité de répondre, étant donné qu'il contient les mêmes informations que la phrase de référence (il contient également des informations bruitées dans le cas où l'on considère plusieurs hypothèses de réécriture), il faut le maximiser. Les résultats pour les trois premières hypothèses montrent que si l'amélioration en terme de taux d'erreurs est déjà importante dans l'absolu (on le divise par 2,5 par rapport à l'initial), elle l'est aussi par rapport à un correcteur orthographique performant (le taux d'erreur est 1,5 plus bas). Les résultats en terme de taux d'erreurs 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'apporterait 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

<sup>4</sup><http://www.nist.gov/speech/tools>

## Réécriture de phrases dysorthographiées

Mesure	Initial	Asp 1	Asp 3	FSM 1	FSM 3
Taux d'erreur	<b>51</b>	35,7	30,8	23,0	<b>19,9</b>
% phrases correctes	5,4	13,5	18,9	<b>43,2</b>	45,9

TAB. 3 – Taux d'erreur 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), si l'on considère la première ou les trois premières hypothèses.

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.

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, et il est intéressant d'observer leur répartition en fonction des individus et des thèmes abordés (les thèmes étant les questions d'origine). Cette répartition consignée dans le tableau 4 montre que si les variations existent entre les individus, à part pour 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 montre en revanche 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 on peut imaginer y ajouter les phrases contenant des mots inconnus du lexique et du modèle initial.

Pers	Taux d'erreur initial	Taux d'erreur	% phrases correctes
1	36	9	75
2	67	39	40
3	49	18	40
4	50	30	20
5	73	12	60
6	30	27	40
7	46	21	40
8	60	30	50

Thème	Taux d'erreur initial	Taux d'erreur	% phrases correctes
1	58	47	0
2	52	42	0
3	40	0	100
4	43	17	62
5	65	11	57

TAB. 4 – Distribution des performances de notre système sur la première hypothèses proposée, par personne (P) ou par thème (T) de question, selon les mesures de taux d'erreur mot (WER) et de pourcentage de phrases identiques à la correction, par rapport au taux d'erreur mot initial (IWER)

## 5 Conclusion

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 correctes une fois lemmatisées, et donc d'autant les performances d'un système de questions réponses pour des questions tapées par des enfants dyslexiques. 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 analyse en profondeur des erreurs résiduelles montre qu'il est encore possible d'améliorer nettement les performances à l'aide de modèles de langages et de lexiques plus adaptés, soit plus complets soit dynamiques.

## Références

- ALLAUZEN C. & MOHRI M. (2005). The design principles and algorithms of a weighted grammar library. *International Journal of Foundations of Computer Science*, **16**(3), 403–421.
- AYACHE C., GRAU B. & VILNAT A. (2006). Equer : the french evaluation campaign of question answering system equer/evalda. In *5th international Conference on Language Resources and Evaluation (LREC 2006)*, p. 1157–1160, Genoa, Italy.
- BECHET F. (2001). Lia\_phon - un système complet de phonétisation de textes. *Traitement Automatique des Langues (T.A.L.)*, **42**(1).
- BRILL E. & MOORE R. C. (2000). An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting of the ACL*, p. 286–293.
- DEOROWICZ S. & CIURA M. G. (2005). Correcting spelling errors by modelling their causes. *International journal of applied mathematics and computer science*, **15**(2), 275–285.
- GILLARD L., BELLOT P. & EL-BÈZE M. (2005). Le lia à equer (campagne technolanguage des systèmes questions-réponses). In *Actes de TALN'05*, Dourdan.
- GILLARD L., SITBON L., BELLOT P. & EL-BEZE M. (2006). Dernières évolutions de squallia, le système de questions/réponses du lia. *Traitement Automatique des Langues (TAL)*.
- GILLON G. T. (2004). *Phonological Awareness- From Research to Practice*. Guilford Press.
- JAMES A. & DRAFFAN E. (2004). The accuracy of electronic spell checkers for dyslexic learners. *PATOSS bulletin*.
- LÉTÉ B., SPRENGER-CHAROLLES L. & COLÉ P. (2004). Manulex : A grade-level lexical database from french elementary-school readers. *Behavior Research Methods, Instruments, and Computers*, **36**, 156–166.
- LOOSEMORE R. P. W. (1991). A neural net model of normal and dyslexic spelling. In *International Joint Conference on Neural Networks*, volume 2, p. 231–236, Seattle, USA.
- MOHRI M., PEREIRA F. C. N. & RILEY M. (2002). Weighted finite-state transducers in speech recognition. *Computer Speech and Language*, **16**(1), 69–88.
- MOHRI M., PEREIRA F. C. N. & RILEY M. D. (1997). At&t fsm librarytm – finite-state machine library.
- PEDLER J. (2001). The detection and correction of real-word spelling errors in dyslexic text. In *Proceedings of the 4th Annual CLUK Colloquium*.
- SPOONER R. (1998). *A spelling checker for dyslexic users : user modelling for error recovery*. PhD thesis, Human Computer Interaction Group, Department of Computer Science, University of York, Heslington, York,.
- TOUTANOVA K. & MOORE R. C. (2002). Pronunciation modeling for improved spelling correction. In *Proceedings of the 40th annual meeting of ACL*, p. 144–151, Philadelphia.