Méthodologie d'annotation des erreurs en production écrite. Principes et résultats préliminaires

Ph. Boissière (1), J.-L. Bouraoui (1), F. Vella (1), A. Lagarrigue (1),(2), M. Mojahid (1), N. Vigouroux (1), J.-L. Nespoulous (2)

(1) IRIT (Institut de Recherche en Informatique de Toulouse) Université Paul Sabatier, 118 Route de Narbonne, F-31062 Toulouse Cedex

(2) OCTOGONE / Laboratoire Jacques Lordat, Laboratoire Jacques Lordat, Université de Toulouse II - Le Mirail Pavillon de la Recherche, 5, allées Antonio-Machado, F-31058 Toulouse Cedex

Résumé Nous proposons une grille qui cherche à rendre compte, le plus exhaustivement possible, des erreurs survenues dans la production écrite, manuscrite ou clavier, de personnes présentant divers types de handicaps « centraux » ou « périphériques ». L'objectif est d'obtenir une modélisation fine des erreurs survenant pendant la saisie. Le résultat de cette modélisation sera implémenté dans les systèmes d'assistance à la saisie. L'une des étapes de la modélisation est cette grille d'analyse, que nous présentons ici. Elle se décompose en deux parties : la première décrit l'erreur sa nature. La seconde analyse l'erreur et détermine sa conséquence au niveau linguistique. Mettre en œuvre cette grille dans un tableur permet de calculer automatiquement le nombre et le type de fautes pour chaque individu. Le pédagogue possède ainsi un outil lui indiquant les faiblesses langagières de l'utilisateur et essayer de le rééduquer. Nous donnons les premiers résultats obtenus à partir d'écrits d'adolescents IMC.

Abstract We propose a grid which seeks to make an exhaustive explanation of the errors occurring during the writing production, handwriting or keyboard, for people presenting various types of "central" or "peripheral" handicaps… The objective is to obtain a subtle modeling of the errors occurring during a keyboarding. The result of this modeling will be introduced in assistance systems to the keyboarding. One of the stages of modeling is this analysis grid, which we present here. This grid is made up of two parts: the first describes the error and indicates its nature. The second analyzes the error and establishes its linguistic consequence. To implement this grid in a spreadsheet makes it possible to automatically calculate the number and the type of faults for each individual. Thus the pedagogue has a tool which indicated the linguistic weaknesses of the user and with which he can try to rehabilitate it. We give the first results obtained from IMC teenagers' writings.

Mots-clés : Analyse et typologie d'erreurs textuelles, assistance à la saisie de textes

Keywords: Analysis and typology of textual errors, assistance to texts input

1. INTRODUCTION

Errare humanum est. Cet adage se vérifie partout, y compris évidemment pour la communication et à l'expression. Dans le contexte du dialogue oral, il est fréquent que les erreurs soient repérées et corrigées sans problèmes. Or, quand ce n'est pas le cas, les conséquences peuvent poser de gros problèmes. Il faut donc maximiser l'identification et la compréhension des erreurs d'ordre linguistique, et minimiser leurs conséquences. C'est un gros travail puisqu'il concerne tous les niveaux de communication langagière, notamment les deux principaux : l'oral et l'écrit De plus, il existe différents objectifs et méthodes pour aborder les erreurs propres à chaque niveau, même s'ils peuvent s'harmoniser.

Le travail présenté ici suit une démarche d'ordre neuro-psycholinguistique. Il trouve son origine première dans les travaux menés par l'un de nous (Jean-Luc Nespoulous), en collaboration avec André Roch Lecours (1979) sur les productions écrites déviantes des patients aphasiques. Il est également complété, sur tel ou tel point, par certains éléments en provenance des travaux sur l'écriture de Nina Catach (1980) et de son équipe.

Le thème central de l'article est de proposer une grille d'interprétation et d'annotation des erreurs à l'écrit. L'objectif de cette grille est de rendre compte, de la manière la plus exhaustive possible, des erreurs survenues dans la production écrite, manuscrite et sur clavier de personnes présentant divers types de handicaps « centraux » ou « périphériques ».

Cet article s'inscrit dans le cadre du projet ESACIMC[1]. Notre premier objectif est d'analyser les erreurs des sujets IMC (Infirmes Moteur Cérébraux) en situation de saisie sur clavier de messages écrits. Les corpus utilisés ont été écrits dans ce contexte, mais les conclusions qui sont tirées sont généralisables.

Nous présentons d'abord les principaux enjeux et problèmes liés à notre étude. Nous passons ensuite en revue 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[2]. Nous donnons enfin quelques résultats.

2. REFLEXIONS ET ANALYSES PRELIMINAIRES

2.1. De la nécessité (et de la difficulté) d'une analyse « multi-niveaux » des erreurs de la production écrite

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

Rares sont finalement les erreurs qui (a) ne se situent qu'à un seul niveau et (b) n'ont pas d'impact (même indirects) à d'autres niveaux. Ainsi, telle omission « locale » d'une préposition entraîne l'agrammaticalité de la phrase dans laquelle elle intervient (ex. : « Il a posé l'assiette **XXX** la table »). Pareillement, une erreur qui pourrait n'être qu'orthographique peut entraîner, secondairement, une violation morphologique (ex. : « il mangeais »).

[1] Evaluation qualitative de Systèmes d'Aide à la Communication pour les Infirmes Moteurs Cérébraux
[2] La méthodologie d'annotation et ses bases théoriques sont décrites bien plus en profondeur dans (Bouraoui et al., 2007)

Nous adoptons une démarche « multi-niveaux » qui demande d'octroyer à une même erreur « superficielle » plusieurs étiquettes. Dans le premier exemple ci-dessus, on sera donc amené 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 »). Le sujet n'a pas commis plusieurs erreurs. Cela veut dire qu'en commettant une erreur « locale » à tel ou tel endroit du message, il a entraîné plusieurs violations aux conditions de bonne formation des énoncés (ici écrits). Ceci nous conduit à différencier, (a) des « erreurs locales à impact (simplement) local » et (b) des « erreurs locales à effets secondaires », ces dernières présentant un degré de gravité plus important, susceptible de perturber de façon massive l'échange d'informations. On aura bien compris que ce point complique passablement l'analyse dont il est ici question.

2.2. Le passage de l'oral à l'écrit : autre problème

S'il est envisageable, dans certains cas, d'analyser les erreurs de la production orale sans tenir compte des représentations écrites des (séquences de) mots[3], il est inenvisageable d'analyser l'écrit sans prendre en considération l'oral. L'écrit n'est en effet qu'une transcription de l'oral qui a été acquis le plus souvent en premier. Dès lors, bon nombre d'erreurs dans la production écrite sont influencées (« contaminées ») par la nature des représentations orales, souvent « co-activées » lorsque le sujet entreprend sa tâche d'écriture !

Ces problèmes oral/écrit sont d'autant plus importants que les règles de conversion phonèmes/graphèmes sont opaques, comme c'est le cas en français ou en anglais où un même phonème peut s'orthographier de n manières différentes ! L'existence d'homophones non homographes constitue donc un problème majeur dont une grille d'analyse complète doit pouvoir rendre compte.

Dans le passage oral ➔ écrit, il conviendra de différencier, (a) les erreurs qui n'entraînent pas de changement de prononciation du mot écrit erroné (Ex : BATEAU ➔ * BATO) et (b) celles qui modifieraient la prononciation du mot si on avait à le produire (Ex : BOISSON ➔ BOISON).

2.3. La chronométrie de la production écrite : une variable importante

La chronométrie de la production d'un message oral est aisée à réaliser : un magnétophone suffit, lequel enregistrera aussi bien les plages de production véritable que les silences (pauses et hésitations). Mais il n'en va pas de même lorsqu'il s'agit d'analyser la production écrite. La plupart du temps les erreurs sont analysées « off line », quelques instants (au mieux) après leur production. On mesure bien l'intérêt qu'il y a, depuis l'arrivée des tablettes graphiques, et des outils de suivi du regard, de prendre en considération ces paramètres temporels. Ils sont de nature à permettre une analyse plus fine de la production écrite (ex. : arrêt du scripteur entre l'écriture du radical et celle de la désinence d'un verbe… indiquant vraisemblablement que le sujet n'est pas à l'aise dans sa gestion de la morphologie flexionnelle verbale !). Il en va selon nous de même dans l'écriture sur clavier, même si bon nombre d'individus utilisant un clavier ne sont pas experts en dactylographie, (ce qui peut rendre l'interprétation des pauses plus difficile : rechercher la bonne touche… indépendamment de ses connaissances linguistiques…).

[3] Encore qu'il existe des cas où la forme écrite d'un mot vient contaminer la production de sa forme orale !

Même si dans ce travail, une étude « on line » n'est pas envisagée, il faudra prendre en considération, (a) les éventuelles mauvaises segmentations de mots, (b) les problèmes majeurs de ponctuation, voire (c) les tentatives d'autocorrection. Ces dernières, si elles ne sont pas systématiquement relevées peuvent conduire à des erreurs de diagnostic (ex. : analyse de « le plus que je possible » comme énoncé dysyntaxique si on ne prend pas en considération le fait que le sujet s'est arrêté un bon moment après le « je » !).

2.4. Erreurs liées à la configuration spatiale des lettres sur un clavier

La saisie « clavier » a ses propres contraintes qui peuvent être à l'origine d'erreurs que l'on n'observerait pas en production manuscrite.

En écriture manuscrite, comme en production orale, une erreur « segmentale » trouve son origine soit dans la confusion entre segments (en général) proches du point de vue de leurs propriétés intrinsèques (phonologiques ou orthographiques) : cf. /p/ VS /b/, soit dans la survenue de « contaminations contextuelles » (ou syntagmatiques). Ces « contaminations » ont pour effet la réduplication à courte distance de segments de l'environnement antérieur (*persévérations*) ou postérieur (*anticipations*). Les dyslexiques en savent quelque chose !

En écriture sur clavier, à ces erreurs (toujours possibles) s'ajoutent celles qui peuvent émaner de la proximité spatiale de certaines lettres, et ce, même si « lettres substituantes » et « lettres substituées » n'ont rien en commun dans le système alphabétique. Par exemple, les touches correspondant aux lettres E, R, S, D, F sont toutes situées sur la même zone d'un clavier AZERTY. Mais elles n'ont aucun point commun en termes de graphie (comme par exemple p et q), ou de sonorité (comme m et n). Ce point risque d'être crucial pour divers types de populations pathologiques présentant des problèmes moteurs importants (ainsi d'ailleurs, dans une moindre mesure, que pour tout autre catégorie de public).

Ainsi, si la saisie clavier est susceptible d'aider certains sujets à produire de l'écrit alors qu'ils ne sont pas capables de le faire via leur main dominante, cette même saisie vient, d'un autre point de vue, rajouter une nouvelle source d'erreurs ! On peut envisager de modéliser ce type d'erreurs en attribuant une « pondération » aux lettres faisant l'objet d'une erreur, en fonction de leur proximité plus ou moins grande sur le clavier.

2.5. Erreurs VS Stratégies

Encore un dernier point qui n'est pas toujours facile à gérer. Il est tentant de qualifier d'erreur toute production non canonique. Or, il arrive que certains phénomènes erronés ne soient pas la conséquence directe d'un « déficit », mais plutôt, la mise en œuvre de stratégies (plus ou moins volontaires) susceptibles de faciliter la tâche à l'émetteur ; les écrits SMS en fournissent de bons exemples : « g » pour « j'ai »…). La systématicité d'une erreur peut aussi indiquer la mise en place d'une « stratégie » (cf. s précédent), d'où les difficultés déjà mentionnées ! Pour trancher entre « erreur *de compétence* » et « stratégie » systématique, il faut soumettre le même sujet à une tâche de jugement de grammaticalité. S'il est bon dans cette tâche alors qu'il recourt spontanément (et systématiquement) à certaines formes erronées, c'est vraisemblablement qu'il a adopté une stratégie, et ce consciemment ou pas…. En TALN, cette distinction est fondamentale puisque c'est à partir de phénomènes linguistiques invariants que l'on peut établir des modèles et des algorithmes.

3. TYPOLOGIE DES ERREURS

Nous présentons d'abord les différents niveaux auxquels une analyse des erreurs intervenant à l'écrit (avec un focus sur la saisie clavier) peut être menée. Nous nous intéressons ensuite à l'un de ces niveaux en particulier, pour des raisons que nous expliquons. Nous nous livrons enfin à une catégorisation détaillée des erreurs pouvant survenir au niveau choisi.

3.1. Approche macroscopique

A l'écrit, les différents niveaux auxquels interviennent les erreurs sont au nombre de 5. On peut les hiérarchiser, du plus global au plus particulier, de la manière suivante :

1. Recours (ou non) à de la MFM[4] (au plan spatial essentiellement).

2. Problèmes de segmentation en phrases (i.e. énoncés phrases, même si celles-ci contiennent des erreurs) VS énoncés non-phrases (i.e. « style télégraphique ») ;

3. Problèmes au niveau du substantif : il peut s'agir aussi bien du substantif dans sa globalité que de ses entités constituantes : lettres, morphèmes, etc. ;

4. Problèmes de gestion de la ponctuation : point VS virgule, au niveau phrastique et intra-phrastique ;

5. Problèmes de gestion des blancs inter-mots, voire intra-mots, ces derniers sont appelés « erreurs logogrammiques » par (Catach, 1980).

Dans ce travail, nous ne nous sommes penchés que sur le niveau 3 pour plusieurs raisons. D'une part, il s'agit du niveau le plus facile à modéliser et pour lequel on peut implémenter des améliorations dans des logiciels d'assistance. D'autre part, on retrouve des analogies avec les productions orales. Enfin, il a fait l'objet de nombreuses études et expérimentations, notamment psycholinguistiques, sur lesquelles nous pouvons nous baser.

3.2. Approche analytique

3.2.1. Types d'erreurs

Le but est de permettre une description aussi précise que possible des erreurs, du niveau le plus concret jusqu'au plus abstrait. Pour cela, on utilise une hiérarchie de catégories. Celle-ci est représentée dans la figure 1, que nous explicitons immédiatement après.

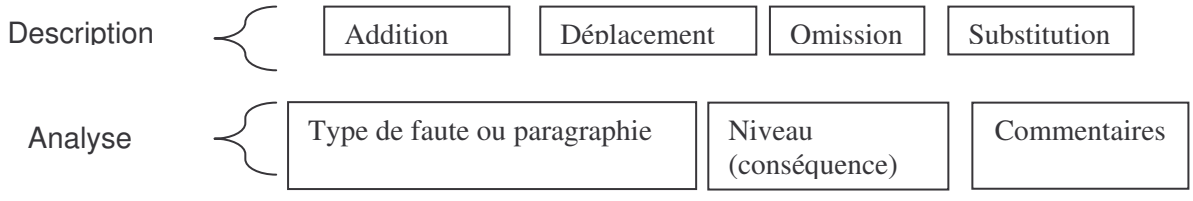

Figure 1 : hiérarchie des catégories d'erreurs

[4] Mise en Forme Matérielle du texte. Pour résumer, cette théorie (Virbel, 1989) postule notamment la prise en compte de la mise en forme du texte écrit (indentation, mise en gras, etc.) dans la transmission du sens. Par exemple la mise en italique d'un mot indique un focus particulier porté sur celui-ci. Mentionnons les travaux de (Luc, 2000) sur l'architecture des énumérations et la typographie des paragraphes, qui fournit des éléments de réflexion pour le niveau 1.

Le premier niveau, « *description* », correspond à la manifestation de la faute. En effet, toute faute appartient systématiquement à l'une et une seule de ces catégories : *Addition*, *Déplacement*, *Omission*, *Substitution*. Nous nous y réfèrerons via l'acronyme *ADOS*. Cette catégorisation a été modélisée par Levenshtein (1966), et a donné lieu à de nombreux développements et applications dans diverses disciplines liées au TALN.

Le deuxième niveau, « *analyse* », correspond à une tâche plus délicate. Il s'agit de catégoriser plus finement l'erreur, de situer le niveau qu'elle affecte, et d'avancer des hypothèses motivant son apparition. Dans la partie suivante, nous décrivons chacune de ces étapes.

Quand une seule erreur de l'un de ces types apparaît dans un énoncé, l'analyse ne pose pas de problème. Cependant, si plusieurs erreurs de l'un ou l'autre de ces types apparaissent (surtout dans le cas des omissions), l'analyse devient plus complexe, à fortiori s'il s'agit de l'omission de plusieurs mots grammaticaux. Dans cas, il vaudra mieux recourir à l'étiquette d'« agrammatisme » pour caractériser l'énoncé en question (sans chercher à quantifier le nombre d'omissions de morphèmes… ce qui de plus, s'avèrerait quasiment impossible)[5].

3.2.2. Types d'unités linguistiques perturbées (et leur interprétation)

S'agissant du langage écrit, les unités suivantes sont pertinentes et sont toutes susceptibles d'être « malmenées » en situation de production écrite :

Reprenons les deux niveaux évoqués dans la figure 1, avec pour chacun les différentes catégories correspondantes :

Niveau « Description » *:* Les différentes catégories référencées *ADOS* sont assez explicites, et des exemples sont donnés dans la section suivante. Mais il faut apporter deux précisions.

D'une part, l'unité pouvant faire l'objet d'un des phénomènes catégorisés est en général la lettre (ou le caractère du clavier), mais selon les circonstances, il pourra s'agir de plusieurs caractères, ou encore d'un ou plusieurs phonèmes. On notera qu'en français, lettre et phonème ne sont pas systématiquement dans une relation bijective. En effet, des phonèmes peuvent être réalisés orthographiquement par plusieurs lettres pour aboutir à un « graphème ». Un exemple typique est le phonème /ã/, qui s'écrit « empts » dans « exempt » ou le /a/ de « femme ».

D'autre part, les erreurs concernant les diacritiques font l'objet d'un traitement particulier quand on utilise un clavier d'ordinateur pour produire le texte. On fera en effet la distinction entre d'un côté les lettres accentuées présentes « d'un bloc » sur une seule touche de clavier (en français : ç, é, è, à, ù), et de l'autre celles formées par la combinaison de 2 touches (en français, diacritique ^ ou ¨, et une voyelle). Dans le premier cas, une seule faute sera comptée. Dans le second, on pourra compter 1 à 2 fautes : une pour le diacritique, et une, éventuellement, pour la lettre accentuée.

On l'a vu, pendant l'écriture, une lettre ou un groupe de lettres peuvent être altérés ou écrits d'une autre façon avec une autre graphie. Ainsi, peuvent se produire des *paragraphies* provenant d'origines diverses et dont nous analyserons les conséquences.

Ici, la classification est faite selon l'unité linguistique perturbée :

- La « lettre » : une erreur de ce type sera appelée « *paragraphie littérale* » (**PL**)

[5] Alors que les omissions, ici ou là, de lettres peuvent, elles, être aisément quantifiées.

Ex[6] : TORNADE ➔ *TORNADRE addition (avec persévération)
CULTIVATEUR ➔ *CURVILATEUR déplacement
CHERCHER ➔ *CHECHER omission
FOURNIR ➔ *FOURFIR substitution (avec persévération)

C'est à cette première catégorie d'erreurs que viendront se rajouter les Paragraphies Littérales engendrées par proximité des touches sur le Clavier (**PLC**). Pour une prise en compte des PLC dans un système d'assistance à l'écriture, cf. (Boissière & Dours 1996, p.170).

Concernant la catégorie des PLC, il est important de préciser qu'on ne doit l'employer seulement que pour les erreurs d'addition, ou de substitution entre une lettre cible et un périmètre situé immédiatement autour[7] (faute de frappe).

Appartiennent également à cette catégorie, les erreurs portant sur les diacritiques (accents) : omissions VS substitutions (ces dernières venant modifier la lecture à haute voix des mots écrits de manière erronée. Ex : « fenêtre » ➔ « * fénêtre »).

- Le « graphème » : une erreur de ce type sera appelée *« paragraphie graphémique »* (**PG**).

  Il s'agira essentiellement de substitutions. Le « graphème » est l'équivalent (ortho)graphique d'un phonème. Une « lettre » peut correspondre à un graphème mais souvent, surtout dans une langue comme le français, un graphème nécessite l'emploi de plusieurs lettres (parfois nombreuses) comme nous le signalons plus haut.

  Cette catégorie est donc surtout utile pour qualifier les erreurs de production écrite de *phonèmes hétérographes* : le sujet utilise une variante graphémique erronée (qui, néanmoins, permet de renvoyer au bon phonème à l'oral).

  Ex : BIBLIOTHEQUE ➔ *BIBLIOTEC
  FRANÇAIS ➔ *FRANÇAIT
  COMMENCEMENT ➔ *COMANSEMENT

- Le « morphème » : une erreur de ce type sera appelée *« paragraphie morphémique »* (PM).

  Il s'agit des « erreurs morphogrammiques » de Catach. Dans sa forme la plus simple à analyser, il s'agira d'une substitution de morphèmes (le plus souvent flexionnels) : « il chantait » ➔ « * il chantais ».

  Rentrent également dans cette catégorie les erreurs d'accords morphologiques, les omissions ou additions de morphèmes grammaticaux et les erreurs de préfixes (par exemple IMPOSSIBLE ➔ * INPOSSIBLE).

  Il y a enfin une autre catégorie, celle des *« Perturbations Morphémiques »* (PeM). Elle se divise en deux sous-catégories :

[6] Exemples tirés de (Lecours et al.1979).
[7] En attendant qu'éventuellement des études plus poussées élargissent le champ d'application des PLC.

- *« Perturbation Morphémique fusion»* (PeMf) : omission d'une lettre (caractère) séparant deux substantifs (espace, trait d'union, apostrophe), et aboutissant ainsi à la réunion de ceux-ci. Ex TRAIT D'UNION: ➔ * TRAIT DUNION

- *« Perturbation Morphémique segmentation»* (PeMs) : addition d'une lettre (caractère) séparant (espace, trait d'union, apostrophe) à l'intérieur d'un substantif, aboutissant ainsi à la segmentation de ce dernier en deux unités. Ex : PROPOSER ➔ * PROP OSER

Niveau « analyse » : les différents niveaux pouvant être affectés sont les niveaux orthographique, phonologique, morphologique et syntaxique.

En règle générale, les fautes relevant de la PL affectent seulement le niveau orthographique et/ou phonologique ; les PG et les PM peuvent affecter, selon les cas, les 4 niveaux. Voici quelques indications supplémentaires pour chaque catégorie (hors orthographe) :

- *Phonologique* : pour déterminer si une erreur concerne le niveau phonologique, il suffit de comparer la représentation phonétique du mot erroné avec celle du mot cible, par exemple en lisant à haute voix. Si les deux représentations sont identiques, alors on ne parle pas « d'erreur à conséquence phonologique ». L'interprétation peut différer selon les variantes locales, comme la prononciation ou non du e final.

- *Morphologique* : à ne pas confondre les erreurs morphologiques et syntaxiques. Les erreurs d'ordre morphologiques, comme leur nom l'indique, concernent uniquement la morphologie grammaticale du mot. Il peut s'agir par exemple d'une faute de flexion (conjugaison, genre, nombre …).

- *Syntaxique* : le terme de « syntaxe » est pris ici au sens de la position et de la fonction des mots dans l'énoncé. Donc, on ne parlera d'erreur de syntaxe que dans le cas d'erreurs à ce niveau. Par exemple, l'omission d'un substantif, ou son déplacement à une autre position que celle qui lui est assignée dans la langue.

Précisons enfin que nous avons été amenés à définir des règles pour déterminer ce qui constitue une erreur ou pas. Nous n'avons pas voulu centrer notre publication sur ce point ; le lecteur intéressé se reportera à (Bouraoui, 2007).

4. Statistiques préliminaires

Nous avons extrait d'un corpus de 472 phrases fournies par le centre KERPAPE dans le cadre du projet ESACIMC, un ensemble de 13 phrases provenant d'un même sujet. Ces phrases ont été analysées selon notre méthodologie. L'annotation se faisant sous Excel, nous pouvons obtenir directement à l'aide de macros les résultats suivants : nous avons séparé, comme dans notre grille, la partie description de celle d'analyse. Dans chaque figure, les nombres en gras situés à gauche des pourcentages correspondent aux nombres d'occurrences.

4.1. Description

Sur la figure 2, nous constatons que près de la moitié des fautes sont des omissions (46 %) suivies des substitutions (39 %). La lecture de la figure 3 montre que seulement 1 % de ces erreurs sont des fautes de frappes (ce qui peut surprendre pour un sujet IMC). Les paragraphies littérales (27 %), graphémiques (34 %) et morphémiques (22 %) représentent à

elles seules (83 %) des erreurs. Cela indique probablement des difficultés orthographiques importantes.

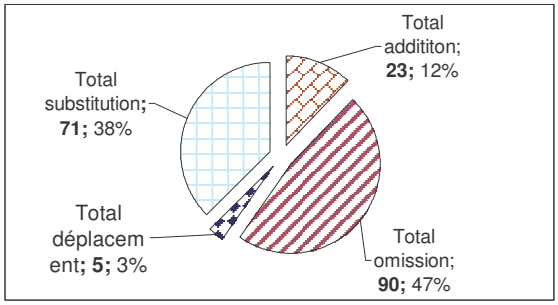

Figure 2 : Répartition des fautes

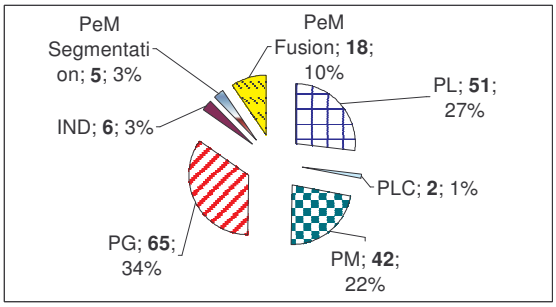

Figure 3 : distribution des types de fautes

4.2. Analyse

Au niveau de l'analyse, on constate (Cf. Figure 4) qu'effectivement, 51 % des fautes sont à conséquence orthographiques, et 26 % à conséquence morphologiques. Seulement 3 % sont des fautes de syntaxe (oubli de mot) et 20 % des fautes entraîneraient une mauvaise lecture par une synthèse vocale

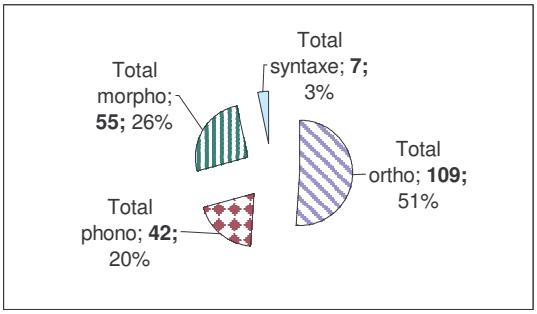

Figure 4 : Distribution des niveaux affectés

5. CONCLUSION ET PERSPECTIVES

Le travail que nous avons présenté n'en est encore qu'à ses débuts. Nous avons fait annoter un corpus d'une douzaine de phrases par deux annotateurs différents, à partir des règles décrites dans ce document, et détaillées dans (Bouraoui et al., 2007). A court terme, nous envisageons de mener une campagne d'annotation à plus grande échelle, sur plus de corpus et avec plus d'annotateurs. Cela nous permettra d'obtenir une mesure de « l'accord inter annotateurs ». Cette mesure nous permettra d'évaluer de manière fiable et objective la stabilité de notre grille d'annotation et sa capacité à être utilisée de manière systématique.

Nous comptons également mieux mettre en place un corpus « artificiel », conçu à partir de diverses tâches de production écrite identiques pour tous les sujets, éventuellement réutilisées chez les mêmes sujets à différents moments (étude longitudinale). Ce corpus serait complémentaire de l'actuel. Ce dernier, collecté dans des conditions de production spontanée, séduit par son caractère naturel, mais les données obtenues auprès d'un sujet sont difficilement comparables avec celles glanées chez un autre sujet dans une situation différente !

Enfin, à plus long terme, on envisage d'automatiser ce travail pour pouvoir l'intégrer dans un logiciel d'aide à la saisie. Ce logiciel, développé par l'un de nous (Boissière, 1996) est prêt à prendre en compte les résultats obtenus.

Remerciements

Cette étude a été réalisée grâce aux financements de l'APETREIMC et de la Fondation Motrice. Les auteurs expriment également leur reconnaissance à nos partenaires du Centre Mutualiste de Rééducation et de Réadaptation Fonctionnelles de Kerpape qui nous ont gracieusement fait parvenir les textes écrits par certains de leurs patients. Que ces derniers en soient ici remerciés.

Références

BAUDOT J.A. (1968) *Information, redondance et répartition des lettres et des phonèmes en français*, Rapport, Université de Montréal, mars 1968.

BOISSIÈRE PH., DOURS D. (1996) "VITIPI : Versatile Interpretation of Text Input by Persons with Impairments". In *5th ICCHP (International Conference on Computers for Handicapped Persons)*. pp.165-172, Linz July 1996.

BOURAOUI J.-L., BOISSIÈRE PH., VELLA F., LAGARRIGUE A., MOJAHID M., LAUR D., VIGOUROUX N., NESPOULOUS J.-L. (2007) *Prolégomènes à l'étude des erreurs en production écrite - Propositions en vue de la mise au point d'une grille d'analyse*, Rapport IRIT/RR—2007-7-FR, Mars 2007.

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

LECOURS A. R. (1966) "Serial order in writing – a study of misspelled words in "developmental dysgraphia"", *Neuropsychologia*, Vol.4, pp. 221-241.

LECOURS A. R., DELOCHE G., LHERMITTE F. (1973) Paraphasies phonémiques – description et simulation sur ordinateur –, in *Colloque INRIA-Informatique Médicale*, pp.311-351, Rocquencourt.

LECOURS A. R., LHERMITTE F. (1969) Phonemic paraphasias: linguistic structures and tentative hypotheses, *Cortex*, 5, pp.193-228.

LECOURS A. R., LHERMITTE F. ET AL. (19) *L'aphasie*, Paris, Flammarion.

LECOURS, A.R., DORDAIN, G., NESPOULOUS, J-L. & LHERMITTE, F. (1979) « Le vocabulaire de la neurolinguistique », in A.R. Lecours & F. Lhermitte (Eds) *L'aphasie*, Paris, Flammarion.

LECOURS, A.R., NESPOULOUS, J-L (1982) « Biologie de l'écriture », *Etudes Françaises*,18/1, 33-45, Les Presses de l'Université e Montréal.

LEVENSHTEIN V.I. (1966) Binary codes capable of correcting deletions, insertions, and reversals, *Cyber. Contr. Theory*,10 (8), pp. 707-710.

LUC C. (2000) *Représentation et composition des structures visuelles et rhétoriques du texte. Application à la génération de textes formatés.* Thèse de doctorat, Université Paul Sabatier, novembre 2000.

MOUNIN G. (1970) *Introduction à la sémiologie*, Paris, Editions de Minuit.

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

NESPOULOUS, J-L. & VIRBEL, J. (2004) Apport de l'étude des handicaps langagiers à la connaissance du langage humain, *Revue Parole*, N°29-30, pp. 5-42.

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