HAL
open science

Problématique d'analyse et de modélisation des erreurs en production écrite. Approche interdisciplinaire (poster)

Jean-Léon Bouraoui, Philippe Boissière, Mustapha Mojahid, Nadine Vigouroux, Aurelie Lagarrigue, Frédéric Vella, Jean-Luc Nespoulous

To cite this version:

Jean-Léon Bouraoui, Philippe Boissière, Mustapha Mojahid, Nadine Vigouroux, Aurelie Lagarrigue, et al.. Problématique d'analyse et de modélisation des erreurs en production écrite. Approche interdisciplinaire (poster). 16ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2009), LIPN (Laboratoire d'Informatique de Paris-Nord, Université Paris 13 & CNRS); ATALA, Jun 2009, Senlis, France. hal-03620121

HAL Id: hal-03620121

https://hal.science/hal-03620121

Submitted on 29 Mar 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Problématique d'analyse et de modélisation des erreurs en production écrite. Approche interdisciplinaire

J.-L. Bouraoui*— Ph. Boissière* — M. Mojahid* — N. Vigouroux* — A. Lagarrigue*, F. Vella* — J.-L. Nespoulous**

** IRIT (Institut de Recherche en Informatique de Toulouse), UMR CNRS 5505*

Université Paul Sabatier

118, Route de Narbonne, F-31062 Toulouse Cedex

{bouraoui,boissier, mojahid, vigourou, alagarri, vella}@irit.fr

*** OCTOGONE, (E.A 4156) / Laboratoire Jacques Lordat*

Université de Toulouse II - Le Mirail Pavillon de la Recherche,

5, Allées Antonio-Machado, F-31058 Toulouse Cedex

nespoulo@univ-tlse2.fr

Résumé L'objectif du travail présenté ici est la modélisation de la détection et la correction des erreurs orthographiques et dactylographiques, plus particulièrement dans le contexte des handicaps langagiers. Le travail est fondé sur une analyse fine des erreurs d'écriture commises. La première partie de cet article est consacrée à une description précise de la faute. Dans la seconde partie, nous analysons l'erreur (1) en déterminant la nature de la faute (typographique, orthographique, ou grammaticale) et (2) en explicitant sa conséquence sur le niveau de perturbation linguistique (phonologique, orthographique, morphologique ou syntaxique). Il résulte de ce travail un modèle général des erreurs (une grille) que nous présenterons, ainsi que les résultats statistiques correspondants. Enfin, nous montrerons sur des exemples, l'utilité de l'apport de cette grille, en soumettant ces types de fautes à quelques correcteurs. Nous envisageons également les implications informatiques de ce travail.

Abstract The aim of our work is modeling the detection and the correction of spelling and typing errors, especially in the linguistic disabilities context. The work is based on a fine analysis of clerical errors committed. The first part of this article is devoted to a detailed description of error. In the second part, we analyze error in (1) determining the nature of the fault (typographical, spelling, or grammar) and (2) by explaining its consequences on the level of linguistic disturbance (phonological, orthographic, morphological and syntactic). The outcome of this work is a general model of errors (a grid) that we present, as well as the corresponding statistical results. Finally, we show on examples, the usefulness of this grid, by submitting these types of errors to a few spellcheckers. We also envisage the computer implications of this work

Mots-clés : Typologie et analyse d'erreurs textuelles, assistance à la saisie de textes
Keywords : Typology and analyze of textual errors, writing assitance systems.

1 Introduction

Un des principaux enjeux de la prise en compte des erreurs à l'écrit (production manuscrite ou clavier) sont leurs traitements computationnels immédiats ou différés (par le biais de correcteurs orthographiques) et leurs préventions à long terme grâce à des dispositifs d'apprentissage (particulièrement pour des enfants ou des étrangers apprenant le français). Or, ces enjeux, s'ils sont de mieux en mieux traités pour les sujets « sains », et ce malgré l'explosion de la communication par SMS par exemple qui postule pour un nouveau mode (Véronis et al. 2006), posent d'importants problèmes pour les personnes souffrant de handicaps langagiers, en fonction de la nature de la pathologie (cérébraux-lésés, sujets dyslexiques, parkinsoniens, aphasiques, paralysie cérébrale, etc.) ou d'incapacités motrices des membres supérieurs (myopathes, sujets parkinsoniens, etc.). Leurs productions linguistiques comportent de nombreuses erreurs, bien plus nombreuses que la moyenne comme nous le verrons dans cet article, et auxquelles les correcteurs orthographiques s'appliquent peu ou prou. Il est, par conséquent, nécessaire de disposer d'une identification et à terme d'une modélisation computationnelle des diverses catégories d'erreurs.

L'une des contributions principales de ce travail est la proposition d'une grille d'interprétation et d'annotation des erreurs à l'écrit. Celle-ci vise à rendre compte, de la manière la plus exhaustive possible, des erreurs survenues dans la production écrite, manuscrite ou par saisie clavier, de personnes présentant divers types de handicaps « centraux » ou « périphériques » du langage. Cette modélisation nous permet d'améliorer les processus de détection et de correction d'erreurs car en effet, les outils existants se trouvent défaillants face à ces types erreurs.

Ce travail s'inscrit dans les projets ESACIMC et CALAME[1]. Nous procédons à l'analyse (neuro)linguistique des erreurs des sujets IMOC/IMC[2] en situation de saisie sur claviers. Nous présentons les principaux enjeux et problèmes liés à notre étude. Nous recensons ensuite les différentes catégories d'erreurs et de perturbations qui leur sont associées. Nous montrons notamment que lors de la saisie par clavier, les erreurs peuvent avoir une motivation phonétique, morphémique, voire spatiale (relative à la configuration du clavier). Nous analysons le comportement des correcteurs automatiques face aux erreurs commises pour illustrer leurs limites; ici nous nous focaliserons sur les erreurs liées aux unités lexicales et à la gestion des blancs. Nous illustrons également l'apport de notre grille à l'aide d'exemples extraits du jeu d'essai. Nous montrons enfin quelques pistes concernant les stratégies pour améliorer les processus de détection et de correction.

2 Réflexions théoriques et méthodologiques

Compte tenu de l'existence de différents niveaux d'organisation de la structure des langues naturelles, il apparaît indispensable de rendre compte de l'ensemble des erreurs susceptibles de survenir à chacun de ces niveaux: littéral, graphémique, lexical, morphologique, syntaxique portant sur des entités linguistiques allant de la « lettre » à la « phrase » et au « texte ».

Ceci étant, rares sont finalement les erreurs qui (a) ne se situent qu'à un seul niveau et (b) n'ont pas d'impact (même indirect) à d'autres niveaux. Ainsi, une omission « locale » d'une préposition entraîne inéluctablement l'agrammaticalité de la phrase dans laquelle elle intervient (exemple : Il a posé l'assiette XXX la table). Pareillement, une erreur qui pourrait

---

[1] Évaluation qualitative de Systèmes d'Aide à la Communication pour les Infirmes Moteurs Cérébraux (http://www.irit.fr/ESACIMC/); CALAME: http://www.irit.fr/calame/.

[2] Infirmes Moteur d'Origine Cérébrale/ Infirmes Moteurs Cérébraux

n'être qu'orthographique (dans son déterminisme sous-jacent) peut entraîner, secondairement, une violation morphologique (exemple : il mangeais).

Il convient, par conséquent, d'adopter une démarche « multi-niveaux », laquelle requiert d'octroyer à une même erreur « superficielle » plusieurs étiquettes. Dans le premier exemple ci-dessus, nous serons ainsi amenés à étiqueter à un premier niveau, l'omission de morphème grammatical en tant que telle, avant d'ajouter une deuxième étiquette, au plan syntaxique cette fois (« agrammatisme »). Cela ne veut certes pas dire que le sujet a commis plusieurs erreurs. Cela veut simplement dire qu'en commettant une erreur à tel ou tel endroit du message, le sujet a entraîné plusieurs violations aux conditions de bonne formation des énoncés écrits. Par conséquent, ceci nous conduira à différencier, (a) des « erreurs à impact (simplement) local » et (b) des « erreurs à effets secondaires » ; ces dernières présentant, de toute évidence, un degré de gravité plus important, susceptible de perturber de façon massive l'échange d'informations entre l'émetteur (scripteur) et le récepteur (lecteur).

La « portée » de l'erreur complique passablement l'analyse dont il est ici question. Dans ce qui suit, nous présentons les points nécessitant une analyse multi-niveaux, dans la perspective que nous venons de décrire.

3 Typologie des erreurs et exploitation de la grille

A l'écrit, les erreurs peuvent intervenir à différents niveaux. Ils correspondent aux différentes dimensions de l'écrit, de la plus globale (mise en forme du document) aux plus fines (phonèmes, caractères de ponctuation, …). Nous pouvons les hiérarchiser, par granularité croissante, de la manière suivante :

– Mise en Forme Matérielle (Virbel 1989) du document (au plan spatial essentiellement). A ce niveau, les erreurs peuvent consister, par exemple, en des sauts de page ou retours à la ligne gênant dans l'organisation de l'ensemble du texte, des erreurs dans l'attribution des titres… ;

– Segmentation en phrases : il s'agit de la composition des énoncés produits en phrases bien formées (même si celles-ci comportent des erreurs aux niveaux inférieurs) par opposition à ce que l'on peut appeler « énoncés non-phrases » (par exemple, écrits en « style télégraphique ») ;

– Unité lexicale : il peut s'agir aussi bien de l'unité lexicale dans sa globalité que de ses entités constituantes : lettres, morphèmes, etc. ;

– Gestion de la ponctuation : ce niveau recouvre, par exemple, l'utilisation erronée d'un point à la place d'une virgule ou inversement, l'absence de ponctuation, etc. Cela concerne aussi bien les niveaux phrastique que intra-phrastique ;

– Gestion des blancs inter-mots, voire intra-mots, ces derniers sont appelés « erreurs logogrammiques » par (Catach 1980).

Dans cet article, nous ne nous sommes penchés que sur les niveaux : unité lexicale et gestion des blancs. Concernant le niveau unité lexicale : il se retrouve également dans les productions orales, ce qui peut être intéressant dans la perspective d'une typologie adaptable à d'autres modalités; d'autre part, il a fait l'objet de nombreuses études et expérimentations, notamment psycholinguistiques, sur lesquelles nous pouvons nous baser (cf. par exemple (Bonin 2007)). Quant au niveau, gestion des blancs, il est relié à celui des unités lexicales, puisqu'il peut affecter leur formation, comme nous le verrons plus bas. Il nous paraissait indispensable de le traiter dans le cadre de ce travail.

Une fois que nous avons délimité les niveaux sur lesquels nous allons travailler, il nous faut ensuite élaborer une typologie des erreurs qui s'y manifestent. Notre objectif est de permettre une description aussi précise que possible des erreurs, du niveau le plus concret au plus abstrait. Nous avons ainsi défini un ensemble de catégories, regroupées sous deux grandes étiquettes, correspondant respectivement à la description des erreurs, et à l'analyse de celles-

ci. Ces différentes catégories sont représentées dans la deuxième colonne du Tableau 1. Nous l'explicitons ci-après.

Nous avons réalisé une étude comparative des annotateurs et les résultats se sont avérés intéressants et prometteurs. Cependant, une objection pourrait être soulevée : l'annotation effectuée n'est-elle pas sujette à la subjectivité de celui qui la réalise ? Dans ce cas, les résultats obtenus varieraient grandement selon l'annotateur, et seraient donc inexploitables. Pour pallier cette argumentation le « coefficient Kappa » appelé aussi calcul de l'accord inter-annotateurs est un moyen utilisé pour évaluer la divergence ou la convergence des annotateurs. 4 spécialistes ont annoté les mêmes phrases issues des jeux d'essai (cf. infra), et la formule de (Carletta 1996) a été appliquée. Les résultats donnent un coefficient de 0,866 pour les types de fautes et 0,842 pour les niveaux d'erreurs. Notre schéma d'annotations peut être ainsi considéré comme robuste, et peut permettre des interprétations rigoureuses.

Notre premier objectif scientifique est d'étudier les erreurs provenant de pathologies affectant la production du langage, dans le but de les modéliser. Nous avons conduit cette première étude sur la base de la pratique antérieure de certains d'entre nous en matière d'étude de productions pathologiques, généralement chez des sujets cérébrolésés (voir par exemple, (Nespoulous et al, 1982)).

Lors de cette étude, nous nous sommes focalisés sur des textes écologiques de patients IMC. Ces textes nous ont été fournis d'une part, par le centre Kerpape dans le cadre du projet ESACIMC, et d'autre part, par le centre Les Iris de Carpentras.

L'échantillon de production fourni par Kerpape est composé de 472 phrases, produites par six sujets IMC[3]. Il était proposé aux sujets de s'exprimer librement. Nous en avons extrait 13 énoncés, provenant tous du même sujet[4] ; c'est à partir de ces énoncés qu'a été conçue notre grille d'annotation. Le choix des énoncés parmi l'ensemble de l'échantillon de production a été fait selon des critères de spontanéité et d'intelligibilité des énoncés écrits. Les phrases agrammaticales répondant directement à des questions scolaires ont été éliminées.

Concernant l'échantillon de production fourni par le centre de Carpentras, il s'agit d'un dialogue écrit – le sujet est atteint d'anarthrie – entre la patiente IMC adolescente, et son psychothérapeute, il comporte 9 énoncés. Il nous a servi à calculer la robustesse de notre grille d'annotation; combiné aux 13 énoncés du centre Kerpape, il a également permis de calculer la distribution des différentes catégories d'erreurs.

3.1 Grille et conventions d'annotation

Nous posons la question de savoir quand catégoriser un phénomène comme étant une erreur ou non, puis nous présentons les conventions d'annotation utilisées.

L'annotation se fait dans un tableau; une illustration est donnée dans le Tableau 2 : l'énoncé à analyser est présenté dans une ligne du tableau, et chaque unité lexicale faisant l'objet d'une ou plusieurs erreurs est mise en gras. Chaque ligne suivante est consacrée à une erreur. Nous appliquerons également cette règle dans le cas où une seule unité lexicale présente plus d'une erreur, à l'exception de la colonne « mot en faute », avec utilisation de la fonction « fusion des cellules ». Les colonnes correspondent aux différentes catégories d'erreur. Quand tous les mots en faute ont été ainsi présentés et traités, nous passons à l'énoncé suivant, et ainsi de suite. Les notations pour chaque niveau sont décrites ci-après.

---

[3] Nous ne disposons malheureusement pas de plus de détails sur le cadre de production de ces énoncés, notamment le type de clavier utilisé.

[4] Né en 1984, la cause de son handicap est une souffrance périnatale ayant entraîné une dyskinésie et, au plan de l'expression orale, une anarthrie. La motricité est de niveau V (mobilité très restreinte, même avec fauteuil roulant).

	Types de fautes	Typologie	Exemples
Description	Addition	Paragraphie Littérale (PL)	*fautteuil → fauteuil
	Déplacement	Paragraphie Littérale (PL)	* bein → bien
	Omission	Paragraphie Littérale (PL)	* deuxièm → deuxième
	Substitution	Paragraphie Littérale (PL)	*génécologue → gynécologue
Analyse	Faute de frappe	Paragraphie Littérale Clavier (PLC)	*bo, → bon
	Faute d'orthographe usuelle	Paragraphie Graphémique (PG)	*foyé → foyer
	Faute d'accord	Paragraphie Morphémique (PM)	* des famille → des familles
	Fusion de mots	Perturbation Morphémique Fusion (PeMf)	* papa ma expliqué → papa m'a expliqué
	Coupure de mots	Perturbation Morphémique Segmentation (PeMs)	*j'ai merai→ j'aimerai

Tableau 1 : Typologie et exemples des types de fautes

Dans le niveau « Analyse », l'annotation des différentes parties de ce niveau est faite sur trois colonnes :

– Colonne « Type de faute ou paragraphie » : utilisation des abréviations décrites ci-dessus. Dans le cas de la seule faute à conséquence orthographique, on laissera la case vide. S'il s'agit d'une erreur « indéchiffrable » (toute séquence de lettres/caractères ne pouvant être identifiée à un mot existant), c'est dans cette colonne que l'on notera « IND » ;

– Colonne « Niveau (conséquence) » : les commentaires fait pour la colonne « Type de faute ou paragraphie » s'appliquent également ici ;

– Colonne « Commentaire » : la rédaction est laissée à l'appréciation de l'annotateur, avec une préférence pour des commentaires succincts.

Les conventions énoncées ci-dessus sont appliquées à l'analyse de l'énoncé :

« il **estg** super **facip** à **utilisre** ». Le Tableau 2 illustre le résultat de l'annotation.

Mot en faute	**Description**				**Analyse de l'erreur**		
	Addition	**Omission**	**Déplacement**	**Substitution**	**Type de faute ou paragraphie**	**Niveau (conséquence)**	**Nature / Commentaire**
il **estg** super **facip** à **utilisre**							
Estg	G				PLC	Phono	
facip				p→l	PLC	Phono	
		E			PL	Phono	oubli [e] final
utilisre			r (-1)e(+1)		PLC	Phono	Métathèse réciproque

Tableau 2 : Exemple d'annotation

3.2 Analyse du comportement des correcteurs automatiques

Nous avons cherché à déterminer l'apport de notre approche d'identification des erreurs par rapport à celles déjà implémentées dans les correcteurs orthographiques, libres ou payants, du marché. Notre but est de déterminer dans quelle mesure les systèmes existants et « grand public » sont adaptés, ou non, aux erreurs produites par des sujets IMC. Notre analyse se compose de deux parties. D'une part, une comparaison statistique des performances par ces correcteurs et d'autre part, une analyse plus fine des performances obtenues par ces correcteurs, et de ce qu'elles pourraient être si l'on appliquait notre méthodologie.

3.2.1 Comparaison de correcteurs existants

Pour établir ces comparaisons, nous avons retenu trois correcteurs très largement répandus: Hunspell, Cordial (version démo: toutes les fonctionnalités activées, mais seulement pendant un mois), et le correcteur intégré à Microsoft Word 2003.

Nous avons soumis à chacun les 13 énoncés de l'échantillon de production décrit en début de section 3, et comptabilisé le nombre de fautes détectées, ainsi que les « faux positifs », c'est-à-dire les mots signalés comme faux alors qu'ils ne comportent pas d'erreurs. Dans ce contexte, nous employons le terme de « corpus » par raccourci, mais nous avons plutôt considéré l'ensemble des énoncés comme un jeu d'essai pour évaluer les limites des systèmes de correction. Cette méthodologie est couramment employée en génie logiciel, et consiste à définir les conditions dans lesquelles une application donnée est peu performante. Souvent, ces conditions sont choisies parce que l'application n'a pas été conçue pour les prendre en compte. Nous considérons que c'est la même problématique que nous adoptons avec les correcteurs orthographiques du marché, qui ne sont pas adaptés aux fautes produites par les sujets IMC/IMOC.

Bien que le logiciel Cordial et le correcteur de Word disposent de modules de corrections syntaxiques (à portée morphologique), les erreurs morphologiques ne sont pas détectées. La raison vient du fait que les erreurs commises en amont d'une faute à portée morphologique Problématique d'analyse et de modélisation des erreurs en production écrite. font perdre les repères au correcteur orthographique. Il suffit de corriger les erreurs antérieures pour que les fautes morphologiques deviennent détectables. De plus, Word comme Cordial signalent une faute si la première lettre de la phrase n'est pas en majuscule. Ce type d'erreurs (très fréquentes dans nos phrases tests, puisqu'une seule des phrases analysées commence par une majuscule), contribue également à « déstabiliser » le progiciel de correction.

Exemples :

(1) « *tu a vu quil ya des famille* … ». En (1), *tu* est détecté comme erreur de syntaxe et *quil* comme faute. En écrivant : (2) « *Tu a vu quil ya des famille* … » *a* est détecté comme faute de syntaxe. Si l'on corrige les trois premières fautes *a quil* et *ya*, on obtient : (3) « *Tu as vu qu'il y a des famille* … ».
Ici, tout le groupe des famille est souligné comme erreur morphologique (abusivement appelé faute de syntaxe par Word).

La couverture lexicale de Word est supérieure à celle de Hunspell et Cordial ce qui paradoxalement fait faire au premier un oubli supplémentaire. Dans (4) « *je pences* .. » Word sait que *pences* est le pluriel de *penny* et ne détecte pas la faute contrairement à Hunspell et Cordial. On voit donc qu'une trop grande couverture lexicale peut être nuisible quand les textes sont fortement entachés d'erreurs.

Enfin, seul Hunspell a détecté une erreur sur le *h* aspiré de (5) « *l'handicap de* … »

Les résultats chiffrés sont présentés dans le Tableau 3.

	Avec oubli des majuscules et retours à la ligne intempestifs (S1)		Sans oubli des majuscules et retours à la ligne intempestifs (S2)	
	Pourcentage de fautes détectées	Pourcentage de faux positifs	Pourcentage de fautes détectées	Pourcentage de faux positifs
Hunspel	54,42%	0,00%	55,10%	0,68%
Word	74,15%	21,09%	59,86%	2,72%
Cordial	65,99%	12,24%	63,27%	4,76%

Tableau 3 : Performances des principaux correcteurs du marché

On observe que les performances générales des correcteurs confrontés à ces erreurs sont assez faibles puisqu'elles sont de l'ordre de 50% d'efficacité. Pour plus de raffinement dans l'analyse, nous avons calculé les *taux de précision* (nombre de détections correctes par rapport au nombre total de détections (y compris les « faux positifs »)) et les *taux de rappel* (nombre d'erreurs correctement détectées par rapport au nombre total d'occurrences d'erreurs) obtenus. Ces deux métriques sont généralement utilisées dans l'évaluation des performances de module d'identification du TAL. Globalement, les correcteurs provoquent peu de « faux-positifs » ; par contre, les taux de rappel montrent des performances parfois basses, en étant parfois proche des 50%.

	(S1)		(S2)	
	Taux de précision	Taux de rappel	Taux de précision	Taux de rappel
Hunspel	100,00%	54,42%	98,77%	54,42%
Word	71,56%	53,06%	95,45%	57,14%
Cordial	81,44%	53,74%	92,47%	58,50%

Tableau 4 : Taux de précision et de rappel des trois correcteurs testés

On remarque également que les performances sont meilleures quand les erreurs portant sur les majuscules et les retours à la ligne (colonne S2) sont prises en compte. Comme nous allons le montrer dans la section suivante, la prise en compte de ce type d'erreurs est l'un des points sur lesquels notre méthodologie est un apport, dans le contexte d'écrits par des personnes souffrant de pathologies, par rapport aux correcteurs existants.

3.2.2 Apports potentiels de notre grille

Pour cette analyse détaillée, nous avons approfondi les mesures présentées précédemment. Le but est de déterminer concrètement si la grille que nous proposons, et le modèle qui peut en être tiré, permettrait de meilleures performances, que celles des correcteurs existants. Pour cela, nous étudions en détails les erreurs non ou mal traitées par ces systèmes, les raisons de ces problèmes de traitement, et les apports de notre grille à ces problèmes. Plus précisément, nous avons déterminé (uniquement sur le correcteur intégré de Microsoft Word) :

- Pour les fautes détectées par le correcteur orthographique, si aucune proposition n'est faite, si une voire plusieurs corrections sont proposées, et le cas échéant si une ou plusieurs sont pertinentes. La distinction entre le traitement de détection et de correction est importante. Comme le fait remarquer (Kukish, 1992, p. 378-379), le premier est bien plus facile que le second, en raison notamment de la complexité de la production morphologique de la langue. Nous distinguerons par ailleurs l'identification et l'étiquetage des erreurs au sein d'un même mot quand elles existent. Dans le domaine d'étude qui nous concerne ici, la difficulté de la correction est accrue par la présence d'erreurs absente des cadres de productions non pathologiques ;
- Pour les fautes non détectées par le correcteur, les raisons de l'absence de détection ;
- Pour les « faux positifs » les raisons de l'erreur.

A chaque défaillance avérée du correcteur, nous avons déterminé si notre grille permettrait une amélioration.

Il s'avère que c'est effectivement le cas, notamment en ce qui concerne les fautes relatives à la mise en forme des documents. En effet, beaucoup d'erreurs concernant certaines propriétés de mise en forme sont commises : ajout ou suppression d'espace de tabulation ou de retour à la ligne, et où le correcteur se trouve défaillant.

Par exemple dans « an visagè » l'espace inséré n'est pas considéré par le correcteur Word et l'erreur sur « an » n'est pas trouvée ; par ailleurs toutes les propositions (visage, vissage, visages) sont erronées. Dans un deuxième exemple d'erreur provoqué par un retour à la ligne non volontaire : « consulta ↳ ion »[5], le correcteur a proposé cinq corrections totalement hors sujet. Par ailleurs, il a pu détecter une erreur syntaxique concernant « ion » puisqu'il réclame systématiquement une majuscule en début de chaque paragraphe.

---

[5] Le ↳ représente le retour à la ligne.

Dans notre jeu d'essais, nous avons relevé un type d'erreurs fréquentes lié à la méconnaissance de l'orthographe et où les sujets cherchent à transcrire à l'écrit ce qu'ils pourraient énoncer à l'oral. Par exemple « musure daprand » ou « lord de mition ». Dans le premier exemple, toutes les propositions du correcteur sont fausses, et dans le deuxième exemple l'erreur sur « lord » n'est pas détectée.

En plus des erreurs de type PLC, PL et PG que VITIPI[6] corrige déjà, la grille des erreurs telle que nous la proposons permettra dans ces cas d'améliorer la détection et la correction des erreurs. En effet, la grille représente les particularités de la production écrite des sujets IMC/IMOC, à partir desquelles de nouveaux modèles et de nouveaux algorithmes de correction seront élaborés, en combinant les trois approches suivantes qui nous semblent complémentaires : (1) la prise en compte, dans les calculs de distance[7] de caractères typographiques comme l'espace ou le retour à la ligne, ou encore à l'intégration de règles spécifiques au contexte de production; (2) l'utilisation d'un lexique adéquat aux situations de productions considérées (par exemple, utilisation d'un langage « soutenu » ou au contraire plus « relâché » comme dans les productions que nous avons étudiées) et (3) le recours à un phonétiseur[8]. Ce dernier permettrait de déterminer les séparations entre mots, non en fonction de données écrites telles que les espaces ou les retours à la ligne, mais de l'enchaînement des phonèmes. De ce fait, il serait ainsi possible de considérer la suite « an visagè » comme une seule unité, indépendamment de l'espace inséré.

4 Conclusion

Ce travail transdisciplinaire (linguistique, psycholinguistique, rééducation, TAL) repose sur la mise en œuvre d'une démarche méthodologique d'annotation de production écrite, manuscrite et clavier fondée sur les divers niveaux d'organisation de la langue. Le premier résultat est la proposition d'une typologie des erreurs à deux facettes : l'une descriptive et l'autre analytique de catégorisation obtenue sur un jeu d'essai d'écrits d'adolescents IMC. Nous avons, ensuite, appliqué cette grille à l'annotation d'un jeu d'essai par quatre annotateurs. Le coefficient Kappa de 0.8 permet de confirmer la robustesse des consignes d'annotation. L'analyse de trois correcteurs a montré que nous obtenons au maximum un taux de rappel de 58, 50 % sur les corpus d'essai. Ce second résultat nous conforte dans la nécessité de considérer les particularités de la production écrite des sujets IMC dans les systèmes d'assistance à l'écriture (facilitation, correction et rééducation). Nous avons ainsi montré que, étant donné les spécificités des erreurs produites par les sujets, les correcteurs automatiques existants sont défaillants et que la grille que nous proposons apporte des améliorations tant qu'au niveau de la détection qu'au niveau de la correction.

A court terme, nous envisageons d'enrichir le schéma d'annotation et de mener une campagne d'annotation à plus grande échelle, sur plus de corpus (corpus à thématiques différents, à modalités de communication différente (dictée, dialogue, etc.) et sujets avec des handicaps langagiers différents, etc.) afin de couvrir toutes les erreurs de la production écrite. Nous poursuivrons également la multi-annotation afin de s'affranchir des subjectivités des

---

[6] VITIPI produit un texte sans avoir à taper toutes les lettres qui le composent, car il prédit des parties de mots et les affiche dès qu'il n'y a plus d'ambiguïté. Il est capable de prendre en compte les fautes de frappe, certaines fautes d'orthographe au fur et à mesure de la saisie , ainsi que des mots n'appartenant pas à son vocabulaire de base tout en continuant de prédire des lettres.

[7] Le calcul de la distance entre le mot cible et sa correction potentielle est une des méthodes les plus couramment employées pour les correcteurs orthographiques. Cf. par exemple (Kukish, op. cit.).

[8] Par exemple LIA_PHON (développé par Frédéric Béchet : http://lia.univavignon.fr/fileadmin/documents/Users/Intranet/chercheurs/bechet/download_fred.html), qui convertit un texte en la suite de phonèmes correspondants.

annotateurs dans cette tâche difficile d'interprétation. L'analyse des comportements des correcteurs a montré l'imperfection de la prise en compte de la mise en forme des textes, de la gestion de la ponctuation, dimensions qui nous semblent indispensables dans l'étude cognitive de la saisie de textes. Nous envisageons également de finaliser et de tester nos stratégies de détection et de correction; cet objectif est central dans le cadre de notre projet ANR PALLIACOM en cours .

Ce cadre méthodologique de détection et de catégorisation est un préalable indispensable à la modélisation des connaissances linguistiques représentant les troubles langagiers. Enfin à plus long terme, nous l'utiliserons à des fins de détection/correction au sein de systèmes d'assistance à l'écriture, pour personnes handicapées, comme VITIPI (Boissière et al., 2006), Sibylle (Wandmacher et al., 2007).

Références

BOISSIERE PH, SCHADLE I, ANTOINE J-Y. "A methodological framework for writing assistance systems: applications to sibylle and VITIPI systems", In *Modelling, Mesurement & Control, Série C, (bioenginaring). Edited by Association for the Advancement of Modelling & Simulation Techniques in Entreprise*, AMSE-journals (Barcelona Spain), Modellling C, Vol 67 (Supp Handicap 2006), pp. 167 - 176, 2006.

BOISSIERE PH., BOURAOUI J.-L, VELLA F., LAGARRIGUE A., MOJAHID M., VIGOUROUX N., NESPOULOUS J.L. « Méthodologie d'annotation des erreurs en production écrite. Principes et résultats préliminaires », *TALN07, Atelier « Reconstruire la langue dans les communications alternatives et augmentées »*, vol.2, p.529-538, Toulouse, juin 2007.

BONIN P., *Psychologie du langage : Approche cognitive de la production verbale de mots*, De Boeck, Collection « Ouvertures Psychologiques », 2007.

CATACH N. *L'enseignement de l'orthographe*, Paris, Nathan, 1980.

CARLETTA J. "Assessing agreement on classification tasks: the kappa statistic", *Computational Linguistics*, Vol.2, Issue 2, p.249-54, juin 1996.

KUKICH K., "Technique for automatically correcting words in text", *ACM Computing Surveys (CSUR)*, Vol. 24, Issue 4, p. 377-439, Year of Publication: 1992

NESPOULOUS J-L., LECOURS A.R., « Les troubles de l'écriture dans l'aphasie », *Etudes Françaises*, 18/1, p. 47-59, Les Presses de l'Université de Montréal, 1982.

VIRBEL J. "The contribution of linguistic knowledge to the interpretation of text structure". Dans Andre, J., Quint, V. et Furuta, R., (Eds) *Structured Documents*, p. 161–181. Cambridge University Press. 1989.

VERONIS J., GUIMIER DE NEEF E. « Le traitement des nouvelles formes de communication écrite ». In Sabah, G. (Ed.), *Compréhension automatique des langues et interaction*, p. 227-248. Paris: Hermès Science, 2006.

WANDMACHER T., ANTOINE J-Y., « Modèle adaptatif pour la prédiction de mots Adaptation à l'utilisateur et au contexte dans le cadre de la communication assistée pour personnes handicapées ». In *Revue TAL*. Volume 48 – n° 2/2007, p. 71 – 95