TALN 2010, Montréal, 19–23 juillet 2010

Recueil et analyse d'un corpus écologique de corrections orthographiques extrait des révisions de Wikipédia

Guillaume Wisniewski Aurélien Max François Yvon
LIMSI—CNRS, B.P. 133 91 403 ORSAY CEDEX
Université Paris Sud 11
{guillaume.wisniewski,aurelien.max,francois.yvon}@limsi.fr

Résumé. Dans cet article, nous introduisons une méthode à base de règles permettant d'extraire automatiquement de l'historique des éditions de l'encyclopédie collaborative Wikipédia des corrections orthographiques. Cette méthode nous a permis de construire un corpus d'erreurs composé de 72 483 erreurs lexicales *(non-word errors)* et 74 100 erreurs grammaticales *(real-word errors)*. Il n'existe pas, à notre connaissance, de plus gros corpus d'erreurs écologiques librement disponible. En outre, les techniques mises en œuvre peuvent être facilement transposées à de nombreuses autres langues. La collecte de ce corpus ouvre de nouvelles perspectives pour l'étude des erreurs fréquentes ainsi que l'apprentissage et l'évaluation des correcteurs orthographiques automatiques. Plusieurs expériences illustrant son intérêt sont proposées.

Abstract. This paper describes a French spelling error corpus we built by mining Wikipedia revision history. This corpus contains 72,493 non-word errors and 74,100 real-word errors. To the best of our knowledge, this is the first time that such a large corpus of naturally occurring errors is collected, which open new possibilities for the evaluation of spell checkers and the study of errors patterns. In the second part of this work, a first study of french spelling error patterns and of the performance of a spell checker is presented.

Mots-clés : ressources, correction orthographique, Wikipédia.

Keywords: resources, spelling correction, Wikipedia.

1 Introduction

Cet article décrit la création d'un corpus d'erreurs orthographiques à partir des révisions des pages de Wikipédia en français. Ce corpus contient 72 493 erreurs lexicales *(non-word errors)* et 74 100 erreurs grammaticales *(real-word errors)* écologiques. C'est, à notre connaissance, la première fois qu'un aussi gros corpus d'erreurs (et de corrections) est collecté et rendu disponible, ce qui ouvre de nouvelles perspectives pour l'étude des erreurs ainsi que pour l'évaluation et l'apprentissage de correcteurs automatiques.

Les corpus jouent un rôle moteur dans le développement et l'analyse de systèmes de traitement des langues ; il n'est que plus regrettable que, pour de nombreuses tâches, de tels corpus ne soient pas publiquement disponibles. C'est en particulier le cas en correction orthographique, la plupart des évaluations, (Islam & Inkpen, 2009), utilisent des petits corpus artificiels, car les sources de données usuelles (articles de journaux, rapports, ...) ne comportent que peu de fautes et leur annotation a un cout élevé.

L'utilisation de corpus artificiels limite toutefois l'utilisation de méthodes d'apprentissage statistique, qui nécessitent des corpus représentatifs de la tâche. Il en va de même pour leur utilisation pour l'évaluation : il n'y a aucune garantie qu'un correcteur capable de détecter des erreurs introduites artificiellement ne soit biaisé et ne sache reconnaitre que celles-ci. *À contrario*, les erreurs introduites artificiellement risquent d'être plus compliquées à détecter et à corriger que les erreurs réelles.

Nous souhaitons montrer, dans cet article, que le développement des wikis tel Wikipédia et, plus généralement, des systèmes de gestion de révisions peut aider à la construction de corpus « naturels ». Une des particularités de ces systèmes est, en effet, de conserver un historique complet de toutes les modifications. Il est donc possible d'accéder aux révisions successives d'un document pour en extraire, non seulement les modifications, ajouts ou suppressions d'informations, mais également toutes les corrections et modifications de style. La collecte de ce dernier type de modifications permet de constituer aisément des corpus pour plusieurs tâches de TAL (Nelken & Yamangil, 2008) et notamment pour la correction orthographique.

Notre contribution est triple. Dans la première section, nous décrivons la construction d'un corpus d'erreurs écologiques à partir des révisions de Wikipédia. Puis nous présentons les premières analyses de ce corpus qui nous permettent d'identifier les types d'erreurs fréquentes. Finalement, dans la troisième section, nous conduisons une première évaluation d'un correcteur libre de l'état de l'art, Hunspell, et montrons comment notre corpus d'erreurs pourrait faciliter la construction d'un correcteur performant.

2 Construction du corpus

Notre corpus est une sous-partie de WICOPACO, un corpus qui regroupe des corrections orthographiques, des corrections typographiques, ainsi que des reformulations. Nous décrivons dans cette section la construction de WICOPACO, puis de notre corpus d'erreurs.

2.1 Le corpus WICOPACO

Le corpus WICOPACO *(Wikipedia Correction and Paraphrase Corpus)* (Max & Wisniewski, 2010) est un corpus de modifications locales extrait des révisions des articles de Wikipédia. Sa construction repose sur l'observation que la plupart des révisions « mineures » d'un article (celles qui ne portent que sur

quelques mots) sont des corrections d'erreurs (orthographiques, grammaticales, typographiques, ...) ou des améliorations du style. L'extraction de ces modifications permet de constituer des corpus pour plusieurs tâches de traitement automatique des langues (Nelken & Yamangil, 2008). La construction de ce corpus se compose de deux étapes. Dans une première étape, un ensemble de modifications locales est extrait[1]. Pour cela, nous calculons l'ensemble des différences textuelles entre deux versions d'une même page à l'aide d'un algorithme de recherche de plus grandes sous-séquences communes[2]. L'objectif étant d'extraire des modifications locales, seules les modifications portant sur sept mots au plus sont prises en compte.

Cette première étape permet d'extraire un très grand nombre de modifications locales. Nous appliquons, dans une seconde étape, un ensemble de filtres afin de ne sélectionner que les plus intéressantes. En particulier, les modifications qui ne conservent pas un minimum de mots et qui ne concernent que des signes de ponctuation ou des changements de casse sont exclues. Le premier filtre permet de rejeter (de manière grossière) des corrections « sémantiques » ne conservant pas le sens ; le second permet de limiter la taille du corpus.[3]

Les modifications extraites sont ensuite normalisées (notamment en supprimant toutes les informations de mise en page), segmentées et sauvegardées dans un format XML. Lors de l'extraction, les informations permettant de faire le lien entre la modification et la page Wikipédia sont conservées, et le contexte (le paragraphe dans lequel la modification est effectuée) est également extrait. La figure 1 présente un exemple extrait du corpus WICOPACO et illustre les informations disponibles.[4]

```
<modif id="23" wp_page_id="7" wp_before_rev_id="4649540"
       wp_after_rev_id="4671967" wp_user_id="0"
       wp_user_num_modif="1096911" wp_comment="Définition">
  <before>On nomme <m num_words="1">Algebre</m> linéaire la branche
  des mathématiques qui se penche...</before>
  <after>On nomme <m num_words="1">Algèbre</m> linéaire la branche
  des mathématiques qui se penche...</after>
</modif>
```

FIG. 1 – Exemple d'entrée de WICOPACO. Les attributs commençant par `wp_` correspondent aux index de Wikipédia et permettent de retrouver la révision correspondant à la modification ; les segments modifiés sont signalés par la balise `m`.

La version actuelle de WICOPACO[5] comporte 408 817 modifications. Une analyse des révisions extraites permet de distinguer trois grands types de modifications présentés dans le Tableau 1. Nous exploitons actuellement ce corpus pour l'analyse des erreurs orthographiques et l'évaluation des correcteurs. Ce corpus présente également un intérêt particulier pour l'étude des phénomènes de reformulation, et nos travaux futurs incluent l'identification automatique de paraphrases à l'intérieur du corpus.

[1] Les modifications effectuées par les « robots de correction » de Wikipédia sont ignorées (voir `http://fr.wikipedia.org/wiki/WikipÃl'dia:Bot`).

[2] Nous avons utilisé une implémentation identique à celle du programme `diff` standard.

[3] Une autre version du corpus ne comportant que des corrections typographiques est en cours de réalisation.

[4] Pour plus de clarté, le contexte de la modification a été réduit.

[5] Téléchargeable à l'adresse `http://wicopaco.limsi.fr`. Elle a été extraite à partir 85 000 articles de la version

Correction	
Normalisations	◇ ***[Son 2ème disque → Son deuxième disque]***
Erreurs lexicales	◇ *c'est-à-dire la* ***[dernrière → dernière]*** *année avant l'ère chrétienne*
Corrections des diacritiques	◇ *la jeune Natascha Kampusch,* ***[agée → âgée]*** *de 18 ans*
Corrections grammaticales	◇ *dans le but de* ***[sensibilisé → sensibiliser]*** *sur les changements*
Reformulation	
sans changement de sens	◇ *Le tritium* ***[existe dans la nature . Il est produit → se forme naturellement]*** *dans l' atmosphère*
	◇ *"Gimme Gimme Gimme" et "I Have A Dream"* ***[contribueront au gigantesque succès de → viendront alimenter la gloire d']*** *Abba*
avec changement de sens	◇ *alors* ***[que l' ordinateur → qu'un processeur de la famille x86]*** *reconnaîtra ce que l' instruction machine*
	◇ ***[Le principal du collège M. Desdouets → Un de ses professeurs]*** *dit de lui*
	◇ *Des opérations de base sont disponibles dans* ***[tous les → la plupart des]*** *jeux d' instructions*
Vandalisme	
Vandalisme agrammatical	◇ *Süleyman Ier s'* ***[empare de l' Arabie et fait entrer dans l' → emp kikoo c moi ca va loll ']*** *empire ottoman Médine et La Mecque*
Vandalisme subtil	◇ *pour promouvoir la justice , la solidarité et* ***[la paix → l'apéro]*** *dans le monde*

TAB. 1 – Typologie des modifications présentes dans le corpus WICOPACO

2.2 Constitution d'un corpus de corrections orthographiques

Le corpus WICOPACO permet de construire facilement un corpus de corrections orthographiques : comme le montrent les extraits présentés Table 1, il suffit pour cela de distinguer, parmi toutes les entrées du corpus, les corrections des reformulations et du vandalisme. Il est également possible d'extraire la correction de cette erreur, en faisant une hypothèse forte : il faut supposer qu'aucune « petite » modification avec la dernière version d'un article (au moment du téléchargement) n'introduit d'erreur et que le contenu après la modification peut donc être considéré comme une « référence ». Une étude rapide du corpus montre que cette hypothèse est valable dans la grande majorité des cas.

Nous souhaitons également classer les erreurs selon les deux catégories traditionnellement identifiées (Kukich, 1992) : les erreurs « lexicales » *(non-word errors)*, pour lesquelles le mot mal orthographié n'est plus un mot valide de la langue (p. ex., lorsque « maman » est écrit « maaman ») et les erreurs « grammaticales » *(real-word errors)* qui correspondent aux cas où le mot mal orthographié reste un mot valide de la langue. En plus des erreurs grammaticales à proprement parler (p. ex., lorsque « mangés » est écrit « manger »), cette dernière catégorie regroupe également certaines erreurs lexicales (p. ex., lorsque « pour » est écrit « pur »). Les erreurs grammaticales ne peuvent être détectées qu'en prenant en compte le contexte dans lequel le mot apparait.

Le corpus d'erreurs est construit en sélectionnant les modifications ne comportant ni signe de ponctuation, ni chiffre, ni nombre écrit en toutes lettres (sauf « un » et « une »), ni plus d'une lettre en majuscule. Ces critères permettent d'écarter des modifications ne portant pas sur l'orthographe du mot et notamment

d'octobre 2007 de la Wikipédia francophone. Ce corpus est naturellement appelé à s'enrichir pour prendre en compte des versions plus récentes.

certaines corrections de nature sémantique (p. ex. lorsque « sept » est corrigé en « six »). Les modifications portant sur plus d'un mot sont également rejetées.

Comme l'on dispose, pour chaque modification, du mot corrigé (avant et après correction), la distinction des erreurs peut être faite quasi automatiquement en deux étapes simples. Dans une première étape, nous utilisons un correcteur orthographique[6] pour identifier si les mots avant et après correction sont des mots valides. Trois cas de figure peuvent se présenter :

1. le mot avant modification n'est pas un mot valide, mais le mot après modification l'est ; ce cas correspond à la correction d'une erreur lexicale ;
2. le mot avant modification et le mot après modification sont des mots valides ; ce cas correspond à la correction d'une erreur grammaticale *ou* à une reformulation ;
3. le mot après modification n'est pas un mot valide ; ce cas correspond soit à l'introduction d'une erreur (mauvaise correction ou introduction de vandalisme), à la modification d'un nom propre, d'un mot étranger ou d'un mot inconnu.

Ce premier traitement simple nous permet donc d'extraire de WICOPACO les modifications qui sont des corrections d'erreurs lexicales (cas 1) et de rejeter certaines modifications, dont la validité est plus difficile à établir, telles les corrections de noms propres et de mots étrangers (cas 3).

Pour distinguer, dans le cas 2, les reformulations des corrections d'erreurs, nous utilisons un critère fondé sur la distance d'édition.[7] En effet, plusieurs travaux ont montré que les mots mal orthographiés sont, en général, proches (au sens de la distance d'édition) de leur forme correcte (Kukich, 1992). L'étude d'un échantillon du corpus corrobore ce résultat : nous observons que dans une modification correspondant à au moins 4 éditions, le mot est généralement complètement réécrit et la correction est donc une reformulation. Il apparait également qu'une modification correspondant à plus que 5 éditions correspond en général à diverses formes de vandalisme. Le corpus d'erreur est donc construit à l'aide des règles suivantes :

– les **erreurs lexicales** sont les entrées correspondant à l'édition d'un mot inconnu en un mot connu, la correction donnant lieux à strictement moins que 6 éditions ;
– les **erreurs grammaticales** sont les entrées dans lesquelles un mot connu est remplacé par un autre mot connu suffisamment proche (la distance d'édition doit être strictement plus petite que 4).
– tous les autres cas sont rejetés.

L'application de ces règles permet d'extraire un corpus de 74 100 erreurs grammaticales et 72 493 erreurs lexicales en contexte. Ce corpus est également téléchargeable à partir de la page de WICOPACO.

3 Erreurs fréquentes en français

WICOPACO fournit des informations précieuses sur les distributions de patrons d'erreurs en français. Dans cette section, nous présentons les résultats de nos premières analyses statistiques de ces patrons.

La principale analyse que nous avons effectuée consiste à identifier les corrections (et donc les erreurs) les plus fréquentes dans le corpus. Étant donné le mot initial et le mot corrigé, la correction peut facilement être déterminée en calculant la distance d'édition entre ces deux mots et la suite d'opérations (ou *trace d'édition*) qui lui est associée.

[6] Dans toutes nos expériences, nous utilisons la version 1.2.8 du correcteur libre Hunspell `http://hunspell.sf.net` avec la version 3.4.1 du dictionnaire français « *Classique et réforme 90* ».

[7] Nous avons utilisé la distance de Levenshtein (Wagner & Fischer, 1974), avec un cout identique pour les trois opérations.

Il existe de nombreuses manières de définir la distance d'édition et la liste des modifications lui correspondant (Navarro, 1999). Nous avons utilisé l'algorithme de Ratcliff & Metzener (1988) qui produit des séquences d'édition plus facilement interprétables dans le cadre de la correction orthographique. Notons que la plupart des corrections fréquentes ne portant que sur un caractère, les résultats présentés dans cette section ne seraient pas fondamentalement modifiés si l'on considérait un autre algorithme.

3.1 Étude des erreurs lexicales

Les corrections des erreurs lexicales les plus fréquentes sont présentées dans la partie gauche de la Table 2. Les corrections les plus fréquentes portent sur des accents : au total, 32,4% de ces corrections consistent à ajouter, supprimer ou modifier un accent. La plupart de ces corrections peuvent difficilement être qualifiées d'erreurs : elles sont plus probablement dues à une méconnaissance des règles typographiques du français (accentuation des majuscules) ou à une mauvaise maitrise des dispositifs de saisie (les caractères accentués ou encore la ligature œ).

Lorsque l'on considère le « contexte[8] » des corrections, on observe qu'hormis les fautes d'accents, les corrections les plus fréquentes portent sur les doublements de consonnes (ajout d'un `n` après ou avant un autre `n`, par exemple). Plusieurs travaux ont déjà signalé la complexité des règles afférentes en français ainsi que la fréquence de ce type d'erreurs.

Il est finalement intéressant de noter que, lorsqu'une correction est fréquente, sa correction « inverse » l'est également : par exemple, l'ajout et la suppression d'un `s` sont, toutes deux, des opérations fréquentes.

3.2 Étude des erreurs grammaticales

La partie droite de la Table 2 présente les corrections les plus fréquentes pour les erreurs grammaticales. On peut observer que le corpus est essentiellement constitué de quelques corrections, très fréquentes, alors que de nombreuses corrections n'apparaissent qu'une ou deux fois. Il apparait donc que la plupart des erreurs se répètent, ce qui ouvre la possibilité de construire des *ensembles de confusions* regroupant les mots souvent mal orthographiés et leurs corrections (Kukich, 1992). Comme nous le détaillerons dans la section 4.1, ces ensembles permettent de faciliter la correction automatique.

Les corrections les plus fréquentes sont causées par des erreurs d'accentuation ou par des erreurs dans les accords des féminins (ajout/suppression du `e` final) et des pluriels (ajout/suppression du `s` final ou de la terminaison `nt`). Il est également intéressant de noter que, comme pour les erreurs lexicales quand une modification est fréquente, la modification inverse l'est également.

La Table 3 indique dans quelle partie du mot se situent les erreurs. Cette information se déduit directement de la séquence d'opérations permettant de corriger un mot. Les erreurs grammaticales se situent quasiment toutes dans la deuxième moitié des mots. En fait, une observation plus précise montre que 47,0% des corrections portent sur la partie finale du mot, ce qui confirme le fait que la plupart des erreurs correspondent à des confusions entre formes d'un même paradigme (typiquement, des problèmes d'accord).

Comme le suggère la Table 2, de nombreuses corrections (au moins 3%) consistent à récrire un mot, correctement orthographié selon la réforme de l'orthographe de 1990, selon la norme orthographique qui

[8] Défini ici comme la lettre précédant et la lettre suivant le lieu de la modification

Erreurs lexicales				Erreurs grammaticales			
e → é	6,7%	-l	1,9%	+s	16.2%	-t	1.5%
E → É	6,7%	+i	1,9%	+e	9.9%	e → a	1.4%
oe → œ	4,6%	a → â	1,8%	-s	8.8%	é → er	1.0%
+n	4,3%	-e	1,7%	A → À	5.6%	er → é	0.9%
+s	2,8%	-n	1,7%	-e	4.9%	u → ù	0.9%
+r	2,7%	+t	1,6%	i → î	2.7%	à → a	0.9%
é → è	2,7%	+m	1,6%	a → à	2.2%	e → é	0.8%
-s	2,5%	e → è	1,4%	+nt	1.9%	é → è	0.7%
+e	2,2%	+l	1,3%	+t	1.7%	s → t	0.7%
é → e	2,1%	-r	1,3%	a → e	1.5%	û → u	0.7%

TAB. 2 – Les vingt corrections les plus fréquentes. Ces corrections représentent 65,0% des corrections d'erreurs grammaticales du corpus et 53,5% des corrections d'erreurs lexicales

	erreurs lexicales	erreurs grammaticales
première moitié du mot	34,06%	4,08%
seconde moitié du mot	62,81%	93,26%
erreurs dans les deux moitiés	3,13%	2,63%

TAB. 3 – Localisation des erreurs à l'intérieur d'un mot

s'imposait antérieurement. Par exemple, évènement est presque toujours récrit en événement. De même, de nombreuses corrections réintroduisent un î à la place de i, alors que la réforme de 1990 fait pratiquement disparaitre les accents circonflexes sur le i. La fréquence de ces corrections peut s'expliquer soit par la présence de contributeurs « puristes » qui pensent qu'une encyclopédie doit respecter une orthographe plus « stricte » que celle de la réforme de 1990, soit à une ignorance des simplifications introduites par la dernière réforme.

3.3 Bilan

Les résultats présentés dans les sections précédentes illustrent quelques-unes des limites que notre méthode de détection automatique des erreurs rencontre dans un contexte où la norme autorise des *variations* et où l'usage est de plus en plus tolérant (par l'exemple sur l'accentuation des majuscules) : faute d'utiliser des ressources lexicales adéquates pour mieux filtrer les corrections, notre corpus agrège un certain nombre de modifications qui ne sont pas des corrections d'erreurs. Il est également important de noter que les statistiques présentées ci-dessous, si elles s'accordent globalement avec d'autres observations sur les difficultés du français (Catach, 1980), ne doivent pas être prises trop littéralement : notre corpus ne permet de mesurer que les fautes les plus courantes effectuées *au clavier* par des scripteurs dont on peut penser qu'ils sont globalement plus éduqués, plus familiés des nouvelles technologies, etc. que l'ensemble des scripteurs du français. Par ailleurs, l'utilisation de nombreux filtres nous a conduit à ignorer dans cette première version du corpus les modifications de ponctuation, les corrections qui conduisent à fusionner deux mots, ou au contraire à remplacemer un mot par deux. Il reste ici un gros effort d'analyse à accomplir pour mieux caractériser notre corpus. Nous montrons dans la suite qu'en dépit de ces limites, une exploitation raisonnée de ce corpus permet d'aider au développement d'outils de correction automatique.

4 Évaluation et apprentissage de correcteurs orthographiques

Nous proposons, dans cette section, de montrer comment le corpus d'erreurs que nous avons construit peut être utilisé à la fois pour évaluer les performances d'un correcteur orthographique automatique et pour apprendre certains des paramètres de celui-ci. Jusqu'à présent, l'évaluation des correcteurs ne s'est faite que sur des corpus artificiels (Islam & Inkpen, 2009). En utilisant un corpus d'erreurs écologiques, nous espérons développer un correcteur mieux adapté aux erreurs réelles des utilisateurs.

Aujourd'hui, la plupart des correcteurs orthographiques sont des systèmes d'« aide à la correction » : ils sont capables d'identifier les mots non valides d'une langue, de suggérer un ensemble de corrections possibles pour un mot donné (que ce mot ait été identifié comme une erreur ou non), mais ils laissent le choix de la bonne correction à l'utilisateur. Dans ce travail, nous nous concentrons sur l'évaluation de la qualité des corrections suggérées, puis introduisons, dans la deuxième partie de cette section, une première expérience portant sur la possibilité d'automatiser la correction.

4.1 Qualité de l'ensemble des suggestions

Deux critères doivent être pris en compte dans l'évaluation de l'ensemble des suggestions : le nombre de fois où la correction d'une faute est suggérée et le nombre de suggestions du système. Nous proposons donc d'utiliser la mesure suivante :

$$\text{qualité} = \frac{1}{N} \sum_i \frac{\delta_i}{\#s_i}$$

La somme se fait sur les N exemples de l'ensemble d'évaluation, δ_i vaut 1 si la i$^{\text{e}}$ correction est dans l'ensemble des suggestions, 0 sinon et $\#s_i$ est la taille de l'ensemble de suggestions correspondant.

Dans ce travail préliminaire, nous considérons trois moyens de construire l'ensemble des suggestions :

1. en considérant l'ensemble des suggestions proposé par Hunspell (méthode *hunspell*). Ces suggestions sont engendrées par une liste de règles qui décrivent des corrections possibles du lemme et des variantes morphologiques des formes corrigées[9]. Toutes ces règles sont écrites à la main.
2. en appliquant un ensemble de patrons des corrections au mot à corriger et en ne conservant que les applications générant un mot valide selon Hunspell (méthode *motif*). Nous utilisons les 20 patrons les plus fréquents du corpus, qui sont décrits dans la Table 2. Ces patrons (`i` $\rightarrow$ `î` par exemple) sont appliqués indépendamment les uns des autres et sans tenir compte du contexte.
3. en considérant la liste des corrections apparaissant dans le corpus (méthode *liste*). Cette liste est construite simplement en associant, pour chaque exemple du corpus, le mot mal orthographié à l'ensemble de ses corrections. Cette liste est semblable aux *ensembles de confusions* utilisés dans plusieurs correcteurs orthographiques (Islam & Inkpen, 2009; Carlson & Fette, 2007). Mais, comme nous disposons d'un corpus d'erreurs, la construction de cette liste est triviale et rapide, alors que dans la plupart des travaux, les ensembles de confusions sont construits manuellement.

La première approche permet d'évaluer un correcteur libre de l'état de l'art utilisé dans de nombreux produits « grand public » tels que le navigateur Firefox ou la suite bureautique OpenOffice. Les deux autres approches ont pour objectif de montrer l'impact que peut avoir l'utilisation d'un corpus d'erreurs avec les caractéristiques de celui que nous avons construit sur le développement de correcteurs.

[9] Ces règles sont écrites par des volontaires, en suivant un mode de développement similaire à celui des logiciels libres.

Pour évaluer ces trois approches, nous séparons aléatoirement le corpus en un ensemble d'apprentissage contenant 80% des exemples et un ensemble de test contenant les 20% restants. L'ensemble de test sert à mesurer la qualité des suggestions ; l'ensemble d'apprentissage à construire la liste des corrections.

Les résultats de cette expérience sont présentés Table 4. Ils montrent clairement qu'il est possible, en combinant ces trois approches, d'assurer que l'orthographe correcte d'un mot est toujours présente dans l'ensemble des suggestions. Notons également qu'Hunspell obtient de très bonnes performances pour la correction des erreurs lexicales, mais est moins utile pour la correction des erreurs grammaticales. Les deux types d'erreurs partagent pourtant plusieurs caractéristiques (notamment leur distance d'édition) et l'on aurait pu espérer que les règles de corrections d'Hunspell seraient suffisamment générales pour traiter certaines des erreurs grammaticales. Il est, par exemple, surprenant qu'Hunspell ne propose quasiment aucune variation morphologique des mots corrigés.

	Erreurs lexicales				Erreurs grammaticales			
méthode	qualité	moyenne	max.	corr.	qualité	moyenne	max.	corr.
hunspell	40,0%	4,5	15	95,0%	13,0%	8,6	15	65,1%
liste	51,1%	1,3	7	58,7%	32,1%	8,3	41	75,7%
motif	35,5%	1,7	11	48,7%	31,3%	2,3	7	53,2%
combi.	38,9%	4,7	22	96,8%	13,7%	14,9	47	92,6%

TAB. 4 – Qualité des ensembles de suggestions évaluée par la mesure introduite *(qualité)*, la taille l'ensemble de suggestion (*moyenne* et *max*), et le pourcentage d'erreurs dont la correction est suggérée *(corr.)*

4.2 Correction automatique des erreurs orthographiques

Nous présentons, Table 5, nos premiers résultats sur la correction automatique des erreurs orthographiques[10]. Ces résultats sont obtenus en ordonnant par un modèle de langage statistique de type n-gram, estimé sur le corpus d'apprentissage, les suggestions proposées par la combinaison des trois méthodes introduites dans la section précédente.

Malgré la simplicité du modèle (il n'y a aucune modélisation des erreurs et seul le contexte est utilisé pour classer les suggestions), les résultats sont plutôt encourageants : dans la majorité des cas, notre correcteur est capable de trouver automatiquement la bonne correction.

	erreurs lexicales		erreur grammaticales	
	corpus	corrections possibles	corpus	corrections possibles
modèle 3-gram	58,9%	60,8%	46,7%	51,9%
modèle 5-gram	75,2%	77,7%	66,9%	71,3%

TAB. 5 – Pourcentage de mots correctement corrigés sur tout le corpus de test et lorsque l'on ne considère que les mots dont la correction est possible

[10]Par manque de place ni ces résultats ni notre méthode ne sont détaillés.

5 Conclusion

Nous avons présenté et analysé un nouveau corpus d'erreurs d'orthographe écologiques, qui ouvre de nouvelles possibilités pour une étude *in vivo* de l'orthographe du français tel qu'il s'écrit électroniquement, ainsi que que pour l'évaluation et l'apprentissage de dispositifs de correction automatisée. Un des intérêts de ce corpus pour la correction orthographique est qu'il permet de développer des systèmes de correction plus adaptatifs, c'est-à-dire s'appuyant sur l'usage réel du français dans diverses situations de communication plutôt que sur une norme définie de manière abstraite et rigide. La méthodologie utilisée pour le construire est générique et ne requiert que la disponibilité de révisions successives d'un document : elle pourrait facilement être utilisée, par exemple, pour acquérir des corpus permettant d'adapter un correcteur aux erreurs typiques d'un locuteur à partir des révisions effectuées.

Des expériences préliminaires illustrant l'utilisation de ce corpus dans un contexte de correction automatique ont été présentées. WICOPACO peut être également utilisé dans de nombreuses autres tâches de TAL et nous sommes actuellement en train de l'exploiter pour l'étude des reformulations et l'identification des paraphrases, pour l'apprentissage et l'évaluation d'un système de correction automatique complet ainsi que pour le développement de système de normalisation de documents pour la traduction automatique.

Remerciements

Ce travail a été partiellement financé dans le cadre d'une Action Incitative du LIMSI et du projet TRACE (ANR CONTINT 2009). Les auteurs remercient Julien Boulet et Martine Hurault-Plantet pour leur aide.

Références

CARLSON A. & FETTE I. (2007). Memory-based context-sensitive spelling correction at web scale. In *ICMLA '07*, p. 166–171, Washington, DC, USA : IEEE Computer Society.

CATACH N. (1980). *L'orthographe francaise : traite theorique et pratique avec des travaux d'application et leurs corriges (avec la collaboration de Claude Gruaz et Daniel Duprez)*. Nathan, Paris.

ISLAM A. & INKPEN D. (2009). Real-word spelling correction using Google Web 1T 3-grams. In *Proceedings of EMNLP'09*, p. 1241–1249, Singapore.

KUKICH K. (1992). Techniques for automatically correcting words in text. *ACM Comput. Surv.*, **24**(4), 377–439.

MAX A. & WISNIEWSKI G. (2010). Mining naturally-occurring corrections and paraphrases from wikipedia's revision history. In *LREC'10*.

NAVARRO G. (1999). A guided tour to approximate string matching. *ACM Computing Surveys*, **33**, 2001.

NELKEN R. & YAMANGIL E. (2008). Mining wikipedia's article revision history for training computational linguistics algorithms. In *AAAI Workshop on Wikipedia and Artificial Intelligence*.

RATCLIFF J. W. & METZENER D. E. (1988). Pattern matching : The gestalt approach. *Dr. Dobb's Journal*.

WAGNER R. A. & FISCHER M. J. (1974). The string-to-string correction problem. *Journal of the ACM (JACM)*, **21**(1), 168–173.