COORTE : Une trousse d'outils pour mettre en pratique une proposition de réforme de l'orthographe française

Valentin D. Richard, Emmanuel Fruchard, Valentin Gatien-Baron

COORTE : Une trousse d'outils pour mettre en pratique une proposition de réforme de l'orthographe française

Richard, Valentin D.[1,2], Fruchard, Emmanuel[2] & Gatien-Baron, Valentin[2]

[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
[2]Association Érofa
valentin.richard@loria.fr, contact@erofa.org, valentin.gatienbaron@gmail.com

1 Introduction

1.1 Contexte

Les évaluations du système éducatif et les études scientifiques pointent du doigt les nombreuses lacunes du système orthographique français et de son enseignement. Pourtant, l'orthographe de la langue française et la potentialité d'une réforme de celle-ci sont des sujets clivants. Avant 1990, elle n'avait pas connu de réforme depuis celles de 1835 et 1878. La réforme de 1990 introduit des rectifications mineures et éparses. Malgré cela, elle peine à être acceptée et utilisée (Humphries, 2019), notamment en France.

1.1.1 Complexité du système orthographique français

Le système orthographique français comporte trois difficultés majeures. Premièrement, il contient un grand nombre de correspondances graphème-phonème, ex. <a>-/a/, <eau>-/o/, <x>-/ks/, etc. On en compte 261 (dans le sens de la lecture, incluant les irrégularités et les lettres muettes finales), selon Manulex-Infra (Peereman et al., 2007), contre seulement 37 phonèmes (pour le français standard). Deuxièmement, la vaste majorité de ces correspondances est rare. Seules 44 correspondances ont une fréquence textuelle suffisante[1] (Peereman & Sprenger-Charolles, 2018). Enfin, en contexte, l'orthographe grammaticale introduit beaucoup de terminaisons qui n'ont pas d'équivalent à l'oral (Jaffré, 2005). Dans cet article, nous nous concentrons sur l'orthographe lexicale. Nous ignorerons donc les questions liées aux terminaisons grammaticales.

La maitrise des correspondances graphème-phonème est prépondérante dans l'apprentissage de la lecture (Gentaz et al., 2015) (*inter alia*). La consistance graphème-phonème d'un mot influence aussi la maitrise de son écriture chez les plus jeunes (Bosse et al., 2021). Le système orthographique français est notamment bien plus complexe que d'autres systèmes alphabétiques européens (ex. espagnol, italien, mais moins complexe que l'anglais) (Ziegler, 2018). Par rapport à d'autres pays, les enfants apprenant l'orthographe française accusent donc un retard que le système éducatif doit compenser sur plusieurs années.

1.1.2 Perception d'une réforme orthographique

L'enseignement de l'orthographe française prend un temps et une énergie certaines, pour des résultats parfois insatisfaisants. Les études DEPP (Eteve et al., 2022) attestent du faible taux de maitrise de l'orthographe en dictée au primaire, et de sa dégradation au fil des générations en France.

Une réforme de l'orthographe aiderait à réduire ces difficultés, notamment en réduisant le nombre de correspondances graphème-phonème ou en diminuant le nombre d'irrégularités lexicales. Pourtant, cette solution est loin de faire l'unanimité. En France, les sondages Ipsos de 1990 et Ifop de 2009, cités par

(Dister & Groupe RO, 2012b), recensent 40 % à 43 % de réponses favorables à une réforme, contre 56 % à 59 % défavorables, dont 26 % à 29 % très opposées.

Malgré tout, l'acceptabilité sociale d'une réforme est plus élevée parmi certaines catégories de personnes. L'enquête du groupe RO sur 1738 enseignant·es de 6 pays francophones en 2012 montre une majorité de réponses favorables à un aménagement de la norme graphique (Dister & Groupe RO, 2012b). Le nombre de personnes « pour » s'élève à plus de 70 % dans tous les pays (sauf en France : 64,5 %), dans l'hypothèse où ces aménagements concerneraient (seulement) les points jugés difficiles. Ces chiffres justifient la pertinence de l'étude approfondie de réformes candidates.

1.1.3 Propositions d'aménagements orthographiques

De nombreux appels à une réforme ont été formulés depuis 1878, par exemple en 1901 par Georges Leygues (Leygues, 1901). Cependant, comme la réforme de 1990 (qui en reprend certains points), elle ne propose que des ajustements spécifiques, et n'applique pas de modification systématique du jeu de graphèmes.

Des réformes radicales ont été proposées, comme Alfonic (Martinet, 1976) ou celle de Mario Périard (Jeandillou, 2009; Kolèktif Ortograf, 2014). Mentionnons aussi l'initiative Ortograf Altêrnativ (Jolicoeur, 2017). Contrairement aux autres projets, ce dernier a pour objectif de donner l'accès à la littératie aux personnes en situation d'incapacité intellectuelle moyenne à sévère, et non pas de changer la norme. Ces trois suggestions changent grandement le système graphique français. En dépit de la simplification majeure qu'une de ces propositions présenterait, l'opinion publique est fermement opposée à une réforme « phonétique » ou radicale (Dister & Groupe RO, 2012b, 2012a). Les avis penchent plutôt vers la rectification d'irrégularités ou la simplification de points difficiles.

C'est dans cette optique que l'association Érofa a proposé, à partir de 2009, plusieurs pistes de régularisation du système orthographique. Elles réduisent le nombre de graphèmes, et uniquement lorsqu'ils n'ont aucune fonctionnalité (grammaticale ou dans la famille de mots). Ces propositions sont détaillées en section 2.

1.2 Motivations

Nous avons choisi de nous concentrer sur les propositions d'Érofa, car elles nous semblaient les plus abouties et les plus à même d'être acceptées par la société. Notre objectif est de fournir des outils informatiques permettant d'utiliser l'orthographe suggérée par Érofa. Notre projet Coorte[2] (correcteur-convertisseur en orthographe d'Érofa) vise à rendre accessible la lecture et l'édition de textes dans cette orthographe. Notre initiative recouvre trois visées.

1. Visée évaluative
2. Visée scientifique
3. Visée pédagogique

1.2.1 Visée évaluative

L'objectif de nos programmes est de pouvoir être utilisés par un large public. Ils doivent donc prendre en compte une grande variété d'entrées textuelles. De ce fait, la description des changements qu'il contient doit être robuste et couvrante. La mise au point de nos outils vient donc tester si les propositions d'Érofa sont cohérentes entre elles et peuvent se généraliser. De plus, le dictionnaire de conversion produit par

l'association (le DOR, voir section 3.1) peut comporter des erreurs (par ex. fautes de frappe ou erreurs d'application des règles), qu'une méthode numérique peut venir mettre en lumière et corriger.

1.2.2 Visée scientifique

Les études sur les réactions à des réformes hypothétiques se basent souvent sur la description de celles-ci. Cela est dû au manque de longs textes transcrits, faute de moyens. La création d'un convertisseur automatique permet donc de produire des données qui peuvent être utiles à la réalisation d'études ciblées. Cette visée a été soulignée lors de la conception du logiciel de conversion Recto/Verso pour la réforme de 1990 (Beaufort et al., 2009). Ceci ouvre la possibilité de faire émerger et de tester des hypothèses scientifiques à propos de l'orthographe d'Érofa. Par exemple, il devient possible d'étudier les effets d'habituation à cette orthographe sur le long terme et sur un grand nombre de personnes.

1.2.3 Visée pédagogique

Dans le but de montrer les propositions d'Érofa, ou si elles sont adoptées, de les enseigner, un convertisseur vient en renfort mettre en application l'ensemble des règles. Il devient possible de visualiser et tester ces règles sur des entrées arbitraires. Cela offre des cas concrets, qui enrichissent l'illustration. De plus, si l'outil est employé au quotidien, il peut soutenir l'apprentissage de la lecture et de l'écriture dans cette graphie. Cette approche numérique est d'autant plus pertinente que les jeunes générations utilisent majoritairement des outils numériques pour vérifier leurs productions écrites (Le Levier, 2018).

1.3 Plan de l'article

Nous commençons par présenter un bref état de l'art d'outils destinés à un changement de norme graphique. Dans la partie 2, nous présentons la proposition de réforme d'Érofa et le lexique que cette association a créé : le DOR. La partie 3 aborde la numérisation du DOR et en mentionne quelques statistiques. Dans la partie 4, nous présentons les trois outils du pu projet Coorte, et nous les évaluons en partie 5. Nous concluons en partie 6.

1.4 État de l'art

D'autres outils d'aide à l'écriture et la conversion existent pour d'autres normes graphiques du français. Nous avons déjà mentionné le site Recto/Verso, pour la réforme de 1990 (Beaufort et al., 2009). Il met à disposition un convertisseur de texte brut en ligne ainsi qu'un service web de conversion de pages HTML. Pour des raisons pédagogiques, le convertisseur rajoute un marquage typographique et des infobulles d'information sur l'application des règles. En plus d'un lexique de conversion, un module de désambigüisation et un module syntaxique basique (pour le participe passé après « laisser ») sont utilisés.

Pour leur suggestion de réforme « phonétique » respective, les sites Fonétik[3] et Sinplegraf[4] proposent un convertisseur en ligne. Leur fonctionnement n'est pas décrit, mais ils se basent probablement sur un lexique comportant une transcription phonologique.

Des outils de conversion et de correction orthographique dans une nouvelle norme ont aussi été créés pour d'autres langues. En vue de l'accord orthographique de la langue portugaise de 1990, plusieurs universités ont proposé des outils numériques (Almeida et al., 2010; Ferreira et al., 2012). Le programme Lince permet de convertir du texte et des fichiers en format textuel (ex. HTML, DOCX, PDF). Il utilise un lexique de conversion et quelques règles morphologiques. Dans le cadre de la réforme de

l'orthographe allemande de 1996, un lexique de conversion a été conçu et inclus dans un correcteur utilisant l'outil LanguageTool[5].

L'objectif de Coorte est de reproduire ce genre d'outils. Pour l'instant, nous nous sommes concentrés sur un correcteur, un convertisseur en ligne de texte brut et autres formats textuels (PDF, Word, epub, etc.) et une extension de navigateur de conversion de pages web. À terme, nous envisageons aussi l'ajout de marquage typographique sur le texte converti, à des fins pédagogiques.

2 Proposition de réforme d'Érofa

L'association Érofa (Études pour une Rationalisation de l'Orthographe Française d'Aujourd'hui) propose la simplification de l'accord des participes passés (qui n'est pas l'objet de cet article), et trois règles pour réformer l'orthographe lexicale, auxquelles nous nous intéressons :

- La suppression des consonnes doubles (Gruaz, 2013)
- La simplification des lettres grecques et similaires (Gruaz, 2015)
- Le remplacement des <x> finaux muets par des <s> (Gruaz, 2009)

Une description synthétique de chaque proposition suit. Elles ne concernent pas les noms propres.

2.1 Suppression des consonnes doubles

Les consonnes doubles sont simplifiées par suppression d'une des deux consonnes, lorsque cela préserve la prononciation[6]. Il s'agit par exemple d'écrire le verbe « donner » avec un seul <n> : « doner ». Les doubles consonnes sont donc conservées dans les cas suivants :

- <ill> pour la transcription du yod, comme dans « feuille »
- <ss> pour distinguer les phonèmes /s/ et /z/ entre deux voyelles, par ex. « aussi »
- <n> et <m> pour la nasalisation de la voyelle qui précède : par ex. « immangeable»
- d'autres cas de prononciation de deux consonnes, comme dans « accident » ou « surréalisme »

Lorsque la double consonne est précédée d'un <e>, on l'accentue selon la *loi de position* (Detey et al., 2010), c.à.d. en fonction de l'ouverture de la syllabe[7] : par ex. « elle » → « èle », « cellule » → « célule ».

2.2 Remplacement des <x> muets finaux par des <s>

Le <x> en fin de mot est remplacé par un <s>, au singulier comme au pluriel, sauf s'il est prononcé, comme dans « index ». Cette règle s'applique à des noms et adjectifs invariables, par ex. « choix » → « chois », « deux » → « deus », et à des pluriels, par ex. « chevaux » → « chevaus », « ceux » → « ceus ».

2.3 Remplacement des graphèmes complexes grecs ou similaires

À l'instar des autres langues romanes, il s'agit de remplacer les graphèmes utilisés pour transcrire des lettres grecques ou des graphèmes complexes originaires d'autres langues (latin, etc.). Les cas les plus courants sont les suivants :

- Remplacement de <ph> par <f> pour le phonème /f/, comme dans « fotografie »
- Remplacement de <th> par <t> pour le phonème /t/, comme dans « téâtre »
- Remplacement de <y> par <i> pour le phonème /i/, comme dans « tipe »
- Suppression des <h> initiaux non aspirés, comme dans « istoire »

Il existe d'autres cas moins fréquents : <ch> (prononcé /k/) par <c> ou <qu> selon la voyelle qui suit, <rh> par <r>, <œ> par <é> ou <eu> selon sa prononciation, etc. À l'inverse, le <y> n'est pas remplacé lorsqu'il ne fait pas le son /i/ seul[8] : « payer », « pays » et « mayonnaise » maintiennent leur <y>.

2.4 Dictionnaire de l'Orthographe Rationalisée du français (DOR)

Le Dictionnaire de l'orthographe rationalisée du français (DOR)[9] a été constitué par Érofa en appliquant les règles précédentes sur le corpus du Robert électronique 2012 et du Petit Robert 2013, soit environ 60 000 entrées. Il fait correspondre chaque entrée dans son orthographe initiale (avec les rectifications de 1990) avec sa graphie dite rationalisée. Les entrées du Petit Robert qui ne sont pas modifiées ne sont pas présentes.

Pour illustrer les changements proposés, voici la conversion de la dictée du ROC (Lecourvoisier et al., 2006). Nous avons mis en gras les mots modifiés.

> Je vais vous raconter l'**istoire** d'un gentil petit garçon qui s'**apèle** Jo. Il **abite** chez son oncle, un **vieus** monsieur qui vit dans un bourg. Cet enfant possède un don extraordinaire. En **éfet**, grâce à ses **ieus** verts, il voit beaucoup plus loin et précisément que tout le monde ! Dans ses pupilles se trouvent des **jumèles** intégrées, microscopiques et invisibles…

3 Bases de données lexicales

3.1 Numérisation du DOR

Le DOR a été constitué par une revue de toutes les entrées du Petit Robert par des linguistes d'Érofa. Cette méthode présente l'avantage d'une vérification exhaustive de l'application des règles par leurs concepteur·rices. Elle a permis de préciser le contour des exceptions, telles que mentionnées ci-dessus. La méthode présente par contre l'inconvénient de produire certaines erreurs. Elles sont de deux natures : des fautes de frappe et des oublis d'application des règles. Cette seconde source d'erreur est d'autant plus probable que la norme graphique actuelle est ancrée profondément dans les habitudes.

Nous avons voulu confronter cette méthode à une méthode automatique. Le rapprochement de ces deux approches de génération d'un lexique de transcription est d'autant plus pertinent qu'elles sont indépendantes et accusent de faiblesses différentes.

3.2 Méthode automatique de conversion d'un lexique

Nous avons élaboré un algorithme d'apprentissage machine d'alignements graphème-phonème. Il est supervisé par la liste des alignements les plus fréquents, et tient compte de la fréquence des mots en usage. Les alignements prédits s'apparentent à ceux de Manulex-Infra (dans le sens de l'écriture) (Peereman et al., 2007). Il a été appliqué sur un lexique de lemmes et flexions contenant des transcriptions phonologiques, et qui est plus couvrant que Manulex-Infra : Lexique 3.83 (New et al., 2004) (142 695 entrées). Comme Lexique n'encode pas le <h> aspiré, nous avons aussi eu recours au Wiktionnaire et à Google Ngram Viewer (cherchant la fréquence des élisions devant le <h> initial) pour récupérer cette information.

Nous avons ensuite codé les règles d'Érofa et les avons appliquées sur ces alignements. La plupart des règles consistent à substituer un graphème par un autre lorsqu'il est aligné avec un certain phonème. Par

exemple, <ph> aligné avec /f/ est réécrit <f>. Nous utilisons aussi des règles sur la composition phonologique du mot pour le segmenter en syllabes. Ceci nous permet d'implémenter la loi de position, pour accentuer le <e> précédent une ancienne consonne double. Les exceptions, notamment concernant les noms communs dérivés de noms propres, sont gérées grâce à une liste construite manuellement.

Comme mentionné en section 1.2.1, la méthode automatique nous a permis d'identifier les erreurs d'application et oublis dans le DOR. Cela a aussi permis d'étendre le DOR de deux manières. D'une part, d'inclure d'autres mots réformés : des mots plus rares et des flexions absentes (ex. conjugaison des verbes). D'autre part, d'inclure dans le lexique cible les mots non-réformés. Cela est nécessaire pour la conception des outils. Dans la suite, nous appellerons ce nouveau lexique le DOR étendu.

3.3 Statistiques sur le DOR

3.3.1 Fréquence d'application des règles

La version numérique du DOR permet de compiler automatiquement les modifications proposées sous forme de statistiques. La Table 1 renseigne le pourcentage de mots présents dans le DOR en fonction de la (ou des) rectification(s) qu'ils subissent. Comme un mot peut avoir subi plusieurs règles, la somme des pourcentages excède 100 %.

Table 1: Pourcentage de mots du DOR concernés par les différentes règles.

Règle	Consonnes doubles	Finale en <x> muet	Graphèmes grecs et sim.
Mots affectés	52 %	10 %	42 %

On observe que la majorité des mots (52 %) sont affectés par la simplification des consonnes doubles.

Le DOR d'origine ne comportant pas les mots non-réformés, nous avons calculé le taux de mots à modifier en nous basant sur Lexique. Pour avoir une idée de la portée de ces changements en contexte d'écriture, nous reportons aussi les résultats pondérés par la fréquence des mots (sur l'écrit) donnée par Lexique. Ces chiffres sont disponibles en Table 2. Comme certains mots sont affectés par plusieurs règles, la prise en compte de toutes les règles à la fois produit un taux plus faible que leur somme.

Table 2: Pourcentages de mots affectés par les règles, dans Lexique et pondérés par les fréquences.

Règle	Consonnes doubles	Finales en <x> muet	Graphèmes grecs et sim.	Tout
Mots affectés dans Lexique	4,3 %	0,7 %	2,4 %	7,1 %
Mots affectés, pondérés par les fréquences	3,7 %	1,6 %	1,0 %	5,9 %

On observe dans les deux tableaux que la règle qui s'applique le plus est celle des consonnes doubles. En nombre absolu de mots affectés, elle est suivie par celle sur les graphèmes complexes, puis celle sur le <x> final. Cependant, ces deux s'inversent en contexte. En situation d'écriture, plus de mots contiennent un <x> final muet qu'un graphème complexe grec ou similaire. Cela est probablement dû au fait que les néologismes gréco-latins sont nombreux, mais peu fréquents (ce sont pour beaucoup des mots techniques), alors que les mots ne finissant pas <x> sont peu nombreux, mais courants. En somme, la réforme d'Érofa ne concerne en moyenne que 6 % des mots écrits, soit à peu près 18 mots par page de 300 mots.

3.3.2 Réduction du nombre d'erreurs

En tant que réforme, la proposition d'Érofa a notamment pour objectif de réduire le nombre d'erreurs d'orthographe à l'écriture. Pour évaluer ce gain, nous nous basons sur l'étude de la banque d'erreurs orthographiques ORTHOTEL (Auberge et al., 1999). Parmi les 168 erreurs de correspondances graphème-phonème listées, 33 seraient évitées. Pondéré par le nombre d'occurrence de ces erreurs, cela revient à éviter environ 40 %[10] des erreurs d'orthographe lexicale.

4 Correcteur et convertisseurs en orthographe d'Érofa

On aborde dans cette section les trois outils conçus pour appliquer les règles définies par Érofa.

4.1 Difficultés et objectifs de programmes de transcription

L'élaboration des outils de Coorte répond à six désidératas :

1. Performance sur la tâche en aval (ici, respect des règles d'Érofa)
2. Rapidité d'exécution
3. Portabilité sur plusieurs types de machines (notamment, légèreté du programme)
4. Facilité à inclure des mises à jour du DOR
5. Simplicité d'utilisation
6. Confiance (le programme ne doit pas communiquer des données privées à un serveur distant)

Pour satisfaire ces contraintes, nous avons choisi une approche de conversion mot à mot, sans analyse morphologique (systématique) ou syntaxique. Ainsi, n'importe quelle proposition de réforme basée sur un changement lexical pourrait être mise en œuvre avec une variante de ces outils[1112]. Ce choix implique que des mots homographes et non homophones, comme « *(ils) excellent* » et « *(c'est) excellent* », ne peuvent être distingués, en tout cas dans l'état actuel des outils.

Les cas d'usage d'un système d'écriture se répartissent en deux situations :

- l'écriture dans la nouvelle norme, aidée par un correcteur orthographique qui applique les règles nouvelles.
- La lecture, pour laquelle l'application d'une nouvelle norme requiert l'usage d'un convertisseur, de la norme graphique actuelle à la norme proposée. Nous avons réalisé trois outils de conversion : un convertisseur de texte brut, un convertisseur de document et un convertisseur de page web.

4.2 Correcteur orthographique

Un correcteur orthographique a pour fonction d'identifier les mots qui ne sont pas présents dans une norme et de proposer des mots voisins qui font partie du lexique selon cette norme. Nous avons élaboré un correcteur qui se base sur la norme proposée par Érofa. Par exemple, en graphie d'Érofa, le mot « graphie » est erroné. En , on voit comment il souligne en rouge ce mot et lui propose l'orthographe « grafie » après par un clic droit de la souris.

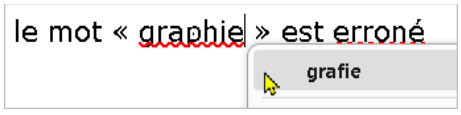

Figure 1: Capture d'écran du correcteur de Coorte dans Thunderbird.

4.2.1 Solution choisie

La solution technique choisie consiste à utiliser le moteur de correction de Hunspell, présent dans Firefox et dans Thunderbird (gestion de courriels). Cet outil utilise deux fichiers : l'un de lemmes et l'autre de règles de flexions. Plusieurs lexiques formatés pour Hunspell existent pour vérifier l'orthographe du français. Nous sommes partis de la version post-1990 de Grammalecte.

Nous avons adapté les deux fichiers de Grammalecte pour appliquer les propositions d'Érofa selon la méthode décrite en section 3.2. Par exemple, dans le fichier de lemmes, le verbe « donner » a été modifié en « doner ». Toutes ses formes fléchies restent déduites par Hunspell selon des règles de conjugaison inchangées. Au total, sur 83 000 lemmes, 16 000 environ sont modifiés, soit 20 %. Cette proportion est très supérieure à celle mentionnée ci-dessus car le lexique source comprend beaucoup de mots d'origine gréco-latine.

Dans le fichier de règles de flexions, nous avons par exemple modifié les règles de pluriel en <x> pour le remplacer par un <s>, ainsi que de nombreuses règles de suffixes de marques du féminin, pour supprimer une consonne double. 500 lignes de règles sur 6 000 ont été modifiées.

4.2.2 Performance

La performance de vérification est celle du moteur Hunspell, qui permet une frappe continue et une vérification au fil de l'eau sans aucune sensation de latence. Ni le nombre de lemmes, ni le nombre de règles n'ayant été augmentés, la rapidité d'exécution est donc satisfaisante. La qualité des suggestions de correction est évaluée en section 5.2.3.

4.2.3 Limitations actuelles et extensions

La principale limitation réside dans la portée de cet outil, qui est réservée aux logiciels extensibles sous licence libre. Il n'est (actuellement) pas disponible dans Word de Microsoft, par exemple, qui impose sa solution de vérification orthographique. Le système Hunspell est utilisable, en plus de Firefox et Thunderbird, sur Chrome et LibreOffice / OpenOffice, sur lesquels nous souhaitons également proposer cette extension.

4.3 Convertisseurs pour navigation sur le web

Un grand nombre de textes numériques consultés au quotidien se trouvent sur internet. Ces textes sont en orthographe traditionnelle, et il serait irréaliste d'exiger que leurs auteur·rices les changent en orthographe d'Érofa. Même si cette proposition était ratifiée par les autorités compétentes, leur transcription prendrait du temps, voire ne serait jamais effective pour les sites non entretenus.

La solution que nous apportons au besoin de conversion des textes numériques en ligne pour une utilisation quotidienne est une extension pour navigateur internet, qui transcrit les pages visitées à la volée. Nous avons construit une telle extension pour Firefox, Chrome et Safari satisfaisant les contraintes énoncées en section 4.1.

Pour éliminer tout effort de l'utilisateur·rice après l'installation, la transcription se fait automatiquement[13], et sur toutes les pages par défaut. Pour assurer la confiance en ce programme, la transcription est effectuée en local, sans communication réseau.

La réécriture se base presque entièrement sur l'utilisation du DOR étendu, ce qui assure une bonne efficacité. Ce processus peut convertir une page entière d'un seul coup sans ralentissement notable, même sur téléphone. Notamment, nous avons préféré éviter une conversion asynchrone (réécrivant le texte au fur et à mesure), qui pourrait gêner le confort visuel.

Voici un exemple de navigation sur Wikipédia avec notre extension :

Figure 2: Capture d'écran de la page Wikipédia « Photographie » avec l'extension web active

4.3.1 Fonctionnement

Pour une page web donnée, l'extension procède en plusieurs étapes. D'abord, elle détermine si la page contient suffisamment de français. Sinon, elle s'arrête pour ne pas interférer avec d'autres langues (le mot anglais « *comment* » par exemple ne doit pas être réécrit)[14]. Ensuite, elle transcrit le texte de la page destiné à la lecture, mais les zones de formulaire (qui peuvent être éditées par l'utilisateur) sont laissées inchangées. Le texte est grossièrement segmenté en mots, et chaque mot est si possible réécrit à l'aide du DOR étendu, avec en plus la gestion de l'absence de ligature <oe> ou <ae>. Malgré l'utilisation d'un lexique de réécriture, quelques règles morphologiques ou typographiques basiques sont implémentées : recherche d'un potentiel singulier si le mot se termine par un <s> et gestion des majuscules. Enfin, si la page change après le chargement initial, l'extension se relance.

4.3.2 Performances et limitations

Le chargement et l'exécution de l'extension est rapide. Par exemple, sur la page Wikipédia « Éléphant » (9 165 mots affichés), le tout s'effectue en 240 ms avec Firefox sur ordinateur. On obtient des performances similaires sur téléphone ou avec Google Chrome.

La principale limitation est celle de la portée de l'extension. Beaucoup d'applications autres que celles supportées permettent d'accéder à internet, notamment des applications pour smartphone.

4.4 Convertisseur en ligne

Nous avons aussi créé un site web qui réécrit un texte rentré par l'utilisateur·rice instantanément[15] :

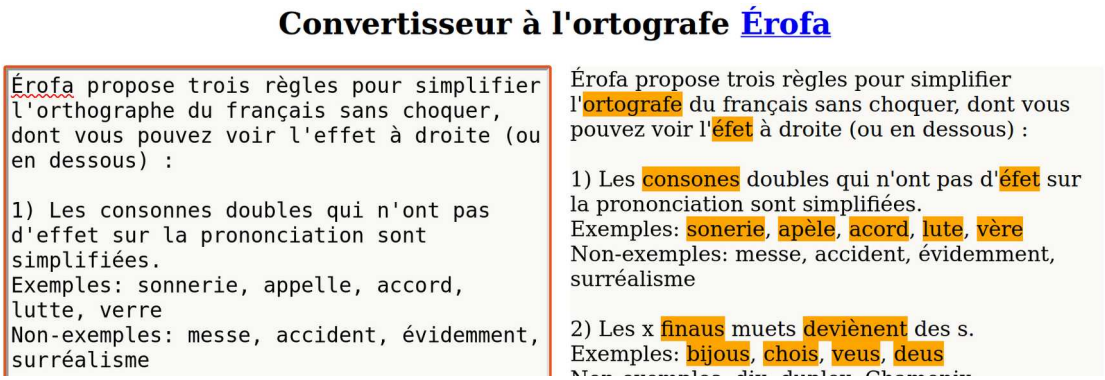

Figure 3: Capture d'écran du site web du convertisseur en ligne

Ce convertisseur en ligne utilise les mêmes techniques que l'extension web. Ce site permet aux utilisateur·rices d'expérimenter avec l'orthographe Érofa sans rien installer, ou de transcrire du texte plus facilement que l'extension. Le site propose aussi la conversion de documents, notamment fichiers Word, livres numériques (epub) et PDF.

5 Évaluation des outils

5.1 Méthode

Nous évaluons les performances du correcteur et du convertisseur sur des tâches de traduction automatique et de correction automatique d'erreurs. Pour ce faire, nous avons constitué un corpus composé de trois textes de genres différents.

Le premier texte, *roman* (472 mots), est le début du roman « Pour moi seule » d'Albin Michel (1919). Le deuxième texte, *journal* (362 mots), est un article de journal de l'Est Républicain du 7 janvier 2011 (Gaiffe & Nehbi, 2020). Le dernier texte, *wiki* (500 mots), comporte les premiers paragraphes de la page Wikipédia « Coévolution antagoniste ». L'objectif est de tester les logiciels sur du contenu pouvant comporter beaucoup de termes techniques. Ces textes ont été sélectionnés aléatoirement[16].

Nous avons fait annoter ce corpus par trois membres d'Érofa ayant travaillé sur ces propositions. Le texte de référence basé sur ces trois annotations (*gold*) est utilisé pour calculer les performances de nos programmes.

Notre évaluation se base sur deux tâches de prédiction. La première tâche est une classification binaire. Elle consiste à prédire si un mot doit être réécrit selon la proposition d'Érofa. Les deux classes sont : 0 - le mot ne doit pas être réécrit, et 1 - le mot doit être réécrit. Nous calculons l'exactitude et la F-mesure par rapport à la référence.

La deuxième tâche est celle de prédiction de la forme rectifiée, étant donné un mot à réécrire. Elle est évaluée sur l'ensemble des 3 annotations de référence et est mesurée par le score BLEU[17] (Papineni et al., 2002). Pour le correcteur, le premier candidat fourni est utilisé (ou un token spécial s'il n'y a pas de candidat). Nous ajoutons aussi une mesure des différents candidats du correcteur en mesurant le rang réciproque moyen (MRR) sur les mots à réécrire.

Pour pouvoir interpréter le score d'exactitude et le score BLEU, nous utilisons l'identité comme baseline. Nous comparons aussi nos résultats à ceux de ChatGPT 3.5. Le prompt donné pour transcrire le corpus

décrit les règles présentées ici en donnant quelques exemples. Cela nous permettra de commenter la pertinence de notre approche symbolique face à l'utilisation d'un gros modèle de langue générique.

5.2 Résultats

5.2.1 Tâche d'annotation experte

Le travail d'annotation experte consistait à identifier et réécrire les mots selon la proposition de réforme. Le score inter-annotateurs Fleiss kappa est de 0,756. Il est un peu faible au vu de l'apparente simplicité de la tâche.

Nous avons remarqué que cette tâche favorisait les oublis. Si le jugement majoritaire avait été utilisé pour former la valeur de référence (*gold*), 22 occurrences de celle-ci auraient été des oublis (ex. *elle*, *croix* et *histoire* n'ont pas été réécrit par deux annotateurs sur trois). Nous avons donc choisi les mots les plus éloignés de l'original comme gold. Autrement dit, si un mot est jugé « à réécrire » par un annotateur, il est pris comme gold, parfois malgré le jugement majoritaire.

Les seuls autres désaccords concernent l'accentuation des <e> (*dérière* vs. *dèrière*), notamment lorsque la syllabe suivante comporte un <e> optionnellement muet (*apèleront* vs. *apéleront*).

5.2.2 Tâche d'identification des mots à réécrire

Les résultats de la tâche d'identification des mots à réécrire sont donnés en Table 3 et Table 4.

Table 3: Scores des différents programmes à la tâche de classification binaire (total sur les trois textes).

	Exactitude[18]	Précision[19]	Rappel[20]	F-mesure[21]
Correcteur	0,976	0,838	**0,966**	**0,897**
Convertisseur	**0,978**	**0,914**	0,876	0,894
ChatGPT	0,856	0,344	0,366	0,355
Identité	0,891	/	0,000	/

En premier lieu, on observe que ChatGPT a une exactitude moins bonne que la baseline, et une très faible F-mesure. Il fait beaucoup de faux positifs (précision basse) en modifiant des mots non demandés (ex. *évolutive* → *évolutiv*). Il fait beaucoup de faux négatifs (rappel bas) en oubliant des mots.

Nos deux programmes ont de bons scores : l'exactitude et la F-mesure sont hauts. Parmi les faux positifs communs, on trouve des mots oubliés par tous les annotateurs (ex. *comme*, *aux*). Le nombre plus élevé de faux positifs du correcteur est dû aux noms propres inconnus de Grammalecte, par défaut identifiés comme incorrects. Le convertisseur compte aussi quelques noms propres homographes de noms communs à réécrire (ex. *Pierre*). Le problème des homographes est aussi présent dans Recto / Verso.

Parmi les faux négatifs communs, on trouve deux artéfacts dus à la méthode de calcul du gold : « *peut-être* » erronément réécrit « *peutêtre »*, et « *A Paris* » réécrit (de manière juste) « *À Paris* », même si la tâche ne demandait pas de corriger ce genre de « faute ». On trouve aussi quelques lacunes du lexique (ex. *s'accomplirent* ou *intercommunalité* absents). Le nombre plus élevé de faux négatifs du convertisseur est

dû à la non-incorporation des orthographes pré-1990 dans le DOR étendu (ex. *événement*) et un problème de segmentation en mots (*hôte/parasite* pas réécrit).

Table 4: F-mesure des différents programmes sur les différents textes.

	Roman	Journal	Wiki
Correcteur	**0,911**	0,795	**0,950**
Convertisseur	0,909	**0,958**	0,848
ChatGPT	0,463	0,455	0,074

La Table 4 montre que le correcteur et le convertisseur sont performants sur les trois genres : littéraire, journalistique et scientifique. Le plus faible score du correcteur sur l'article de journal est dû aux nombreux noms propres. Malgré cela, en pratique, un utilisateur humain devrait comprendre que ces mots ne sont pas à modifier.

5.2.3 Qualité de la réécriture et des candidats du correcteur

Le score BLEU des différents programmes est affiché en Table 5.

Table 5: Score BLEU des différents programmes sur les différents textes.

	Roman	Journal	Wiki	Tout
Correcteur	0,894	0,878	**0,921**	0,900
Convertisseur	**0,930**	**0,977**	0,920	**0,939**
ChatGPT	0,740	0,611	0,806	0,730
Identité	0,893	0,845	0,859	0,867

Comme attendu, les scores BLEU de ChatGPT sont très mauvais. En plus des raisons précédemment mentionnées, les mots correctement identifiés sont parfois mal réorthographiés. Notamment, il oublie d'accentuer les <e> précédant une consonne double. La médiocrité d'un gros modèle de langue pour ce genre de tâche nous conforte sur l'emploi d'une méthode symbolique.

Les scores BLEU de nos programmes sont bons, et suivent globalement les F-mesures. Le correcteur fonctionne légèrement moins bien car son système de recommandation repose sur celui de Hunspell. Il propose parfois un autre mot que le mot cible en premier candidat (ex. pour *professionnelle* : *professionnel* en premier candidat). Dans quelques cas, le mot cible est bien dans le lexique, mais n'est même pas proposé. Malgré tout, le mot cible est la plupart du temps premier candidat, notamment lorsque l'application des règles d'Érofa est légère, comme la suppression d'une seule lettre. Le score MRR sur tout le corpus est de 0,830.

6 Conclusion et perspectives

L'association Érofa a proposé trois règles de réformes de l'orthographe lexicale du français : simplification des consonnes doubles, remplacement des <x> muets finaux par un <s>, et suppression des graphèmes complexes grecs ou similaires. Nous avons présenté trois outils numériques mettant en application cette proposition de réforme. D'une part, la numérisation du dictionnaire produit par Érofa nous a permis d'y identifier des erreurs et d'en extraire des statistiques. Cette proposition touche environ 6% des mots en contexte. D'autre part, elle nous a servi de base pour l'élaboration de nos programmes informatiques : un correcteur et un convertisseur automatique (extension web et conversion de texte). Ils fonctionnent sur la base d'un lexique réécrivant mot à mot. Leurs performances sont très satisfaisantes quel que soit le genre du texte en entrée.

Nous travaillons à améliorer ces outils, notamment en essayant d'inclure le plus de mots possible, par exemple en utilisant le Wiktionnaire (Sajous et al., 2013). Des modules complémentaires (par exemple, reconnaissance d'entités nommées) pourraient permettre de limiter les faux positifs. Nous envisageons aussi d'étendre la correction à l'entrée clavier des smartphones.

Références bibliographiques

Almeida, J. J., Santos, A., & Simões, A. (2010, mai 19). *Bigorna : A toolkit for orthography migration challenges*. https://repositorium.sdum.uminho.pt/handle/1822/16529

Auberge, V., Ghneim, N., & Belrhali, R. (1999). Analyse du corpus ORTHOTEL : Apport du traitement automatique à la classification des déviations orthographiques. *Langue française*, *124*(1), 90-103. https://doi.org/10.3406/lfr.1999.6308

Beaufort, R., Dister, A., Naets, H., Macé, K., & Fairon, C. (2009). Recto /Verso. Un système de conversion automatique anciène / nouvèle ortografe à visée linguistique et didactique. In A. Nazarenko & T. Poibeau (Éds.), *Actes de la 16ème conférence sur le Traitement Automatique des Langues Naturèles. Articles courts* (p. 301-310). ATALA. https://aclanthology.org/2009.jeptalnrecital-court.33

Bosse, M.-L., Brissaud, C., & Le Levier, H. (2021). French Pupils' Lexical and Grammatical Spelling from Sixth to Ninth Grade : A Longitudinal Study. *Language and Speech*, *64*(1), 224-249. https://doi.org/10.1177/0023830920935558

Detey, S., Durand, J., Laks, B., & Lyche, C. (2010). *Les Variétes du Français Parle Dans l'Espace Francofone—Ressources pour l Enseignement*. OPHRYS.

Dister, A., & Groupe RO. (2012a). « Une bonne réforme est possible, à condition de… ». Les maitres s'expriment sur ce que serait une « bonne » réforme de l'orthographe française. *Glottopol*, *19*, 117-129.

Dister, A., & Groupe RO. (2012b). Une réforme de l'ortografe ? Quels positionements ? *Glottopol*, *19*, 37-51.

Ferreira, J. P., Lourinho, A., & Correia, M. (2012). Lince, an End User Tool for the Implementation of the Spelling Reform of Portuguese. In H. Caseli, A. Villavicencio, A. Teixeira, & F. Perdigão (Éds.), *Computational Processing of the Portuguese Language* (p. 46-55). Springer. https://doi.org/10.1007/978-3-642-28885-2_5

Gaiffe, B., & Nehbi, K. (2020). *Corpus journalistique issu de l'Est républicain, v1*. ORTOLANG. https://hdl.handle.net/11403/est_republicain/v4

Gentaz, E., Sprenger-Charolles, L., & Theurel, A. (2015). Differences in the Predictors of Reading Comprehension in First Graders from Low Socio-Economic Status Families with Either Good or Poor Decoding Skills. *PLOS ONE*, *10*(3), e0119581. https://doi.org/10.1371/journal.pone.0119581

Gruaz, C. (Éd.). (2009). *Le X final*. Éditions Lambert-Lucas. http://www.lambert-lucas.com/livre/x-final-le/

Gruaz, C. (Éd.). (2013). *Simplifier les consonnes doubles*. Éditions Lambert-Lucas. http://www.lambert-lucas.com/livre/simplifier-les-consonnes-doubles/

Gruaz, C. (Éd.). (2015). *Les Lettres grecques et similaires*. Éditions Lambert-Lucas. http://www.lambert-lucas.com/livre/les-lettres-grecques-et-similaires/

Humphries, E. (2019). #JeSuisCirconflexe : *The French spelling reform of 1990 and 2016 reactions*. *JOURNAL OF FRENCH LANGUAGE STUDIES*, *29*(3), 305-321. https://doi.org/10.1017/S0959269518000285

Jaffré, J.-P. (2005). L'orthographe du français, une exception ? *Le français aujourd'hui*, *148*(1), 23-31. https://doi.org/10.3917/lfa.148.0023

Jeandillou, J.-F. (2009). De ratione scribendi. *Linx. Revue des linguistes de l'université Paris X Nanterre*, *60*, 73-83. https://doi.org/10.4000/linx.699

Jolicoeur, M. (2017). *COMMUNICATION ÉCRITE 3 – ortograf altêrnativ*. Bibliothèque et Archives nationales du Québec. http://capable.ctreq.qc.ca/wp-content/uploads/2018/04/Communication-%C3%A9critre-3-ortograf-alt%C3%AArnativ-2.pdf

Kolèktif Ortograf. (2014). *Dikortograf 2014* (Lè-z Édision Sédisieuz). https://www.lulu.com/shop/kol%C3%A8ktif-ortograf/dikortograf-2014/paperback/product-21626183.html

Le Levier, H. (2018, mai 30). *Rapport à l'orthographe d'apprenants du secondaire et du supérieur dans les écrits numériques extrascolaires*. Colloque international des Étudiants chercheurs en Didactique des Langues 2018, CEDIL'18. https://hal.univ-grenoble-alpes.fr/hal-02099751

Lecourvoisier, F., Boudinot, A., Davy-Aubertin, C., & Willhelm, C. (2006). *ROC Repérage Orthographique Collectif* (p. 25) [Notice]. Cogni-Sciences – UPMF Grenoble, CHU Montpellier, Académie de Grenoble, Académie de Montpellier, Académie de Rennes. http://compiegne.dsden60.ac-amiens.fr/IMG/pdf/test_roc.pdf

Leygues, G. (1901). 73. 26 février 1901 : Arrêté relatif à la simplification de la syntaxe française. *Publications de l'Institut national de recherche pédagogique*, *5*(2), 198-203.

Martinet, A. (1976). L'accès à la lecture et à l'écriture par l'alfonic. *Communication & Langages*, *30*(1), 21-33. https://doi.org/10.3406/colan.1976.4295

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2 : A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 516-524. https://doi.org/10.3758/BF03195598

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu : A Method for Automatic Evaluation of Machine Translation. In P. Isabelle, E. Charniak, & D. Lin (Éds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (p. 311-318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Peereman, R., Lété, B., & Sprenger-Charolles, L. (2007). Manulex-infra : Distributional characteristics of grapheme—phoneme mappings, and infralexical and lexical units in child-directed written material. *Behavior Research Methods*, *39*(3), 579-589. https://doi.org/10.3758/BF03193029

Peereman, R., & Sprenger-Charolles, L. (2018). Manulex-MorphO, une base de données sur l'orthographe du français intégrant les morpho-phonogrammes. *Langue française*, *199*(3), 99-109. https://doi.org/10.3917/lf.199.0099

Sajous, F., Hathout, N., & Calderone, B. (2013). GLÀFF, a Large Versatile French Lexicon (GLÀFF, un Gros Lexique À tout Faire du Français) [in French]. *Proceedings of TALN 2013 (Volume 1: Long Papers)*, 285-298. https://aclanthology.org/F13-1021

Ziegler, J. C. (2018). Différences inter-linguistiques dans l'apprentissage de la lecture. *Langue française*, *199*(3), 35-49. https://doi.org/10.3917/lf.199.0035

[1]U > 5.000, modulo les proximités /e/-/ɛ/, /o/-/ɔ/, /a/-/ɑ/ et /œ/-/ø/

[2]Le code source est en libre accès : https://gitlab.com/erofa/coorte

[3]http://fonétik.fr/fon%C3%A9tikeur-fr.html

[4]https://sinplegraf.org/transcripteur.html

[5]https://www.korrekturen.de/rechtschreibpruefung.shtml

[6]Il a y deux exceptions dans la version actuelle des règles. Elles ne couvrent ni les adverbes en <-mment> ni le mot « femme ».

[7]Cette loi de position n'est qu'une préférence du système phonologique français, et n'est pas uniforme dans la francophonie. Notamment, on observe de nombreuses variations de ces transcriptions phonémiques selon les lexiques (voir le commentaire de (Sajous et al., 2013, p. 10)). Érofa a choisi cette règle pour sa simplicité, en dépit la variation.

[8]Le DOR comporte une exception : « yeux » → « ieus ». De plus, si un <y> prononcé /i/ est à la fin d'un mot, il est gardé <y>, ex. « rugby » et l'adverbe « y » ne sont pas modifiés.

[9]Ce dictionnaire est disponible sous format papier et électronique : www.lambert-lucas.com/wp-content/uploads/2022/11/OA-dictionnaire-EROFA.pdf

[10]Le nombre d'occurrences mentionné dans l'article est agrégé par catégorie d'erreur (une catégorie par phonème cible), ce qui nous empêche d'avoir le nombre précis d'erreurs évitées, les données brutes étant indisponibles. Nous avons donc fait l'hypothèse grossière que la répartition des erreurs de correspondance était uniforme dans chaque catégorie. Si on retire les catégories comportant des erreurs non couvertes, la pondération atteint quand même un minimum de 30 % d'erreurs évitées.

[11]Nous avons d'ailleurs composé un lexique de conversion orthographe traditionnelle / orthographe de 1990 grâce à Grammalecte et Recto/Verso. Le convertisseur en ligne et l'extension web peuvent donc être utilisés pour la norme post-1990.

[12]Nous avons observé que le site Sinplegraf utilisait déjà notre extension web avec son propre lexique.

[13]Cela s'oppose notamment aux extensions de traduction de pages, où l'utilisateur·rice doit cliquer sur un bouton à chaque page qu'il ou elle souhaite traduire.

[14]Comme l'extension réécrit toute la page, elle peut malencontreusement réécrire des mots ou paragraphes si plusieurs langues sont présentes.

[15]Le convertisseur en ligne est accessible à l'adresse https://orthographe-rationnelle.info/

[16]*Wiki* a été sélectionné en tirant au hasard la première page Wikipédia de sujet scientifique d'au moins 300 mots. *Roman* a été sélectionné en recherchant le roman en français le plus récemment ajouté au site du projet Gutenberg. *Journal* a été sélectionné en piochant un jour et un article au hasard du corpus Est-Républicain (Gaiffe & Nehbi, 2020).

[17]Le score BLEU sur une phrase est calculé en comptant le nombre de bouts contigus de taille k de la phrase transcrite qui sont présents dans au moins une des phrases de référence, pour k allant de 1 à 4. Le score est ensuite moyenné sur l'ensemble du texte.

[18] L'exactitude est la proportion de prédictions correctes parmi le nombre total de cas examinés. Elle peut être facilement haute s'il y a faible nombre d'occurrences du phénomène à identifier, comme c'est le cas ici.

[19] La précision est la proportion de vrais positifs parmi les cas positifs, c.à.d. parmi les mots où le programme a prédit la classe 1. La précision de l'identité n'est pas définie car la classe 0 est prédite pour tous les mots.

[20]Le rappel est le nombre de vrais positifs parmi l'ensemble des mots qui auraient dû être identifiés.

[21]La F-mesure est la moyenne harmonique de la précision et du rappel.