Les nouveaux outils de correction linguistique de Microsoft

Thierry Fontenelle

Microsoft Speech & Natural Language Group, Redmond
thierryf@microsoft.com

Résumé

De nouveaux outils de correction linguistique sont disponibles pour le français depuis quelques mois. Mis à la disposition des utilisateurs de Microsoft Office 2003, un nouveau correcteur orthographique et un nouveau correcteur grammatical permettent d'améliorer le processus de rédaction de documents. En partant d'évaluations externes effectuées récemment, nous présentons les diverses facettes de ces améliorations et de ces outils, en abordant la question de l'évaluation des outils de correction linguistique (qu'évaluer ? quels critères appliquer ? pourquoi développer une nouvelle version ?). La réforme de l'orthographe, la féminisation des noms de métier, l'évolution de la langue figurent parmi les thèmes abordés dans cet article.

Mots-clés : correcteur orthographique, correcteur grammatical, français, outils de correction linguistique, Microsoft, réforme de l'orthographe, féminisation des noms de métier.

Abstract

New French proofing tools were recently made available to Microsoft Office 2003 users. A new spell-checker and a new grammar checker make it possible to improve the document creation process. Based on recent external evaluations, we present specific aspects of these improved tools and discuss the more fundamental issue of how to evaluate proofing tools (What do we need to evaluate? Which criteria should be applied? Why develop a new version?). Current language changes including the spelling reform, and such innovations as feminine job titles are among the themes we discuss in this paper.

Keywords: spell-checker, grammar checker, French, proofing tools, Microsoft, spelling reform, feminine job titles.

1. Introduction

Nous nous proposons de décrire dans cet article quelques-unes des facettes des nouveaux outils de correction linguistique que Microsoft vient de mettre à la disposition des utilisateurs de sa suite Office 2003. Des évaluations externes ayant montré que les améliorations de ces outils sont perceptibles tant par rapport aux versions précédentes que par rapport aux produits concurrents, il semble opportun de se pencher sur la question de l'évaluation des outils de correction linguistique et des critères à appliquer lorsqu'il s'agit de décrire les améliorations d'outils linguistiques tels qu'un correcteur orthographique ou un correcteur grammatical.

2. Évaluation

Il y a quelques mois, Jean Véronis publiait sur son site Web trois blogs relatifs au nouveau correcteur orthographique français de Microsoft, lancé par le biais d'une mise à jour publique en avril 2005 et intégré au Service Pack 2 d'Office 2003 en septembre 2005 (Véronis, 2005a ; 2005b ; 2005c). Dans les deux premiers blogs, il évaluait ce produit par rapport à la Toolbar de Google, qui offre depuis quelques mois une fonctionnalité de correction orthographique. Dans le dernier blog (2005c), il proposait une évaluation du correcteur de Microsoft par rapport à celui d'OpenOffice. Dans les deux cas, il utilisait le même texte de base, à savoir un

article assez court du journal *Le Monde* qu'il avait auparavant soumis à son « pourrisseur » de texte, un programme informatique qui injecte dans un texte des fautes en lui faisant subir toute une série de manipulations textuelles aléatoires (inversion, ajout ou suppression de caractères). Les blogs de Jean Véronis comportaient également une évaluation implicite du nouveau correcteur de Microsoft par rapport à la version précédente en présentant le nouveau correcteur comme un « patch » (il s'agit en fait non d'un patch, mais d'un outil distinct ayant bénéficié d'un développement distinct au sein de notre équipe de développement du *Speech & Natural Language Group* à Redmond).

Dans ces blogs, le nouveau correcteur de Microsoft s'en tirait plus qu'honorablement, les métriques produites par Jean Véronis le plaçant bien au-dessus de la Toolbar de Google, qui se caractérisait par un nombre excessif de fausses alertes (*false flags* ou *false positives*, disent les anglophones). S'agissant du correcteur d'OpenOffice, le score semblait plus serré, probablement parce que l'évaluation effectuée par Jean Véronis avait tenu compte de l'une ou l'autre fausse alerte produite par le nouveau correcteur grammatical de Microsoft, mis lui aussi gratuitement à la disposition des utilisateurs d'Office 2003 depuis quelques mois (l'évaluation semblait aussi plus serrée parce que le texte utilisé ne comportait pas toutes les difficultés de la langue française, notamment les ligatures, que certains correcteurs ne semblent pas accepter, malgré leur présence dans des mots aussi fréquents que *cœur*, *sœur* ou *œuf*). OpenOffice ne disposant pas de correcteur grammatical, les chiffres correspondant au rappel et à la précision en tenant compte du nombre total d'erreurs dans ce petit échantillon étaient assez proches, avec toutefois un avantage global pour le produit de Microsoft (Véronis 2005c).

Même si ce n'était pas le but initial, les évaluations effectuées montraient également que la nouvelle version du correcteur orthographique de Microsoft était très différente de la version précédente. La méthodologie de l'évaluation reposant toutefois sur un texte très court (une page environ) ainsi que sur des erreurs introduites artificiellement dans ce texte, j'ai pensé qu'il serait bon d'approfondir cette question de l'évaluation, en l'abordant par le biais d'une série de questions que se posent les développeurs d'outils de correction linguistique, notamment dans le monde commercial, où les règles sont évidemment différentes de celles qui sous-tendent le développement de prototypes tels que ceux que l'on trouve dans les laboratoires universitaires. Voici quelques-unes de ces questions :

- Quels sont les critères permettant de déterminer qu'une nouvelle version d'un outil est nécessaire ?
- Quelle méthodologie doit-on mettre en œuvre pour évaluer la qualité d'un outil et déterminer qu'il a atteint un stade auquel il peut être diffusé auprès de dizaines de millions d'utilisateurs aux profils et aux exigences très différents ?[1]

Il va de soi qu'un texte d'une page est nettement insuffisant pour se faire une idée correcte de la qualité d'un outil linguistique. Jean Véronis en est parfaitement conscient (il le confirme à plusieurs reprises sur son blog). Une véritable évaluation telle que nous les menons au sein de notre groupe nécessite une équipe de plusieurs personnes compilant des batteries de tests de millions de phrases tirées de textes authentiques. Les textes journalistiques tels que ceux du *Monde* font bien sûr partie des corpus que nous utilisons, mais ils ne représentent qu'une partie des types de données auxquelles nous avons recours. Plus fondamentalement, nous utilisons deux grandes catégories de données linguistiques :

1 Le correcteur grammatical sorti en novembre 2004 a été installé plus de 9 millions de fois et le correcteur orthographique sorti en avril 2005 plus de 10 millions de fois entre le moment de leur sortie et leur intégration dans le Service Pack 2 d'Office 2003 en septembre 2005.

1. Des corpus de textes « édités » : ces textes, de bonne qualité, ne comportent normalement pas de fautes (ou ont été corrigés). Ils proviennent de magazines ou de journaux de qualité, mais aussi de romans, de publications scientifiques, de rapports en tous genres, ainsi que d'articles tirés par exemple d'encyclopédies (notamment Encarta). Ces textes sont nécessaires pour vérifier que les outils linguistiques en développement ne soulignent pas à tort des mots ou des tournures correctes (les fausses alertes ennuyant les utilisateurs, parfois au point de les convaincre de débrancher leur correcteur si elles sont trop fréquentes, comme nous y reviendrons dans quelques paragraphes).
2. Des corpus de textes « non-édités ». Ces textes sont souvent plus difficiles à obtenir. Il peut s'agir de textes provenant de courriers électroniques, de newsgroups, de lettres, de textes obtenus en privant les rédacteurs de tout correcteur linguistique, etc. Nous insistons beaucoup sur l'authenticité des textes et donc des erreurs que nous récoltons. Cela ne signifie pas que nous n'utilisons pas des outils proches du « pourrisseur de textes » de Jean Véronis. Nous les limitons toutefois au développement de batteries de tests très ciblés destinées à guider les développeurs dans leur travail et à tester nos outils pour certains phénomènes très spécifiques. Le profil de l'utilisateur principal de Word étant le « travailleur de la connaissance » (*knowledge worker*), nous mettons aussi l'accent sur l'acquisition de phrases réelles contenant des erreurs réelles commises par ce type d'utilisateur. Pour ce faire, nous avons établi des sites destinés spécifiquement à la collecte de ce type de documents, où ces travailleurs de la connaissance sont invités à créer des documents à l'aide d'une version modifiée de Word ne leur permettant pas de repositionner le curseur, ce qui préserve les erreurs dans les documents.

Tout ceci explique que notre groupe est devenu un gros consommateur de données textuelles, lexicales et acoustiques, acquises par notre *Microsoft Language Resources Center*. Nous sommes ainsi membres de divers consortiums tels qu'*ELRA* (l'Agence Européenne pour les Ressources Linguistiques), l'*ANC* (American National Corpus) ou le *LDC* (*Linguistic Data Consortium*), auxquels nous achetons très régulièrement de grandes quantités de données (ces acquisitions représentent une part significative des investissements que nous consentons en matière de technologie linguistique). Notre bibliothèque de corpus contient, pour un nombre sans cesse croissant de langues, des milliards de mots auxquels nos développeurs ainsi que nos testeurs ont accès pour effectuer la nécessaire veille lexicale basée sur des textes authentiques et récents. Il est crucial de mettre les corpus à jour afin de tenir compte de l'évolution de la langue, comme nous y reviendrons plus bas dans cet article.

C'est la deuxième catégorie de corpus, celle des textes « non édités », qui nous permet de dresser une typologie des erreurs les plus fréquentes. Cette typologie guide notre travail, puisque nos outils sont destinés à répondre à des besoins bien précis. Dans le cadre d'un correcteur grammatical, par exemple, il est nécessaire de se rendre compte que l'accord de l'adjectif pose plus fréquemment des problèmes que les tournures comparatives fautives telles que celles que l'on retrouve dans « Mon gâteau est plus bon que le tien », très rares chez les locuteurs natifs. Comme le développement des outils de correction linguistique s'apparente souvent à une forme moderne du mythe de Sisyphe et que les impératifs économiques ne nous permettent pas de travailler pendant vingt ans sur la question des comparatifs (pour ne citer que cet exemple), ces typologies nous aident à sérier les problèmes et à apporter les solutions qui seront les plus utiles pour nos utilisateurs (puisque l'on sait que la gamme des erreurs possibles est infinie). Pour le correcteur grammatical, par exemple, le profil de l'utilisateur-type peut être défini comme suit, pour paraphraser Riley *et al.* (2004) :

- Travailleur de la connaissance possédant la compétence d'un locuteur natif ou quasi-natif
- Possédant une compréhension moyenne des questions d'ordre grammatical (l'utilisateur type n'est donc pas linguiste)
- Désirant obtenir des suggestions de la part d'un système d'aide à l'écriture
- Écrivant souvent dans un contexte « commercial » à propos de sujets professionnels ou commerciaux
- Utilisant le correcteur orthographique avant de soumettre son texte au correcteur grammatical
- Corrigeant le document en plusieurs phases, entre le brouillon et la version finale
- Ayant des attentes différentes selon que les fonctionnalités sont offertes par défaut ou sont facultatives

Revenons-en au correcteur orthographique. Comme je le signalais plus haut, une des questions cruciales est de savoir quand une nouvelle version doit être développée. Il va de soi que la langue évolue, nous le savons bien. De nouveaux mots apparaissent tous les jours, d'autres tombent en désuétude. Si les mots *blog* ou *altermondialisation* (et *altermondialiste*) étaient encore soulignés par le correcteur orthographique, il est probable que nos utilisateurs estimeraient que l'outil n'est pas à jour. Ce qui était pardonnable il y a trois ans nous serait immanquablement reproché aujourd'hui. Dans ce cas de figure, nous avons d'ailleurs fait le même choix que d'autres lexicographes, comme par exemple ceux de Larousse qui ont, comme nous, choisi de considérer le mot *blog* comme faisant partie du stock lexical français en 2005, allant ainsi à l'encontre des recommandations de la *Commission générale de néologie et de terminologie*, qui préconisait, en mai 2005, soit quelques semaines à peine après la sortie de notre correcteur, l'utilisation du mot *bloc-notes* en lieu et place de *blog*, considéré comme un anglicisme malvenu. Nous avons préféré agir en lexicographes descriptifs et ainsi rendre compte de l'évolution de notre langue ainsi que d'une indéniable assimilation de ce terme. Les millions d'occurrences du mot *blog* (et de ses dérivés tels que *blogosphère*) sur la Toile montrent qu'il est vain de nier l'évidence. Les décrets ont leurs limites et les millions d'utilisateurs de nos logiciels comprennent qu'il était de notre devoir de consigner cette évolution dans notre correcteur.

Les néologismes tels que *blog*, *blogosphère*, *cybercriminalité, cybercommerce, cyberconsommateur, altermondialisation, antimondialisation, comitologie, Eurocorps* ou *bioluminescent* ne sont pas les seules sources d'ajout. L'actualité joue également un rôle. Il n'est bien sûr pas toujours possible de faire en sorte que les noms des ministres d'un gouvernement soient directement intégrés au lexique d'un correcteur. La notoriété est parfois de courte durée, mais les utilisateurs s'attendent à ce que les noms propres familiers soient reconnus et comptent sur le correcteur pour leur confirmer (ou leur rappeler) que *Raffarin* prend deux « f », que *de Villepin* ne s'écrit pas avec -ain comme *pain*, que *Sarkozy* prend un -y et pas un -i, que le nom de l'actuel premier ministre belge *Guy Verhofstadt* s'écrit avec -dt ou qu'une certaine circulaire européenne qui fait couler beaucoup d'encre est à attribuer à un certain *Bolkestein* et pas à *Bolkenstein*, graphie que le nouveau correcteur soulignera.

Dans sa première évaluation du correcteur de Microsoft par rapport à la Toolbar de Google, Jean Véronis (2005a) avait intentionnellement ignoré les noms propres. Dans un commentaire posté sur son blog, j'avais fait remarquer qu'ils devraient à mon avis être pris en compte lors du calcul de la précision et du rappel, qui utilisent la notion de bruit. Lors de la seconde évaluation (Véronis 2005b), tenant compte cette fois des noms propres, les chiffres révélaient

une différence énorme quant au « bruit », c'est-à-dire la proportion de fausses alertes des deux correcteurs. Même sur un échantillon aussi petit que le texte du *Monde*, on notait 9,3 % seulement de bruit pour le nouveau correcteur de Microsoft, contre 34,7 % pour la Toolbar de Google. Là où Google montre que son lexique d'entités géographiques et de noms de personnes est très pauvre (Jean Véronis notait que *Londres, Madrid, New York, Moscou, Singapour* et *Chirac*… sont considérés par le correcteur de Google comme des fautes au même titre que *Londre* et *Chriac*), le correcteur de MS Word ne soulignera que *Londre* et *Chriac* (à juste titre), tout en proposant les bonnes versions de ces noms propres. Cet aspect est également important pour nos utilisateurs, qui ne se préoccupent guère de la nature du mot. Qu'il s'agisse d'un nom commun ou d'un nom propre, l'important pour eux est de laisser le moins de fautes possibles dans leur texte et un correcteur qui leur indique que le « s » manque à *Londre* leur rend un réel service. Si l'on prend l'habitude d'ignorer les soulignements rouges parce que ce sont des noms propres, on n'est plus en mesure de faire la différence entre une faute véritable et un soulignement intempestif dû à la pauvreté du lexique. On comprend dès lors pourquoi nos lexiques comprennent aussi une forte proportion de noms propres. Nos synergies avec nos collègues de l'encyclopédie Encarta nous ont permis de bénéficier de leurs bases de données de termes géographiques et de personnages célèbres pour enrichir notre dictionnaire et ainsi réduire le bruit[2]. Des expériences effectuées à partir de divers textes utilisés dans des concours de dictées (Dictée de Pivot en France, Dictée du Balfroid en Belgique) ont montré que, là aussi, le correcteur de Microsoft se démarquait par rapport à d'autres outils linguistiques assez pauvres en matière d'entités nommées et de noms propres. Certains objectent que l'utilisateur peut enrichir son dictionnaire personnel en y ajoutant sa propre terminologie. C'est évidemment tout à fait vrai et les correcteurs de Microsoft Office offrent aussi cette fonctionnalité, très utile au demeurant. On ne peut toutefois se retrancher derrière cet argument pour cacher le problème de l'utilité toute relative d'un correcteur qui n'offrirait qu'une coquille presque vide à ses utilisateurs et attendrait d'eux qu'ils « nourrissent » eux-mêmes le lexique de leur outil. Ne s'improvise pas lexicographe qui veut… Même avec un outil performant et une bonne couverture lexicale, un correcteur reste un outil imparfait dont l'utilisation réclame une certaine dose de bon sens et une attitude critique. On n'ose donc imaginer les dégâts que peut occasionner un outil au lexique squelettique, particulièrement dans le domaine de l'enseignement, où il n'est bien sûr pas question de demander aux élèves de compiler eux-mêmes leur dictionnaire.

3. Évolution de la langue

3.1. Réforme de l'orthographe

Un facteur crucial qui a présidé à la décision de développer un nouveau correcteur orthographique pour le français fut la question de la réforme de l'orthographe. Sujet sensible s'il en est, cette réforme a fait couler beaucoup d'encre depuis la publication en 1990 des recommandations formulées par l'Académie française et le Conseil supérieur de la langue française. Ignorée initialement, accueillie avec enthousiasme par certains, comme par exemple Claude Hagège, Bernard Quémada, André Goosse ou Daniel Blampain, pour ne citer que quelques-uns de ses ardents supporters, avec scepticisme par d'autres, cette réforme a mis du temps à se faire accepter. Pourtant, elle est de plus en plus souvent utilisée et de nombreuses instances officielles l'ont recommandée. En 1998, le Ministère de la Communauté française

[2] Il ne faut toutefois pas perdre de vue que l'ajout de noms propres doit lui aussi suivre les mêmes règles que celles qui régissent l'acquisition lexicale des mots courants. Il serait vain (et dangereux) de tenter d'inclure dans le lexique tous les noms propres possibles. Cela conduirait immanquablement à masquer de réelles erreurs. Le jugement du lexicographe et les critères traditionnels de fréquence sont donc toujours de mise.

de Belgique, en charge de l'éducation dans la partie francophone du pays, recommandait l'enseignement et l'application des nouvelles règles orthographiques dans tous les réseaux d'enseignement du pays.

La circulaire ministérielle du 30/03/1998 précisait :

> (…) Quoi qu'il en soit, il n'est certainement pas recommandé d'imposer une, et une seule orthographe. Chacun a le droit d'utiliser les différentes graphies. Il s'ensuit que, durant une période de durée indéterminée, les deux orthographes auront à coexister et seront acceptées. En conséquent, lors des contrôles, les deux orthographes seront admises.

Plus près de nous, en mars 2005, le ministère québécois de l'éducation nationale recommandait lui aussi l'adoption de cette nouvelle orthographe et son enseignement systématique, tout en précisant également que les deux orthographes devaient être considérées comme correctes, conformément aux recommandations de l'Académie. Quelques jours plus tard, la revue *Forum* de l'Université de Montréal annonçait qu'elle passait à la « nouvelle » orthographe. En France, depuis 2002, les dictionnaires Hachette accordent une place prépondérante à cette nouvelle orthographe. Le dictionnaire Le Robert ne l'ignore plus non plus, la mentionnant de plus en plus souvent dans le corps des articles en la faisant précéder d'une étiquette quelque peu sibylline : « On écrirait mieux … ». Le Petit Larousse a également intégré bon nombre de ces modifications dans ses éditions les plus récentes. Sans le savoir, de nombreux journalistes ont commencé à l'utiliser (on trouve ainsi depuis plusieurs années de nombreuses références à des *matchs*, plutôt que des *matches*, des *sandwichs* plutôt que des *sandwiches*, ou des *crashs* plutôt que des *crashes*). Ne parlons pas *d'évènement*, qui semble à présent bien installé. Le Bescherelle lui-même (tout comme le Petit Larousse) s'est mis à mentionner les modifications de l'orthographe lorsqu'elles concernent les conjugaisons, comme par exemple l'utilisation de l'accent grave en lieu et place de l'accent aigu dans les formes du futur et du conditionnel de verbes conjugués sur le paradigme de *céder*, *opérer* ou *gérer* (il *cèdera* au lieu de il *cédera*, nous *opèrerions*, au lieu de nous *opérerions*, tu *gèreras* au lieu de tu *géreras*). Ces dernières rectifications ne font finalement qu'entériner un changement de prononciation.

La réforme de l'orthographe a bien sûr ses militants. Ces derniers affichent clairement leur position et, de plus en plus souvent, on trouve des ouvrages rédigés selon les nouvelles règles. Ainsi, Jean-Marie Klinkenberg et Benoît Denis n'hésitent pas à préciser dans l'introduction de leur ouvrage sur *La littérature belge* : « Le présent livre fait usage des rectifications orthographiques proposées par le Conseil supérieur de la langue française (1990) et approuvées par toutes les instances francophones compétentes, dont l'Académie française ». (p.16)

Des sites Web donnent des informations souvent très détaillées sur la réforme de l'orthographe, sur son histoire ou les mots les plus fréquents qui sont touchés. Parmi eux, on trouve le site « orthographe recommandée » (www.orthographe-recommandee.info) mis en place par le groupe de modernisation de la langue française, qui siège à Paris et comprend des représentants de plusieurs pays francophones. Nous avons pu collaborer avec les linguistes responsables de ce site, qui ont testé notre correcteur et lui ont décerné en 2005 un label de qualité réservé aux outils linguistiques qui ont correctement mis en œuvre les recommandations de cette réforme. Ici aussi, nous avons répondu à la demande d'un nombre croissant d'enseignants qui souhaitaient disposer d'un outil pouvant être utilisé pour enseigner la nouvelle orthographe. À partir du moment où de plus en plus d'organes officiels en matière d'éducation insistaient sur le fait que les deux types de graphies (ancienne et nouvelle)

doivent être considérées comme valables, il devenait urgent de mettre à la disposition des enseignants et des étudiants un outil leur permettant de suivre ces recommandations. C'est maintenant chose faite et les réactions très positives du monde enseignant nous ont montré que les décisions que nous avions prises quant aux options par défaut étaient les bonnes.

L'option par défaut à laquelle est confronté l'utilisateur lors de l'installation du nouveau correcteur est effectivement celle qui considère que l'ancienne orthographe et la nouvelle orthographe sont valables. Que l'on dise « *il gèrera* » ou « *il gérera* », « *il parait* » sans accent circonflexe ou « *il paraît* » avec accent, « *il amoncèle* » ou « *il amoncelle* », le correcteur acceptera donc les deux formes sans broncher, puisque telle a été la volonté des organismes officiels. Dans son évaluation du correcteur face à celui d'OpenOffice, Jean Véronis notait que ce dernier n'intègre pas du tout la réforme de l'orthographe. Il notait qu'ajouter les quelques centaines de mots touchés par ces rectifications ne serait pourtant pas très difficile. S'il ne s'agissait que d'ajouter ces formes, ce serait probablement une tâche aisée et assez rapide. Notre nouveau correcteur intègre les formes simples des quelques milliers de mots touchés par la réforme, ainsi que les formes fléchies, ce qui représente tout de même plus de 20.000 nouvelles formes. Toutefois, la valeur ajoutée du nouveau correcteur ne se limite pas à l'intégration de ces nouvelles formes. Les trois options offertes par le correcteur permettent également une souplesse répondant aux souhaits de nombreux utilisateurs. Certains ont effectivement émis le souhait de ne plus écrire qu'en respectant les règles de la nouvelle orthographe, ce qui est possible en choisissant une option appelée « orthographe rectifiée » par le biais d'une simple boîte (ou boite) de dialogue, reproduite ci-dessous. D'autres utilisateurs ont souhaité disposer d'un outil leur permettant de ne considérer que l'ancienne orthographe comme valable, ce qui permet une utilisation pédagogique de l'outil, puisqu'un même texte peut être soumis au correcteur en appliquant des options différentes, ce qui met en évidence les formes différentes au travers du souligné rouge qui apparait sous certaines conditions uniquement.

Figure 1. Boite de dialogue des paramètres du correcteur orthographique de Microsoft Office 2003

Comme on le voit, cette grande souplesse d'utilisation satisfait tous les publics, mais elle a compliqué la tâche du lexicographe, puisque ce qui est correct selon une option peut être souligné selon une autre option. La gestion du dictionnaire s'en est trouvée compliquée puisqu'il a fallu marquer les formes touchées par les rectifications et coder de façon très précise et systématique plusieurs dizaines de milliers de formes pour indiquer dans quels contextes elles étaient valables.

Prenons quelques exemples (*maitre/maître, whisky,* les verbes *gérer* et *amonceler)* pour illustrer les trois cas de figure et la façon dont ils sont codés dans le lexique. Un système basé sur trois codes permet de représenter les formes comme suit :

PRE : forme valable en mode pré-réforme (= ancienne orthographe)
POST : forme valable en mode post-réforme (= nouvelle orthographe)
NORMAL : forme valable dans tous les cas de figure (ancienne et nouvelle orthographe)

Il va de soi que la toute grande majorité du vocabulaire français n'est pas touchée par cette réforme. Toutes les formes non concernées par la réforme seront donc codées NORMAL. On notera toutefois que les codes ne peuvent être attribués à un lemme (forme de base): l'attribution des codes se fait au niveau des formes fléchies. Pour une paire de mots telles que *maitre* et *maître*, il serait concevable d'assigner les codes au lemme (le lemme *maitre* et ses formes fléchies seront codés POST alors que le lemme *maître* sera codé PRE). Cela n'est toutefois pas possible pour les variations ne concernant que certaines formes fléchies pour un paradigme donné. Ainsi, pour le verbe *gérer* (ainsi que pour *opérer, posséder*, etc.), la majorité des formes fléchies seront codées NORMAL (valables dans tous les cas de figure), puisque la réforme n'a concerné que les formes du futur et du conditionnel, comme on l'a rappelé plus haut. On aura donc une représentation telle que la suivante :

Forme	**Codage**	**Signification du codage**
maître	PRE	uniquement pré-réforme
maîtres	PRE	uniquement pré-réforme
maitre	POST	uniquement post-réforme
maitres	POST	uniquement post-réforme
gérer	NORMAL	toujours valable
gère	NORMAL	toujours valable
gèrent	NORMAL	toujours valable
gérerai	PRE	uniquement pré-réforme
gèrerai	POST	uniquement post-réforme
géré	NORMAL	toujours valable
gérerions	PRE	uniquement pré-réforme
gèrerions	POST	uniquement post-réforme
amoncèle	POST	uniquement post-réforme
amoncelle	PRE	uniquement pré-réforme
whisky	NORMAL	toujours valable
whiskies	PRE	uniquement pré-réforme
whiskys	POST	uniquement post-réforme

Tableau 1. Pré-réforme vs. post-réforme : codage du dictionnaire

Outre le fait que le dictionnaire a dû faire l'objet de codages minutieux, il a fallu également créer de nouveaux paradigmes (classes morphologiques) pour générer automatiquement les formes nouvelles, tout en continuant à générer les anciennes. Les nouvelles formes du futur ou du conditionnel des verbes tels que *gérer, espérer* ou *céder* font effectivement appel au même type de fonction morphologique que les verbes comme *mener* (avec utilisation d'un accent grave – *cèdera*, *mènera*), alors que l'accent aigu est conservé dans l'ancienne orthographe.

On est donc bien loin d'un simple ajout de quelques centaines de mots au dictionnaire, qui aurait probablement eu une certaine utilité, mais aurait laissé bon nombre d'utilisateurs sur

leur faim et ne permettrait pas l'évolution future de l'outil. Si l'on s'aperçoit en effet dans quelques années que la nouvelle orthographe a totalement supplanté l'ancienne, il n'est pas exclu de modifier le paramètre par défaut pour ne plus accepter que la nouvelle orthographe, en soulignant les anciennes formes.

3.2. Féminisation des noms de métiers

Un autre aspect sur lequel a porté l'amélioration du correcteur est celui qui concerne la féminisation des noms de métiers. Ici encore, la frontière entre le prescriptif et le descriptif en matière de choix lexicographique n'est pas facile à déterminer. D'une part, il convient de reconnaitre que la féminisation des noms de métiers a fait l'objet de décrets tout à fait officiels dans des pays comme la Belgique, où l'on recommande depuis des années l'emploi de formes féminines comme *échevine*, *écrivaine*, *auteure*, *officière, députée, sénatrice* ou *soldate*. Comme le note le *Guide de féminisation des noms de métier, fonction, grade ou titre* dans sa 2^{e}édition (2005, p. 6), la publication en 1993 du décret relatif à la féminisation des noms de métiers, fonction, grade ou titre a été suivie d'une période d'hésitation compréhensible. Pourtant, depuis 1994, l'usage des noms de métier et de fonction au féminin a évolué et des formes féminines telles qu'*auteure, chercheure* ou *écrivaine* s'écrivent couramment en 2006. Le mot *professeure* est connu et utilisé au Québec depuis très longtemps. Ici, il n'était plus question de créer un système d'options tel que celui mis en place pour la réforme de l'orthographe. Les études de corpus ont montré que cette évolution de l'usage était bien réelle et que cette réforme linguistique était couronnée de succès (voir entre autres Dister (2004), ainsi que l'étude de Dister et Moreau (2006) sur les dénominations des candidates dans les élections européennes en France et en Belgique) : en respectant les observations scientifiques quant à l'utilisation de ces formes dans les corpus de textes francophones et en tenant compte des recommandations politico-linguistiques promouvant l'emploi du féminin dans la langue française, nous n'avons fait que répondre aux attentes de nombreux utilisateurs. Ici, encore, cela a signifié un travail lexicographique d'une certaine ampleur pour, comme le notent Dister et Moreau (2006), « élargir le champ d'application des règles morphosyntaxiques, et notamment créer de nouvelles formes là où elles n'existaient pas ». Cette intégration dans le correcteur d'un traitement de texte aussi répandu que Word devrait contribuer à renforcer encore le succès de cette réforme linguistique puisque les formes féminines sont à la fois acceptées par le correcteur et proposées dans les listes de suggestions en cas de faute. Dister et Moreau vont même jusqu'à suggérer que cette diffusion gratuite de ce correcteur par Microsoft intégrant à la fois les recommandations orthographiques de 1990 et les nouvelles formes féminines est un indicateur de l'évolution de la langue que l'on pourra prendre en considération lorsque l'on écrira l'histoire de ces changements linguistiques.

Cette volonté de prendre en compte ces changements explique les investissements consentis pour développer ce nouveau correcteur. Ici encore, toutefois, la comparaison avec d'autres outils concurrents ne peut se faire sur la base de quelques textes ou de quelques pages uniquement. Les critères généralement utilisés pour l'évaluation des performances linguistiques d'un correcteur sont difficilement d'application. Ainsi, la précision et le rappel concernent la qualité des soulignements :

Précision : nombre de soulignements corrects / nombre total de soulignements
Rappel : nombre de soulignements corrects / nombre total d'erreurs dans le texte

Ces types de statistiques sont utilisées par nos équipes de testeurs pour guider les équipes de développement et déterminer le moment où le produit sera considéré comme suffisamment bon pour être mis sur le marché. On notera que les corpus de plusieurs dizaines de millions de

phrases utilisées par nos testeurs sont distincts des corpus utilisés pour le développement. Nous sommes soucieux d'éviter que les développeurs améliorent leurs produits en fonction des données utilisées par les équipes de test. Comme on l'a vu plus haut, les évaluations effectuées par Jean Véronis utilisaient ces modes de calcul pour conclure que le correcteur de Microsoft donnait de meilleurs résultats que les autres produits qu'il évaluait. Des fonctionnalités plus générales telles que l'intégration de ces réformes linguistiques ne peuvent toutefois pas transparaitre lors de ces évaluations statistiques alors qu'elles ont en fait été les raisons premières du développement de notre nouvel outil. La probabilité de trouver un nom de métier au féminin dans un texte de deux ou trois pages est en effet extrêmement faible et le nombre de mots touchés par la réforme de 1990 est aussi trop réduit pour qu'une telle évaluation puisse mettre en évidence la nature et la portée de ces fonctionnalités d'un correcteur. Des tests ciblés sont alors nécessaires pour vérifier que les recommandations et les directives ont été correctement mises en œuvre et que les modules morphologiques sont correctement adaptés pour tenir compte de nouvelles règles morphosyntaxiques. Ces tests et ces développements nécessitent un travail lexicographique et linguistique important assez peu compatible avec la tentation de construire des dictionnaires par des procédures automatiques et quasi uniquement sur des bases statistiques.

4. Correcteur orthographique et séquenceur

Dans cette section, je me propose d'aborder brièvement le rôle que joue le séquenceur dans la correction orthographique. Le terme « séquenceur » est ici utilisé pour désigner l'outil qui détermine les frontières entre les éléments du texte (les anglophones parlent plus volontiers de « *word-breaker* »). En d'autres termes, il permet de déterminer sur quelles unités (quels mots) va porter la correction orthographique. Il s'agit bien sûr d'un outil dépendant de la langue et il serait vain de récupérer un outil ayant été développé par exemple pour l'anglais en pensant pouvoir l'utiliser sans problème en français. Comme l'ont montré Grefenstette et Tapanainen (1994), Grefenstette (1996) ou Mitkov (2003), la segmentation en mots pose des problèmes particuliers, même si les linguistes ont tendance à regarder de haut cet aspect du TALN. Fontenelle (2005a) montre qu'il est nécessaire de déterminer quel est le statut de certains caractères avant de se lancer dans la conception d'un correcteur orthographique. Il est ainsi crucial de décider si l'on va traiter l'apostrophe comme un caractère séparable en français. Tout le monde est évidemment d'accord sur le fait que l'apostrophe est un caractère séparable lorsqu'une chaine de caractères entourée de blancs commence par l'une des séquences suivantes : *l', d', n', c', m', t', s'* et quelques autres. Cela permettra de scinder *l'avion, d'école* ou *qu'ici* en deux mots distincts (*l'*+*avion, d'*+*école, qu'*+*ici*..). On connait les exceptions traditionnellement citées dans les manuels (*aujourd'hui, prud'homme, presqu'ile*…). Il va de soi que l'on ne souhaite pas inclure dans le lexique toutes les formes élidées possibles (*qu'hier, d'avant, l'horloge*…). Cela entrainerait une augmentation du nombre de formes fléchies d'environ un million d'unités. On a donc le choix entre décider de considérer l'apostrophe comme un caractère séparateur par défaut et donner la liste des exceptions (*aujourd'hui*…, auxquelles il faudrait ajouter les noms d'origine irlandaise tels que *O'Neill*, *O'Brien*, etc. ou d'autres irrégularités comme *no man's land*). Mais on peut aussi décider que, par défaut, l'apostrophe sera considérée comme un caractère non-séparateur (*aujourd'hui* serait donc conforme à la « règle ») et les exceptions seront limitées au traitement de chaines commençant par l'une des quelques séquences citées ci-dessus. L'avantage de la deuxième approche est qu'elle permet de garder comme un seul token, c'est-à-dire de considérer comme un seul mot, des séquences comportant des fautes d'orthographe telles que *Ren'e* ou *Herv'e*, où l'utilisateur a remplacé l'accent aigu par une apostrophe, ce que certains font lorsqu'ils

cherchent à indiquer que le mot possède un accent qu'ils ne trouvent pas sur un clavier non français.

Le trait d'union pose lui aussi des problèmes spécifiques au français et, ici encore, la décision de le considérer ou pas comme un caractère séparable sera lourde de conséquences quant à la nature des mots à inclure dans le lexique. Ainsi, la version précédente du correcteur de Word considérait le trait d'union comme un caractère non séparable. Cette approche est évidemment compatible avec le nombre assez important (plusieurs milliers) de mots comportant un trait d'union que l'on trouve dans un dictionnaire traditionnel (*porte-avions, cul-de-sac, brise-glace, chef-d'œuvre*, etc.). Le gros désavantage est toutefois que les combinaisons productives non lexicalisées étaient soulignées dans la version précédente, alors qu'elles étaient parfaitement correctes. Ainsi, des références au gouvernement *Chirac-Raffarin*, à l'axe *France-Allemagne*, au match *Espagne-Danemark*, au vol *Bruxelles-Barcelone*, aux relations *employeurs-syndicats* ou à un congé *épargne-temps* étaient impitoyablement rejetées par le correcteur qui considérait que la chaine de caractères *France-Allemagne* était un seul mot, que son correcteur ne possédait évidemment pas (on n'ose imaginer le cauchemar du lexicographe à qui on demanderait d'ajouter toutes les paires possibles : France-Italie, France-Belgique, France-Grèce, etc.). Les fausses alertes irritant l'utilisateur (qui, si elles étaient trop nombreuses, en arrivait à ne plus regarder ces soulignés rouges et donc à ne pas voir la différence entre une fausse alerte dans *Chirac-Raffarin* et une alerte réelle dans *Chiarc-Raffarin*), il était nécessaire de revoir la conception du séquenceur qui se trouvait à la base du correcteur. En modifiant la nature du trait d'union et en le considérant comme un caractère séparable, nous sommes parvenus à réduire de 74 % le nombre de fausses alertes par rapport à la version précédente. Ce pourcentage impressionnant recouvre bien sûr la disparition de fausses alertes suite à la lexicalisation d'entités nommées provenant des bases de données (personnages célèbres, termes géographiques) de nos collègues d'Encarta, de la veille lexicographique basée sur les corpus dont j'ai parlé plus haut, mais aussi de cette modification de notre séquenceur qui nous permet d'accepter des combinaisons telles que *Chirac-Raffarin*, *Villepin-Sarkozy* ou *Belgique-Angleterre, aide-cuisinière* et *épargne-logement,* puisque le correcteur va traiter séparément les chaines de caractères de part et d'autre du trait d'union. Dans la foulée, le correcteur ne souligne plus non plus des termes tout à fait corrects tels que *demi-million, demi-milliard* ou *demi-millier* (qui n'étaient pas dans le lexique du correcteur précédent, alors que *demi-douzaine* était accepté).

Cette décision a bien sûr forcé les linguistes à mettre en œuvre des moyens distincts pour continuer à souligner des fautes (bien réelles celles-là) de nature presque grammaticale très fréquentes dans les mots composés. On sait effectivement que de nombreux locuteurs ajoutent erronément un « s » à la première partie d'un mot composé, même si celle-ci est invariable lorsqu'elle provient d'un élément verbal. On connait l'exemple de *tires-fesses, portes-avions, attrapes-mouches* ou *lances-missiles*. Il va de soi que, si l'on considère le trait d'union comme un caractère séparable, on court le risque d'appliquer le correcteur à *tires, portes, attrapes* ou *lances* indépendamment de leurs contextes, auquel cas on devrait les accepter. Nous avons dès lors mis en place des mécanismes particuliers pour faire en sorte que toutes ces erreurs continuent bien évidemment d'être corrigées. En ce sens, cette stratégie démarque notre correcteur des autres produits concurrents qui soit considèrent le trait d'union comme un caractère non-séparable et soulignent alors tous les composés productifs (nombreux et corrects) tels que *France-Allemagne* ou *épargne-logement*, soit considèrent ce signe comme un caractère séparable et ne sont alors plus capables de détecter des fautes réelles telles que *portes-avions*.

5. Améliorations des suggestions

La fonction première d'un correcteur orthographique est de déceler les fautes. Mais il s'agit aussi de proposer des suggestions. Le correcteur n'a généralement pas conscience du contexte et il est dès lors essentiel d'essayer d'identifier au mieux ce que l'utilisateur avait réellement l'intention d'écrire. Les suggestions sont généralement basées sur le concept de distance d'édition, qui calcule le nombre de manipulations nécessaires pour retrouver, à partir d'un input erroné, une chaine de caractères dans le dictionnaire qu'il s'agit de proposer à l'utilisateur. Parmi ces manipulations on trouvera notamment :

Suppression d'un caractère :	débonaire → débonnaire
Ajout d'un caractère :	mourrir → mourir
Permutation :	Chiarc → Chirac
Remplacement :	indépendemment → indépendamment

Outre ces manipulations standards, que l'on retrouve à la base de tous les correcteurs orthographiques, toutes langues confondues, on trouvera d'autres types de manipulations basées sur une typologie des erreurs les plus fréquentes pour une langue donnée. Ainsi, pour certaines langues (y compris le français), on note une propension chez certains utilisateurs à écrire le son /f/ phonétiquement, même si la graphie est normalement [ph]. Pour une erreur telle qu'*éléfant*, il s'agit donc de proposer *éléphant* à l'utilisateur, même si, en termes de distance d'édition, on se retrouve finalement assez loin de l'entrée initiale (trois manipulations ayant été nécessaires pour passer de l'entrée correcte à l'erreur: suppression du *p* et suppression du *h,* ajout du *f*). De telles transitions sont spécifiques à certaines langues et doivent permettre d'accorder une pondération plus importante à certains types d'erreurs. En raffinant ces transitions spécifiques sur la base de notre typologie d'erreurs réelles, nous avons augmenté le pourcentage de cas où la bonne forme apparait en première position dans la liste de suggestions.

Les dictionnaires de nos correcteurs contiennent des informations concernant la partie du discours. Ces dictionnaires étant utilisés dans d'autres applications (par exemple dans des séquenceurs-lemmatiseurs (« word-breakers ») pour le Microsoft SharePoint Portal Server, afin de faciliter la recherche d'information par le biais de l'expansion de requêtes), ils contiennent également des informations grammaticales supplémentaires. Pour une forme verbale donnée, le correcteur sait par exemple qu'il s'agit d'une première personne du pluriel. Cette intégration d'éléments de nature grammaticale dans le correcteur orthographique nous a servi à améliorer le mécanisme des suggestions. Considérons par exemple la phrase suivante :

Mon fils sentend bien avec ses voisins.

La chaine *sentend* est soulignée par le correcteur et la forme correcte *s'entend* apparait en première position dans la liste des suggestions. Notre typologie d'erreurs nous a en effet révélé qu'il était assez fréquent d'omettre l'apostrophe et donc de procéder à une concaténation de deux tokens. La réinsertion de l'apostrophe pose toutefois des problèmes puisque certaines conditions doivent être réunies. La chaine erronée doit évidemment commencer par un des mots (un article, un pronom, une conjonction…) qui déclenchent une élision en français (*d, qu, s, n, t, m, l, puisqu*..). Il s'agit ensuite de n'insérer cette apostrophe que si l'on trouve ensuite une voyelle. Cela permettra de ne pas tenter l'insertion d'une apostrophe dans la phrase suivante, où *straitent* est souligné :

Il straitent leurs élèves méchamment.

Comme l'insertion de l'apostrophe ne sera pas tentée, d'autres manipulations seront effectuées et, parmi celles-ci, le mot « *traitent* » sera proposé (en première position car la

distance qui sépare *straitent* et le mot *traitent* dans le lexique ne représente qu'une seule manipulation). Ce sera ensuite au correcteur grammatical de prendre la relève pour détecter le problème d'accord dans « *il traitent* » et corriger la faute en « *ils traitent* ».

Profitant de l'inclusion d'informations grammaticales dans le dictionnaire, nous avons mis en œuvre une série de contraintes d'ordre phono-syntaxiques permettant par exemple de ne permettre l'insertion d'une apostrophe que si la chaine de caractères suivant un *s'* correspond à une entrée verbale à la 3e personne (comme *il s'imagine, ils s'aiment*…). Ces informations permettent ainsi de bloquer l'insertion d'une apostrophe dans un exemple tel que le suivant :

Paul n'aime pas le senfants.

Il va de soi que proposer *s'enfants* semblerait absurde et ces contraintes permettent d'éviter la suggestion de formes grammaticalement incompatibles. En modélisant les contraintes et la compatibilité des formes élidées avec les formes pouvant les accompagner, nous sommes parvenus à réduire les suggestions totalement impossibles (en précisant que le pronom élidé *j'* ne peut être combiné qu'avec des verbes à la première personne du singulier, on évite ainsi de suggérer *j'épis* lorsqu'un utilisateur écrit *jépis* à la place de *képis*, le *j* et le *k* étant voisins sur le clavier ; comme *épis* n'est présent dans le lexique que comme nom, il est incompatible avec *j'*). Bien sûr, les informations grammaticales sont trop limitées et la sémantique manque pour ne proposer que des formes totalement compatibles dans le contexte de ce que l'utilisateur a réellement voulu dire, mais ce genre de modélisation nous a prouvé que la frontière entre le correcteur orthographique et le correcteur grammatical pouvait être franchie en exploitant une information lexicale plus riche au sein même du correcteur orthographique (voir Fontenelle 2005b pour plus d'informations sur ces contraintes).

6. Un correcteur grammatical amélioré

La première version du correcteur grammatical français développée en interne par notre groupe fut mise à la disposition des utilisateurs d'Office XP en 2001. Nous avions alors décidé de développer ces produits en interne pour diverses langues (l'espagnol, le français, l'allemand et l'anglais) parce que de nombreuses réactions recueillies auprès de nos utilisateurs indiquaient que les correcteurs précédents considéraient dans de trop nombreux cas des phrases correctes comme fautives, ce qui était gênant et peu utile. Nos recherches montraient alors que beaucoup d'utilisateurs débranchaient ces correcteurs grammaticaux à cause des trop nombreuses fausses alertes. Notre but, en remplaçant ces produits, était de fournir des outils plus discrets et plus fiables quant aux erreurs qu'ils étaient en mesure de repérer, et donc moins gênants tout en étant plus utiles. Ce faisant, ces outils ne repéraient pas un certain nombre d'erreurs réelles. On voit donc la tension entre le désir de repérer le plus grand nombre de fautes possibles (avoir un rappel élevé) et de ne pas se tromper en signalant erronément une faute (avoir une précision élevée).

C'est la raison pour laquelle nous avons entrepris de développer une version améliorée qui a été gratuitement mise à la disposition des utilisateurs d'Office 2003 en novembre 2004, via une mise à jour publique (par téléchargement volontaire) et ensuite intégrée au Service Pack 2 d'Office 2003 en septembre 2005. Cette nouvelle version était clairement développée dans le but d'étendre la couverture des fautes repérées tout en augmentant la précision de façon significative.

Ici encore, comme nous l'avons indiqué plus haut, nous sommes partis de nos observations faites sur de gros corpus contenant des erreurs réelles, ainsi que des réactions de nos utilisateurs. Les améliorations ont notamment porté sur les concepts grammaticaux suivants.

La phrase servant d'exemple illustre des fautes réelles signalées par le correcteur grammatical, le souligné reproduisant le trait ondulé vert typique de cet outil.

Type d'erreur	Exemple	Correction proposée
Accord sujet-verbe	Il conduis sa voiture.	conduit
Accord des temps et des modes	Si j'aurais su, je n'aurais pas acheté ce livre.	j'avais
a/à	Il habite a Singapour.	à
Accord dans le groupe nominal	Vous devez souligner les réponse incorrecte.	la réponse incorrecte les réponses incorrectes
Trait d'union	Donnez moi 5 minutes !	Donnez-moi
Expressions à mots multiples	As-tu téléchargé les deux mises à jours ?	jour
Impératif	Manges tes pommes de terre !	Mange
Participe passé/Infinitif	Mon fils a trouver un billet de 100 euros.	trouvé
	J'ai fait fonctionné mon PC sans aucun problème.	fonctionner
Accord du participe passé	Ces mesures ont été introduite en 2001.	introduites
Participe passé et verbes pronominaux	Elle s'est foulée la cheville en descendant les escaliers.	foulé
Auxiliaire	Je m'ai trompé sur toute la ligne.	me suis

Tableau 2. Exemples de fautes corrigées par le correcteur grammatical

Il ne s'agit bien sûr que d'un échantillon. Pour chacune de ces règles, nous calculons la précision et le rappel. Ainsi, nous pouvons déterminer par exemple que telle règle X a une précision de 80 %. Cela signifie donc que sur 10 erreurs signalées, 8 sont correctes et donc 2 sont des fausses alertes. Si le rappel de cette règle est de 60 %, cela signifie que sur 10 erreurs réelles, l'outil en repère 6. Toutefois, ces chiffres de précision et de rappel ne disent rien quant à l'utilité de l'outil. Une précision et un rappel très élevés peuvent s'avérer inutiles si le type d'erreur est peu fréquent. Il se peut donc, comme le notent Riley *et al.* (2004) qu'une règle de 90 % de précision et de 90 % de rappel soit moins utile qu'une autre règle se déclenchant plus fréquemment dont la précision serait seulement de 80 % pour un rappel identique (90 %). L'impact sur l'utilisateur doit donc prendre la fréquence des erreurs en compte. Ainsi, un outil qui permet de corriger une expression telle que « *en terme de* » où le *s* de *termes*, obligatoire dans cette tournure, est oublié, sera utile aux milliers de personnes qui chaque jour postent sur la Toile un document contenant cette forme erronée.

Par rapport à la version précédente de notre correcteur, nous avons réduit le nombre de fausses alertes de 40 % tout en augmentant le nombre de fautes réelles correctement signalées de près de 50 %. Nous avons donc opté pour un produit de grande précision, ce qui signifie qu'il n'est pas en mesure de trouver toutes les erreurs, mais que, lorsqu'il signale une erreur, la probabilité qu'il se trompe est très réduite, contrairement à d'autres produits du marché qui détectent parfois un peu plus d'erreurs, mais peuvent se tromper jusqu'à 25 fois plus souvent, comme l'a montré une évaluation indépendante effectuée sur un corpus de 1000 pages.

Une des innovations de ce correcteur est sa capacité à détecter certains types de fautes dans des phrases inanalysables bien que courtes ou dans de très longues phrases, même si l'analyseur syntaxique n'a pas été en mesure de fournir une analyse complète, vu la

complexité de cette phrase. Ce travail sur l'analyseur, dû à mon collègue Thierry Etchegoyhen, a permis d'augmenter de façon significative le nombre d'erreurs détectées, surtout dans le domaine de l'accord au sein du groupe nominal. Nous renvoyons le lecteur intéressé à Fontenelle (2005b) pour plus de détails sur les types d'informations lexicales utilisés pour traiter certaines des erreurs ci-dessus, comme par exemple les fautes relatives au participe passé invariable des verbes pronominaux. Nous nous sommes ici plus particulièrement penchés sur les classes lexico-sémantiques inspirées de travaux de Levin (1993) pour modéliser les verbes de mutilation ou de soins corporels (*se fracturer, se casser, se luxer, se démettre, se briser, se fouler, se brosser, se peigner, se teindre*…), qui sont invariables lorsqu'ils sont utilisés transitivement avec une partie du corps (comparer : *elle s'est luxé l'épaule ; elle s'est brisé le doigt ;* et *la tasse s'est cassée*). Dans le cas présent, le correcteur détectera les erreurs dans « *Elle s'est fracturée la cheville* » ou « *ma sœur s'est lavée les mains* ».

7. Ressources linguistiques

Les paragraphes qui précèdent étaient relatifs aux outils de correction orthographique et grammaticale. Les outils linguistiques que nous produisons ou utilisons ne s'arrêtent évidemment pas là. Les séquenceurs (« *word-breakers* ») et les lemmatiseurs que nous développons nous permettent également d'affiner les technologies de la recherche en permettant l'expansion de requêtes pour augmenter la pertinence des documents trouvés par les moteurs de recherche. L'extraction d'entités nommées (dates, noms de personnes et de lieux, chiffres, montants monétaires, etc.) sont également au centre de nos préoccupations et, ici encore, des synergies sont établies entre nos différents projets puisque les bases de données lexicales que nous construisons sont utilisables dans divers scénarios.

La compétence de Microsoft se situe bien sûr principalement au niveau technologique et sa mission première n'est pas de construire des bases de données linguistiques. Pour un certain nombre de langues, nous développons toutefois nos propres ressources linguistiques en exploitant entre autres les énormes corpus de textes que nous acquérons constamment. Pour certaines applications, nous obtenons ce contenu linguistique auprès d'éditeurs traditionnels ou de fournisseurs de services linguistiques. Par l'intermédiaire de notre *Microsoft Language Resources Center (MSLRC),* nous acquérons ainsi des données lexicales, par exemple des dictionnaires monolingues ou bilingues qui sont mis à la disposition des utilisateurs par l'intermédiaire du volet de référence d'Office ou via l'encyclopédie Encarta. Ainsi, depuis Office 2003, les utilisateurs peuvent, directement depuis Word, obtenir la définition d'un mot et des exemples via la fonction de consultation qui permet d'accéder au dictionnaire monolingue français d'Encarta, le premier dictionnaire français électronique qui n'a jamais eu d'équivalent « papier » et qui fut dès le départ conçu comme une ressource pouvant être utilisée dans la perspective du TALN (Jacquet-Pfau 2005). L'utilisateur de Word a ainsi à sa disposition des dictionnaires monolingues pour le français, l'anglais, l'allemand et le japonais. Les dictionnaires bilingues, accessibles eux aussi via le volet de référence d'Office dans le traitement de texte Word, ont eux aussi fait l'objet d'une mise à jour récente pour tenir compte des recommandations orthographiques.

Les ressources lexicales sont à la base de la plupart des applications linguistiques que nous développons ou que nous offrons. Qu'il s'agisse des correcteurs orthographiques, des correcteurs grammaticaux, des thesaurus, de la reconnaissance de l'écriture pour le Tablet PC, des lemmatiseurs et des séquenceurs pour la recherche documentaire, de la reconnaissance et de la synthèse vocales, des aides à la rédaction (dictionnaires monolingues et bilingues), ou de la traduction automatique, on trouve à la base de toutes ces applications de riches bases de

données lexicales que nous développons en interne ou que nous acquérons en établissant des partenariats avec des éditeurs ou des associations telles qu'ELRA ou le LDC. J'ai décrit de façon plus détaillée les nouveaux outils de correction linguistique français, en montrant comment on peut juger de la plus-value qu'ils représentent par rapport aux versions précédentes et par rapport à des outils concurrents auxquels on les compare régulièrement. Les critères qui président à la décision de développer de nouvelles versions de ces outils ont été abordés, tout en insistant sur la nécessité de développer des outils robustes et utiles que l'on peut évaluer objectivement. Le français n'est qu'une des très nombreuses langues auxquelles nous nous intéressons (nous venons par exemple de lancer tout récemment un correcteur orthographique népalais ainsi qu'un correcteur maori). Tout ceci représente des investissements très importants. Ils sont nécessaires parce que les utilisateurs sont, à juste titre, exigeants, parce que les langues sont difficilement formalisables, nous le savons, et que chaque application linguistique nécessite de longs mois, voire des années de développement pour passer du stade du prototype tel que ce que l'on trouve dans les labos universitaires au stade d'outil robuste, souple, précis et utile pouvant être utilisé par des dizaines de millions de personnes. Le défi pour nous est d'être à l'écoute de nos utilisateurs et de transformer leurs requêtes en outils réalistes. Les améliorations que nous avons apportées ces derniers mois à nos correcteurs français reflètent ce souci de tenir compte de l'évolution de notre langue et de répondre à une demande explicite de la part de bon nombre de nos utilisateurs. La réforme de l'orthographe, la féminisation des noms de métiers ne sont que quelques-unes des facettes de ces nouveaux outils et le label de qualité qui a été octroyé à notre correcteur nous a encouragés à persévérer. Il reste évidemment encore beaucoup de problèmes à résoudre, mais nous y travaillons et la tâche est passionnante.

8. Remerciements

Je tiens à remercier ici mes collègues Thierry Etchegoyhen et Olivier Gauthier qui ont bien voulu relire mon texte et apporter des remarques constructives. Je reste bien sûr entièrement responsable des imperfections qu'il contient.

Références

CERQUIGLINI B. (réd.) (1999). *Femme, j'écris ton nom… Guide d'aide à la féminisation des noms de métiers, titres, grades et fonctions*. CNRS-INALF, Paris.

COMMISSION DE FÉMINISTATION DU CONSEIL SUPÉRIEUR DE LA LANGUE FRANÇAISE (2005) : *Mettre au féminin. Guide de féminisation des noms de métier, fonction, grade ou titre*. Ministère de la Communauté française de Belgique, Bruxelles. 2e édition.

CONSEIL SUPÉRIEUR DE LA LANGUE FRANÇAISE. (1990). « Les Rectifications de l'orthographe ». In *Journal Officiel de la République Française* 100, 6 décembre 1990.

CHRISTODOULAKIS D. (éd.) (2004). *Pre-Proceedings of the 1st Workshop on International Proofing Tools and Language Technologies*. Patras University, Patras.

DENIS B., KLINKENBERG J.-M. (2005) *La Littérature Belge – Précis d'histoire sociale*. Labor, Bruxelles.

DISTER A. (2004). « La féminisation des noms de métier, fonction, grade ou titre en Belgique francophone. État des lieux dans un corpus de presse ». In G. Purnelle, C. Fairon et A. Dister (éds), *Actes des JADT 2004.* Presses universitaires de Louvain, Louvain-la-Neuve : 313-324.

DISTER A., MOREAU M.-L. (2006). « Dis-moi comment tu féminises, je te dirai pour qui tu votes – Les dénominations des candidates dans les élections européennes de 1989 et de 2004 en Belgique et en France ». In *Langage et Société* 115.

FONTENELLE Th. (2004). « Lexicalization for proofing tools ». In G. Williams et S. Vessier (éds), *EURALEX 2004 Proceedings (11th EURALEX International Congress)*. Université de Bretagne-Sud, Lorient : 79-86.

FONTENELLE Th. (2005a). « Identifying tokens: Is word-breaking so easy ? ». In Ph. Hiligsmann, G. Janssens et J. Vromans (éds), *Woord voor woord. Zin voor zin. Liber Amicorum voor Siegfried Theissen*. Koninklijke Academie voor Nederlandse Taal- en Letterkunde, Gent : 109-115.

FONTENELLE Th. (2005b). « Dictionnaires et outils de correction linguistique ». In Th. Fontenelle (coord.), *Revue Française de Linguistique Appliquée* X (2), numéro spécial sur les « Dictionnaires : nouvelles approches, nouveaux modèles » : 119-128.

GREFENSTETTE G. (1996). « Approximate Linguistics ». In *Proceedings of the 4th Conference on Computational Lexicography and Text Research – COMPLEX'96*. Budapest.

GREFENSTETTE G., TAPANAINEN P. (1994). « What is a word, what is a sentence ? Problems of tokenization ». In *Proceedings of the 3rd Conference on Computational Lexicography and Text Research – COMPLEX'94*. Budapest : 79-87.

HANSE J., BLAMPAIN D. (1994). *Nouveau dictionnaire des difficultés du français moderne*. DeBoeck-Duculot, Louvain-la-Neuve.

JACQUET-PFAU Ch. (2005). « Pour un nouveau dictionnaire informatisé ». In *Éla, Revue de didactologie des langues-cultures et de lexiculturologie* 137 : 51-71.

LEVIN B. (1993). *English Verb Classes and Alternations: a Preliminary Investigation*. University of Chicago Press, Chicago.

MIKHEEV A. (2003). « Text Segmentation ». In R. Mitkov (éd.) *The Oxford Handbook of Computational Linguistics*. OUP : 201-218.

MITKOV R. (2003). The *Oxford Handbook of Computational Linguistics*. Oxford University Press, Oxford.

RILEY M., CRAVEN L., OLSEN M. (2004). « Customer-focused evaluation of a grammar checker ». In D. Christodoulakis (éd.), *Pre-Proceedings of the 1st Workshop on International Proofing Tools and Language Technologies*. Patras University, Patras.

VÉRONIS J. (2005a). *Texte : Correcteurs orthographiques en panne ?* Blog du 6 juillet 2005 : http://aixtal.blogspot.com/2005/07/texte-correcteurs-orthographiques-en.html.

VÉRONIS J. (2005b). *Ortograf – Ça repart chez Microsoft*. Blog d'octobre 2005 : http://aixtal.blogspot.com/2005/10/ortograf-repart-chez-microsoft.html

VÉRONIS J. (2005c). *Ortograf : OpenOffice vs. Microsoft*. Blog du 29 novembre 2005 : http://aixtal.blogspot.com/2005/11/ortograf-openoffice-vs-microsoft.html