“Ortholexies”, une base de données publique pour l’orthographe lexicale

Jean-Luc Manguin

SHS Web of Conferences **27**, 11006 (2016) DOI: 10.1051/shsconf/20162711006
Congrès Mondial de Linguistique Française - CMLF 2016

«Ortholexies», une base de données publique pour l'orthographe lexicale

Jean-Luc Manguin

CNRS, GREYC UMR 6072
jean-luc.manguin@unicaen.fr

Résumé. Nous décrirons ici la construction de notre base « Ortholexies », en particulier nous examinerons les techniques utilisées pour apparier les formes erronées avec leurs formes correctes, et nous expliquerons comment nous avons comparé nos données avec d'autres sources. Enfin nous détaillerons certains travaux où cette base a pu servir de support, avant de conclure par ses perspectives de développement et ses autres possibilités d'application.

Abstract. Here we describe the construction of our database "Ortholexies", in particular we will examine the techniques used to pair the erroneous forms with their correct forms, and explain how we compared our data with other sources. Finally we will detail some works in which this database could be used as support, before concluding by its development prospects and its other possible applications.

1 Introduction

La correction des erreurs orthographiques est un des premiers problèmes auxquels s'est attaqué le traitement automatique des langues, mais il demeure un sujet de recherche de cette discipline (voir par exemple [Baranes, 2015]), car sa simplicité n'est qu 'apparente. En effet, l'erreur orthographique (lexicale) ne résulte pas d'un brouillage aléatoire (comme dans le cas d'une transmission numérique) mais d'un processus cognitif complexe où se mêlent des confusions graphiques et/ou phonétiques. C'est pourquoi les méthodes de correction de mots isolés fondées sur la correction de codes binaires s'avèrent imprécises si elles sont employées seules. Les meilleurs résultats sont souvent obtenus avec des méthodes hybrides qui emploient un dictionnaire d'erreurs. Dans le domaine public, une telle ressource apparaît nécessaire car il n'y a pas de base en libre accès qui puisse indiquer les formes erronées observées, ni donner une idée de leur fréquence ; seuls existent des correcteurs orthographiques en ligne, ou intégrés aux outils bureautiques. La liste des corrections offerte par Wikipedia est par ailleurs totalement indigente (contrairement à sa version anglaise).

Nous décrirons ici la construction de notre base « Ortholexies », en particulier nous examinerons les techniques d'appariement des formes erronées avec leurs formes correctes, et nous expliquerons comment nous avons comparé nos données avec d'autres sources. Enfin nous détaillerons certains travaux où cette base a pu servir de support, avant de conclure par ses perspectives de développement et ses autres possibilités d'application.

2 Constitution de la base

Les données qui ont permis de constituer la base sont les requêtes reçues par le dictionnaire des synonymes du Crisco [1998] pendant les neuf premières années de sa mise en ligne (oct. 1998 - déc. 2007) ; le total représente plus de 200 millions de mots, répartis en un peu plus de 4 millions de formes distinctes. Chaque forme peut en effet avoir été demandée plusieurs fois (de 1 à 290 000 fois), mais nous n'avons retenu pour notre base que les formes demandées au moins 5 fois, ce qui réduit la liste à environ 566 000 formes différentes. Notre fichier de départ contient donc les formes et leurs fréquences :

Fréquence	**Nombre de formes distinctes**
Supérieure à 50 000	264
de 5 000 à 50 000	7 476
de 500 à 5 000	29 116
de 50 à 500	87 445
de 5 à 50	441 810

Tableau 1 : répartition des formes distinctes par tranche fréquentielle

3 Sélection et appariement des erreurs

Notre but étant de réaliser une base de données qui indique quelles sont les formes erronées d'une forme correcte (et inversement), le travail sur les données se divise naturellement en deux étapes : le repérage des formes correctes, puis l'appariement des formes restantes avec les formes correctes, quand cela est possible.

3.1 Sélection des erreurs

Pour repérer les formes correctes parmi nos données, nous employons la base Morphalou [Romary et al. 2004] comme base de référence. A ce jour, en descendant jusqu'au seuil fréquentiel égal à 5, nous avons repéré 124 125 formes correctes. Notre base contient donc, jusqu'à ce seuil f=5, environ 22 % de formes correctes[1], dont la répartition par tranche fréquentielle est donnée ci-après :

Fréquence	Nombre de formes distinctes	Nombre de formes correctes	Pourcentage
Supérieure à 50 000	264	263	99,6 %
de 5 000 à 50 000	7 476	7084	94,8 %
de 500 à 5 000	29 116	23 069	79,2 %
de 50 à 500	87 445	39 039	44,6 %
de 20 à 50	90 478	18 013	19,9 %
de 5 à 20	351 332	36 657	10,4 %

Tableau 2 : répartition des formes correctes par tranche fréquentielle

3.2 Principes d'appariement avec les formes correctes

L'appariement entre une forme erronée et sa (ou ses) correction(s) est un problème étudié depuis une cinquantaine d'années. Sans faire un état de l'art complet de la question, nous pouvons d'abord rappeler qu'on sait depuis les années 1960 que dans environ 80% des erreurs orthographiques (lexicales), la forme erronée et sa forme correcte se situent à une distance de Levenshtein (ou de Damerau-Levenshtein) égale à 1, ce qui semble faciliter la correction [Damerau 1964]. Malheureusement, cette valeur égale à 1 est aussi bien souvent la distance entre les unités lexicales courtes : il suffit de taper « coute » dans un logiciel de traitement de texte standard et de regarder les corrections proposées pour le constater. La correction par « voisinage » n'est donc pas une méthode assez précise pour être automatique. Certains ont alors eu recours à la phonétisation, qui suppose que les erreurs ne sont que des mauvaises résolutions d'inconsistances. Mais là encore, de nombreuses erreurs ne peuvent être corrigées, puisqu'elles résultent souvent d'une méconnaissance des règles de prononciation. Par exemple, l'interface d'un dictionnaire de référence du français, qui fonctionne avec ce type de correction, s'avère incapable de corriger la forme « acceuil », qui est pourtant une erreur fréquente. Remarquons enfin que plusieurs études sur les performances orthographiques des français ont montré que les erreurs les plus courantes sont celles qui concernent les accents et les consonnes doubles [Lucci & Millet 1994].

A partir de ces constatations, nous avons opté pour une méthode hybride en plusieurs passes, qui corrige tout d'abord les erreurs sur les diacritiques, puis emploie le « voisinage » (dans un sens que nous

définirons ensuite) comme champ de recherche, et enfin utilise l'homophonie pour apparier les formes qui ne l'ont pas encore été.

Remarquons enfin que nous cherchons toujours à apparier des formes qui existent dans la base, autrement dit nous faisons l'hypothèse que si une forme est erronée, sa forme correcte doit se rencontrer dans notre base (nous reviendrons sur cette remarque dans la discussion à propos de la correction).

3.3 Correction des diacritiques

La correction des diacritiques est une opération techniquement assez simple pour être automatisée. Le principe est le suivant : pour apparier une forme erronée, on commence par la débarrasser de ses diacritiques, puis on recherche dans la base Morphalou une forme correcte (elle aussi débarrassée de ses diacritiques) qui lui soit égale. Cette méthode donne généralement une seule « solution » d'appariement, les seuls cas d'ambiguïté qui peuvent se produire sont entre participes passés et formes du présent de l'indicatif des verbes du premier groupe (ex. détermine/déterminé). La précision de cette méthode est excellente (supérieure à 99%).

3.4 Correction avec les voisins ou avec les formes réduites

En nous appuyant sur la constatation faite par Damerau, nous pouvons nous servir du voisinage graphique pour proposer automatiquement des formes correctes qui correspondraient à la (ou les) correction(s) possible(s) de chaque forme erronée. Malgré le temps que nécessite le calcul de la distance d'édition (ou de Levenshtein), et bien que ce calcul puisse être abrégé lorsqu'on ne recherche que les termes voisins (donc à une distance égale à 1), cette méthode permet de corriger avec une assez bonne précision un pourcentage satisfaisant des erreurs qui ne concernent pas les diacritiques. Néanmoins, sachant que le réseau lexical par voisinage se densifie énormément quand les formes deviennent plus courtes, la précision de ce type de correction baisse très sensiblement pour les formes de faible longueur, et le contrôle humain obligatoire devient vite long et fastidieux. Par ailleurs, il n'est pas rare de rencontrer des formes erronées combinant à la fois une erreur de type « voisinage » (substitution, omission, ou ajout) avec une erreur de signe diacritique. Une telle forme, pourtant facile à corriger, se place alors à une distance égale à 2, et devient incorrigible par voisinage. Il est en effet illusoire d'élargir la méthode par voisinage aux « voisins de voisins », en raison de l'explosion combinatoire des formes proposées comme candidates.

Ces inconvénients que nous venons de citer nous ont conduit à élaborer une méthode d'appariement du même type que celle utilisée pour les erreurs sur les diacritiques. Cette méthode n'emploie pas de calcul de distance, mais opère par comparaison de « formes réduites ». Ce que nous appelons « forme réduite », c'est une forme dans laquelle nous avons non seulement remplacé les diacritiques par les signes simples, mais encore éliminé certaines particularités qui sont les principales sources d'erreurs [Lucci & Millet 1994] : consonnes doubles, y (remplacé par i), et h après certaines consonnes (t, b, d). En voici deux exemples :

professionnel -> profesionel (forme réduite)

dithyrambique -> ditirambique (forme réduite)

Nous avons comparé les formes proposées par le voisinage classique et celles proposées par cette méthode des formes réduites, en terme de précision et de rappel, et les résultats sont nettement à l'avantage de la seconde. On mesure en effet une précision de 60% si l'on cherche les formes correctes par voisinage, tandis que la même recherche avec les formes réduites donne une précision de 92%.

Exemple : pour la forme erronée *sante, les méthode des formes réduites ne propose que la correction « santé » (qui est la seule pertinente si l'on se réfère aux occurrences dans le corpus frWac), tandis que la méthode des voisins fait 12 propositions (gante, hante, jante, mante, sainte, sanie, santé, saute, sente, tante, usante, vante), que l'on peut à la rigueur ramener à 5 si l'on exclut de modifier la première lettre.

	Formes à corriger	Formes corrigées	Corrections proposées	Précision	Rappel	Facteur F1
voisinage	16072	13046	21687	60,2%	81,2%	69,1%
forme réduite	16072	11297	12245	92,3%	70,3%	79,8%

Tableau 3 : comparaison des deux méthodes de recherche de formes correctes

La différence entre les deux méthodes est surtout sensible pour les unités courtes, du fait que le réseau lexical (des formes correctes) est beaucoup plus dense dans ce cas, comme nous l'avons déjà dit.

3.5 Correction avec un phonétiseur

Après ces deux premières passes (correction des diacritiques et correction par voisinage ou par formes réduites), le recours à un phonétiseur permet souvent de corriger certaines erreurs résiduelles [Véronis 1988]. Toutefois, la confrontation d'un phonétiseur avec les erreurs réellement commises révèle les faiblesses de ce procédé ; en effet, un phonétiseur est « réglé » pour prononcer parfaitement des suites de mots correctement écrits, tandis que les erreurs orthographiques sont souvent dues à une méconnaissance des règles de prononciation du français. L'erreur courante « acceuil » en est un bon exemple, car le phonétiseur applique dans ce cas les règles « normales » et transforme de double c en /ks/ à cause du e qui le suit. Ne trouvant pas d'homophone dans sa base de mots corrects, il ne peut pas corriger cette erreur (ceci est vérifiable sur le site d'un dictionnaire de référence du français qui emploie ce type de correction dans son interface de requêtes).

3.6 État actuel de la base

Nous donnons ici un aperçu de l'état actuel de la base, mais il va de soi que ce tableau évolue constamment au fur et à mesure des corrections.

Fréquence	Formes distinctes	Formes correctes	Formes corrigées	Formes non reliées	Formes indéterminées
Supérieure à 50000	264	263	1	0	0
De 5000 à 50000	7476	7084	390	2	0
De 500 à 5000	29116	23069	5727	320	0
De 100 à 500	47883	26932	17661	3290	0
De 50 à 100	39652	12107	4670	0	22785
De 20 à 50	90478	18013	8354	0	64111
De 5 à 20	351332	36657	11833	0	302842
Total	**566201**	**124125**	**48636**	**3612**	**389738**

Tableau 4 : répartition des formes dans la base actuelle

Nous n'avons pas encore exploré les formes (nombreuses : environ 170 000) de fréquence inférieure à 8, c'est encore une « terra incognita » dans laquelle nous n'avons pas lancé nos automates à la recherche d'appariement avec les formes correctes.

Précisons enfin que les « formes non reliées » sont celles que nous avons reconnues comme erronées, mais que nous n'avons pas reliées à une correction, parce qu'elle ne sont pas à proprement parler des « erreurs orthographiques » : barbarismes (par ex. *accointage), mots étrangers (par ex. *because), etc., ou que leur forme correcte ne se trouve pas dans la base. Les « formes indéterminées » n'ont pas encore été contrôlées, elles possèdent une fréquence inférieure à la limite actuelle de nos investigations.

4 Validation

Notre liste de mots provient des requêtes formulées par les internautes, sa constitution ne suit donc pas les principes « classiques » rencontrées en linguistique[2] (extraction de vocabulaire à partir d'un corpus de textes) ou en psycholinguistique (recueil de productions manuscrites). Il était donc important de la « valider » en la comparant sur certains points avec des données plus habituelles.

4.1 Comparaison avec un corpus

Cette comparaison a déjà fait l'objet d'une étude et d'une présentation en colloque [Manguin, 2009], mais il convient de la mentionner ici. Après avoir remarqué que les erreurs les plus fréquentes, après les diacritiques, sont celles qui portent sur les consonnes doubles, et en particulier sur le double n, nous avons choisi 351 formes contenant le motif « onn » (assez courant en français), et nous avons relevé dans notre base la fréquence de chacune de ces formes, ainsi que la fréquence de la forme correspondante erronée où le motif « nn » était remplacé par « n ».

Nous avons ensuite relevé ces fréquences pour les mêmes formes sur un corpus en ligne de type tout-venant (afin d'éviter le biais de la correction automatique) composé des textes de 10 forums du domaine « .fr ». La comparaison des taux d'erreurs pour chaque forme a montré un taux de corrélation tout à fait satisfaisant. Nous avons ainsi pu considérer que notre liste de mots est un assez bon reflet de la production écrite par les internautes. Nous donnons ci-après la figure qui illustre la corrélation mesurée entre les taux d'erreurs (échelle logarithmique).

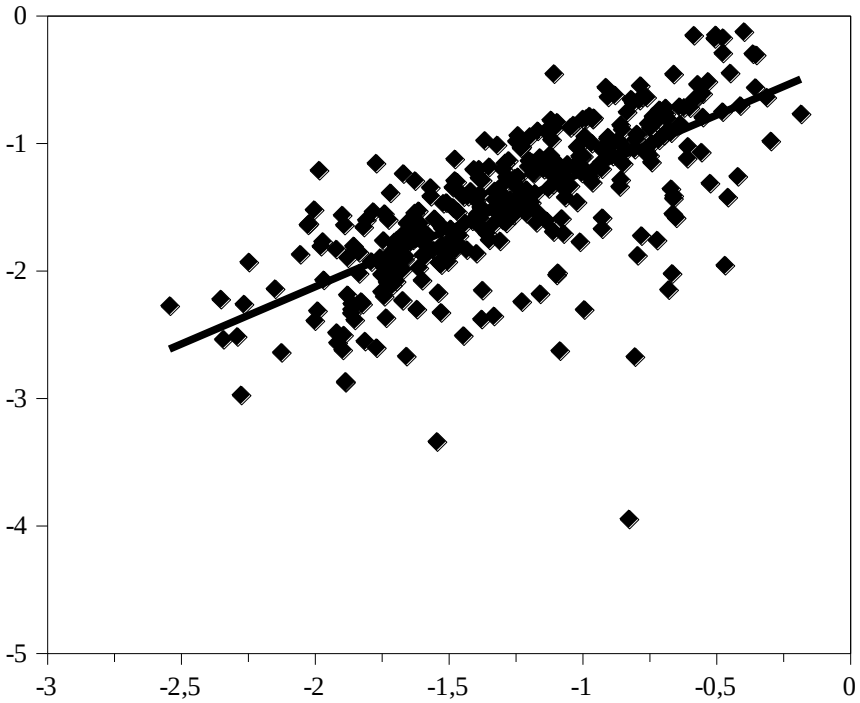

Figure 1 : corrélation base-corpus sur un échantillon d'erreurs

4.2 Comparaison avec une dictée réelle

Cette comparaison a été effectuée par Chloé Olivier [2010] sous la direction d'Arnaud Rey, en demandant à 100 sujets de participer à une dictée de 100 mots, et nous avons déjà commenté cette comparaison dans une étude précédente [Manguin, Rey et Olivier, 2013]. Rappelons que ces 100 mots ont été sélectionnés pour leur capacité à provoquer plusieurs types d'erreurs chez les scripteurs ; par exemple « recenser » offre 3 points principaux de difficulté à la transcription phono-graphémique : « c », « en » et « s », ce qui donne un ensemble de formes erronées plus riche. Le résultat final de la dictée est une liste de 10 000 items, représentant 754 formes distinctes. La comparaison entre les fréquences (relatives) des erreurs entre les deux listes de formes (celle issue de notre base, et celle issue de la dictée) révèle une corrélation assez bonne, mais aussi une grande dispersion des résultats, qui n'est pas explicable par la différence d'ordre de grandeur des fréquences. Nous discuterons cette différence dans le paragraphe suivant.

Mais malgré ces différences quantitatives, les erreurs commises montrent néanmoins un bon accord qualitatif ; pour mesurer celui-ci, nous avons étudié mot par mot le classement des formes produites selon leurs fréquences. Par exemple, pour le verbe « commettre », les rangs des formes sont les suivantes :

	requêtes		dictée	
	fréq.	rang	fréq.	rang
commettre	8579	1	67	1
commetre	1534	2	5	3
comettre	1528	3	23	2
commaître	25	4	2	4
comètre	15	5	2	5
comaittre	1	6	1	6

Tableau 5 : les erreurs sur le verbe « commettre »

Nous remarquons pour ce verbe une seule inversion dans le classement des formes, ce qui indique un très bon accord (93 % pour être précis). En procédant de même pour les 100 mots de la dictée, nous obtenons un accord qualitatif qui vaut 88 %, ce qui est aussi une très bonne valeur.

4.3 Discussion

Deux raisons expliquent les différences quantitatives entre notre liste de formes et celle issue de la dictée manuscrite. Tout d'abord, notre liste est constituée de requêtes qui n'ont pas été forcément tapées au clavier dans le formulaire d'accès ; cette interface en ligne permet en effet de coller le texte de la requête depuis un traitement de texte, et si ce dernier emploie déjà un correcteur orthographique, le risque d'erreur est notablement réduit. Ensuite, le niveau des scripteurs intervient aussi (mais à un degré moindre) : nous savons en effet que le dictionnaire des synonymes est principalement utilisé par des professionnels de la plume. Il est donc normal que ceux-ci commettent moins d'erreurs dans leurs requêtes. Ainsi s'expliquent les différences observées entre nos deux listes.

5 Interface d'accès

Cette interface a été développée par Meriem Khodja [Khodja 2015]dans le cadre de son projet de Master 1 à l'Université de Caen, dont le but principal était d'implanter une nouvelle architecture de la base de données sous moteur de type SQL. Cette implantation sous un moteur standard offre des perspectives de développements ultérieurs qui bénéficieront de la puissance et du vocabulaire liés à cette technique.

L'accès à cette interface se fait par le site https://ortholexies.greyc.fr/ .

5.1 Interface de requêtes

Voici la page d'accueil où l'utilisateur peut formuler sa requête :

Figure 2 : page d'accueil de l'interface de la base Ortholexies

La case à cocher « Recherche des voisins » ajoute les voisins orthographiques au résultat constitué normalement par les formes erronées recensées dans la base.

5.2 Page de résultats

La présentation des résultats est relativement simple : on indique à l'utilisateur si la forme qu'il a demandée est correcte ou erronée, puis selon cette première information, on y ajoute un complément qui peut être de deux sortes. Si la forme est correcte, on donne la liste des formes erronées répertoriées dans la base, avec leurs fréquences respectives ; si elle est erronée, on donne la correction (ou les corrections). En effet, il existe des cas où une forme erronée dérive de deux formes correctes, par exemple « contater », où l'erreur est la suppression d'une consonne de « contacter » ou bien de « constater » (cette double possibilité de correction est confirmée par l'examen des exemples trouvés en corpus).

Comme nous l'avons dit, il est également possible d'accéder aux voisins orthographiques de la forme demandée ; ci-après un exemple de page de réponse avec les voisins, pour la requête « écueil » :

Résultat de votre recherche

Le mot **écueil** (f=7051) est correct !

Des exemples de écueil

Les erreurs sur **écueil**

Forme	Fréquence
ecueil (exemples)	**3764**
eceuil (exemples)	**473**
éceuil (exemples)	**407**
eccueil (exemples)	**354**
ecceuil (exemples)	**190**
écueuil (exemples)	**130**
éccueil (exemples)	**114**

Les exemples sont fournis par le projet Intercorp de l'Université Charles IV à Prague.

Pour y accéder, merci de créer un compte (gratuit) sur le serveur de ce projet.

Les voisins de **écueil**

Forme	Fréquence
ecueil	**3764**
écueils	**730**
acueil	**332**
écureil	**185**
écueuil	**130**

Tapez un mot ici... RECHERCHER

Recherche des voisins

Figure 3 : exemple de page de réponse à une requête

La partie inférieure de cette page de résultats est intéressante, car elle montre que les voisins orthographiques détectés peuvent être des formes correctes (comme « écueils », signalé par un fond vert), ou des formes erronées d'autres mots (comme « écureil »). Par ailleurs, on se rend compte dans cet exemple que les formes erronées courantes ne sont pas détectées par voisinage. Cela confirme ce que nous avons déjà dit à propos du voisinage, et renforce l'intérêt d'employer de préférence les formes réduites.

5.3 Liaison avec un corpus

A chaque forme est ajouté un lien vers les exemples tirés du corpus frWac [Baroni et al. 2009] qui contient environ 1,6 milliard de mots, constitué de textes tirés du Web. Ces exemples sont présentés par un concordancier nommé Kontext, réalisé par l'Université de Prague[3] [cf. Rychlý, P. 2007 et Machálek, T. & Křen, M. 2013], comme ci-après :

dinde femelle sur le havre a donner merci de me	contater	sur msn mailto:voyoux-t-ou@hotmail.fr 19 - Jan - 08 Recherche :
blog ! j' eviterrai le language sms ! pour me	contater	: flo_t16@hotmail.fr merci ! welcome to the AUTO-LAND bienvenu à
. Si vous êtes intéressés par ces offres vous pouvez	contater	à la mairie de Castelmaurou Nicole Loze , maire adjoint
. Pour tout renseignement complémentaire , n' hésitez pas à	contater	le service archéologie de Noyon ou directement la DRAC de
je suis a peu pres sur que l' on pourrait	contater	la meme chose au sujet d' animaux . Merde ,
pour un bon prix . n' hésitez pas à me	contater	en indiquant la référence exact de ce que vous me
cautions . bail minimum 6 mois renouvelables Pour visite me	contater	sur le profil . Réponses : Posté par : Big
d' informations complémentaires ou de propositions devisées , veuillez nous	contater	. Toutes nos offres Audit , Hébergement - Cartographie Webmarketing
l' offre , sur les caractéristiques mais force est de	contater	qu' Apple met sur le marché le premier " mac
, tableau que je reproduis ici (fort heureux de	contater	que le " modèle belge " est loin de faire

Figure 4 : extrait des concordances trouvées par Kontext

Nous donnons ici quelques exemples trouvés avec la forme « contater » qui montre, comme nous l'avons signalé plus haut, que cette forme découle à la fois de « constater » et de « contacter ».

6 Applications et développements

6.1 Études de psycholinguistique

La base que nous avons constituée devait à l'origine servir de support à des études de psycholinguistique, principalement sur l'apprentissage de l'orthographe. Plusieurs informations sont en effet intéressantes dans les données recueillies : au point de vue qualitatif tout d'abord, la présence ou l'absence de certaines formes sont des indices précieux concernant le mécanisme de l'erreur orthographique et de la production écrite en général. Nous avons pu ainsi, dans une étude précédente, montrer que nos données confirment les observations déjà formulées par N. Catach [2001 et 2003] ou par J.P. Jaffré [2008] et qui font appel à la notion de « difficulté phonogrammique ». Outre cela, au point de vue quantitatif, nous avons aussi observé grâce aux fréquences des différentes formes erronées, une concordance avec les résultats de Wing et Baddeley [2009] sur les mots de l'anglais. Ces deux chercheurs ont effectivement remarqué que la partie finale des mots était plus souvent le siège d'erreurs que la partie initiale[4], interprétant ce résultat comme étant un effet de mémoire dans le buffer graphémique[5]. En outre, nous avons également montré que les données de notre base pouvaient confirmer les observations faites par Deacon et Bryant [2006] concernant la fréquence des erreurs dans les morphèmes.

Par ailleurs, nos données quantitatives aident à déterminer les formes les plus sensibles aux erreurs, et fournissent aussi les formes erronées les plus fréquentes. Ces informations sont précieuses et leur efficacité a été démontrée dans une expérience d'apprentissage implicite à laquelle nous avons collaboré. Cette expérience a en effet montré que la soumission à des formes correctes par la lecture induit une amélioration de la production écrite sur ces mêmes formes, et surtout que la soumission aux formes erronées les plus fréquentes induit l'effet contraire, autrement dit une baisse des performances à l'écrit. De plus, le point de vue quantitatif de ces résultats fait ressortir des corrélations nettes entre les points de difficulté phonogrammique et leurs fréquences dans le lexique [Le Goff K., Pacton S., Manguin, J.-L. & Rey, A. (2013)].

6.2 Traitement automatique

Le fait de fournir de manière instantanée une correction (ou du moins un très faible nombre de corrections) pour traiter une forme erronée constitue un atout majeur de ce type de ressource. En effet, les algorithmes de correction basés sur le voisinage ou la phonétisation présentent des inconvénients que nous avons signalés plus haut, et en outre exigent un temps de calcul parfois non négligeable.

D'autre part, la taille de notre ressource est négligeable si on la compare aux gigantesques bases de données constituées de bigrammes ou de n-grammes. Nous pensons donc qu'une utilisation en traitement automatique n'est pas dépourvue d'intérêt, c'est pourquoi nous avons débuté un projet d'annotation automatique de corpus d'apprenants, en collaboration avec l'Université de Prague.

6.3 Développements

Hormis le développement du contenu de la base par enrichissement des appariements, la suite de notre projet va s'orienter selon deux axes : l'évolution de l'interface, et l'intégration d'informations supplémentaires dans la base. L'interface pourra en effet rapidement évoluer en tirant profit de l'architecture SQL de la base, et permettre une interrogation (en mode expert) utilisant les caractères « joker » ou les expressions régulières ; ceci donnera aux utilisateurs la possibilité de travailler sur des familles de mots (par exemple). Outre cela, nous projetons d'inclure des informations morphologiques, ou du moins d'ajouter aux formes correctes leur lemme, afin d'inclure dans l'interface une possibilité de recherche par lemme, ceci dans le but de comparer les erreurs en fonction de la flexion.

7 Conclusion

Après avoir signalé que la construction de notre base répondait à un besoin de la communauté des psycholinguistes et venait combler un vide en matière de données publiques sur les erreurs d'orthographe lexicale, nous avons évoqué les problèmes liés à la constitution de cette base et décrit les solutions que nous avons apportées à ces questions. Nous avons également évoqué les applications déjà mises en oeuvre et les possibles projets susceptibles de tirer profit de nos données. Enfin, nous avons tracé les premières grandes lignes des évolutions de notre base qui soient dignes d'intérêt. Il nous reste à signaler que la finalité de cette base est d'offrir un support au plus grand nombre, et que nous gardons comme point de mire la possibilité de laisser ces données à la disposition du public sous une licence qui permettra aisément l'utilisation de nos données dans les travaux de recherche ou les applications en traitement des langues naturelles.

Références bibliographiques

Baranes, M. (2015). *Normalisation orthographique de corpus bruités.* Thèse de l'Université Paris Diderot.

Baroni M., Bernardini S., Ferraresi A., Zanchetta E. (2009). The WaCky Wide Web : A collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation* 43(3), pp. 209-226.

Bonin, P. (2007). *Psychologie du langage.* Louvain, De Boeck.

Catach, N. (2003 nouvelle éd.). *L'orthographe.* Paris, PUF, collection "Que sais-je ?".

Catach, N. (2001). *Histoire de l'orthographe française.* Paris, Honoré Champion.

Crisco (1998). *Dictionnaire électronique des synonymes.* http://www.crisco.unicaen.fr/

Damerau, F. (1964). A Technique for Computer Detection and Correction of Spelling Errors. *Communications of the ACM, Vol. 7(3),* pp. 171-176 .

Deacon S. H., et Bryant P. (2006). This turnip's not for turning : children's morphological awareness and their use of root morphemes in spelling. *British Journal of Developmental Psychology, Vol. 24,* pp. 567-575.

Jaffré J.-P. (2008). *Nouvelles recherches en orthographe.* Paris, Lambert-Lucas.

Khodja M. (2015). *Enrichissement de la base « lexiques.greyc.fr » et de son interface.* Mémoire de projet annuel Master 1, Université de Caen.

Le Goff K., Pacton S., Manguin, J.-L. & Rey, A. (2013). The interference of misspellings on spelling performance. *poster for the 25th Association for a Psychological Science (A.P.S.) Convention*. Washington D.C.

Lucci V., et Millet, A. (1994). *L'orthographe de tous les jours : enquête sur les pratiques orthographiques des français.* Paris, Honoré Champion.

Machálek, T. et Křen, M. (2013). Query interface for diverse corpus types. *Natural Language Processing, Corpus Linguistics, E-learning*, pp. 166–173. Lüdenscheid, RAM Verlag.

Manguin, J.-L. (2009). Les requêtes sur un site Web : un corpus pour étudier la variation orthographique. *Journées Internationales de Linguistique de Corpus.* Lorient.

Manguin, J.-L., Rey A. et Olivier C. (2013). Corpus orthographiques : une convergence entre linguistique et psycholinguistique. *Journées Internationales de Linguistique de Corpus.* Lorient.

Olivier C. (2010). *Apprentissage sans erreur de l'orthographe : validation d'une base de données informatique,* Mémoire de recherche Master 2, Université de Provence, Aix-Marseille.

Romary C., Salmon-Alt S., et Francopoulo G. (2004). Standards going concrete : from lmf to morphalou. *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*, pp. 22–28, Genève, Suisse.

Rychlý, P. (2007). Manatee/Bonito - A Modular Corpus Manager. *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pp. 65–70. Brno, Masaryk University.

Véronis J. (1988). Computerized correction of phonographic errors. *Computers and the Humanities, 22(1),* pp. 43–56.

Wing, A., et Baddeley, A. (2009). Righting errors in writing errors: The Wing and Baddeley (1980) spelling error corpus revisited, *Cognitive neuropsychology, vol. 26, 2*, pp. 223-226.

---

[1] Mais ce repérage des formes correctes est dépendant de Morphalou ; l'examen manuel des formes « erronées » permet de trouver d'autres formes correctes qui ne sont pas présentes dans la référence. Ce travail humain a été accompli jusqu'au seuil fréquentiel f=100.

[2] N'oublions pas néanmoins que Google (et ses concurrents) considèrent ces listes de requêtes comme un matériau de choix pour certaines applications en traitement automatique des langues.

[3] Je remercie le professeur Olga Nádvorniková qui a permis la mise en relation de nos deux bases de données.

[4] Mais cependant bien moins que la partie médiane.

[5] Le buffer graphémique est une mémoire tampon entre le cerveau et la commande motrice de l'écriture ; il est nécessaire à cause de la différence de vitesse entre la conception du mot dans le cerveau et sa réalisation par le geste. Pour plus de détails, voir [Bonin, 2007].