Modèle de vérification grammaticale automatique gauche-droite

Agnès Souque

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEURE DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Sciences du langage, Spécialité Informatique et Sciences du langage**

Arrêté ministériel : 7 août 2006

Présentée par

Agnès SOUQUE

Thèse dirigée par **Thomas LEBARBÉ**

préparée au sein du **Laboratoire LIDILEM – EA 609**
dans **l'École Doctorale n° 50 – Langues, Littérature et Sciences Humaines**

Modèle de vérification grammaticale automatique gauche-droite

Thèse soutenue publiquement le **12 décembre 2014**,
devant le jury composé de :

M. Thomas LEBARBÉ
Professeur, Université Stendhal - Grenoble 3, Directeur de thèse
Mme Cécile FABRE
Professeur, Université Toulouse 2 - Le Mirail, Rapporteur
M. Geoffrey WILLIAMS
Professeur, Université de Bretagne Sud, Président
M. Olivier KRAIF
Maître de Conférences, Université Stendhal - Grenoble 3, Examinateur

Université Joseph Fourier / Université Pierre Mendès France /
Université Stendhal / Université de Savoie / Grenoble INP

Remerciements

Je tiens tout d'abord à remercier mon directeur de thèse, Thomas Lebarbé, de m'avoir fait confiance pour mener cette thèse à bien et surtout à son terme.

Je voudrais ensuite remercier Cécile Fabre et Geoffrey Williams d'avoir accepté d'évaluer mon travail et Olivier Kraif d'avoir accepté de faire partie de mon jury.

Un immensissime (c'est moche mais j'aime bien ! !) merci également à Bad Cop et Good Cop, sans le coaching desquels cette thèse n'aurait jamais connu de fin.

J'aimerais aussi dire, dans le désordre :

Un grand merci aux copines de galère, de poucave et de refaisage de monde dans le bureau, Aïcha et Paulette ;

Un grand merci aux copains des pauses du matin, du midi et de l'aprés-midi, pour les discussions surnaturelles auxquelles elles donnent généralement lieu : Aïcha, Mathieu, Alexia, Thomas, Ninie, Claude, Aurélie, Bubu, Cristelle, les 2 Isa, Monmon, Tiphaine, Vannina, Hoaï, Arno, Sylvain, Paulette, Alex, Lucie, Eleni, et sûrement d'autres que j'oublie ;

Un grand merci à la fine équipe de CEDIL2010, la petite Isa, Aïcha, Auriane et Tiphaine... Trop bonne expérience que ce colloque avec vous les filles !

Un grand merci aux collègues du DIP et en particulier Maman Roseline pour sa gentillesse (même si elle fait des pouet-pouet quand on dit des gros mots) et Ninie, Thomas, Mathieu, Alexia pour leurs facéties ;

Un grand merci à Gégé, qui m'a offert l'asile quotidien et m'a nourrie, toujours le ~~grognement~~ sourire aux lèvres ! Et P'tit mouton, toujours de bonne humeur mais qui chante comme une casserole ;

Un grand merci à mes relecteurs attentifs et parfois psychopathes de la virgule : Paulette (AINSI !), Ninie (on ne se base pas sa mère), Tiphaine, Vannina, Alex et la grande Isa ;

Un grand merci à Math, pour ses jeux de piste dans ses relectures et ses méta-commentaires désopilants. Tu dois avoir un sacré stock de pincettes ! ! ;

Un grand merci à ma Lucy pour ses conseils en traduction ;

Un grand merci à Thomas (et Sandrine), Gégé et leur chat caractériel respectif pour les résidences d'été studieux ;

Un grand merci à Cécile et Élisabeth pour m'avoir accueillie à la MSH ;

Un grand merci à la communauté d'OpenOffice.org, et Laurent Godard, qui m'ont par hasard conduite à cette thèse ;

Un grand merci à ma famille pour avoir composé avec la rareté de mes visites ;

Un grand merci aux étudiants qui ont réalisé la dictée et les auteurs des mails, des résumés et des commentaires de blog qui m'ont fourni la matière première de mon travail ;

Un grand merci enfin à tous ceux que j'aurais involontairement oubliés...

Cordialement ! ;-)

Table des matières

Liste des figures **vii**

Liste des tableaux **ix**

Conventions **1**

Préambule **3**

Partie I Problématique **5**

1 Approche linguistique de la notion d'erreur de grammaire **7**

1.1 Interprétation de la notion de grammaire 7

1.1.1 La grammaire 7

1.1.2 L'orthographe 15

1.1.3 Conclusion 19

1.2 Définitions de l'erreur et de la faute 19

1.2.1 Définitions générales 20

1.2.2 Erreur et faute en didactique des langues 20

1.2.3 Précisions terminologiques 22

2 Interprétation informatique de l'erreur de grammaire **25**

2.1 Mécanismes de gestion des erreurs tapuscrites 25

2.1.1 La vérification orthographique 26

2.1.2 La vérification grammaticale 27
2.2 De la grammaire académique à la grammaire en bureautique 35
2.2.1 La grammaire des outils bureautiques 35
2.2.2 Les types d'erreurs 36
2.2.3 Conclusion 37

3 Etat des lieux des outils et études sur les erreurs 39
3.1 Documentation et fonctionnement des vérificateurs existants 39
3.1.1 Des outils très peu documentés 40
3.1.2 Un fonctionnement limité 41
3.1.3 Les utilisateurs livrés à eux-mêmes 45
3.2 Panorama des études sur les erreurs tapuscrites 53
3.2.1 Les études existantes 54
3.2.2 Spécificité des tapuscrits 57

Partie II Caractérisation des erreurs tapuscrites 61

4 Choix d'une approche corpus 63
4.1 Justification d'une approche corpus 64
4.1.1 Définition de la notion de corpus 64
4.1.2 Les corpus disponibles 67
4.2 Méthodologie de constitution du corpus 68
4.2.1 Caractéristiques communes des données 68
4.2.2 Variété des scripteurs 69
4.2.3 Variété des situations de scription 70
4.2.4 Variété des types de documents 71
4.3 Caractérisation du corpus de l'étude 72
4.3.1 Écueils de la collecte des textes 72

4.3.2 Contenu du corpus et représentativité 73
4.3.3 Positionnement de notre corpus 75

5 Constitution du corpus 77

5.1 Recueil des textes 77
5.1.1 Dictées 77
5.1.2 Résumés 79
5.1.3 Courriers électroniques 80
5.1.4 Commentaires de blog 80
5.2 Normalisation des données 81
5.2.1 Stockage homogène des données 81
5.2.2 Standards d'annotation : XML, TEI, CES 83
5.2.3 Normalisation des données 86

6 Annotation et analyse des erreurs 89

6.1 Typologies des erreurs et annotation descriptive 89
6.1.1 Adaptation de typologies existantes 89
6.1.2 Balisage du corpus 100
6.1.3 Réajustements de la typologie 103
6.2 Analyse quantitative des erreurs 109
6.2.1 Traitements statistiques des données 109
6.2.2 Description quantitative du corpus 112
6.3 Résumé des principaux résultats 136

Partie III Modélisation de la vérification grammaticale 139

7 Modélisation de la production et de la détection humaine des erreurs 141

7.1 La production du langage écrit 142
7.1.1 Les processus cognitifs mis en œuvre 142

7.1.2 La production d'erreurs dans le corpus 147
7.2 Révision du langage écrit 160
7.2.1 Le processus de révision 160
7.2.2 Hypothèses sur la manière de détecter une erreur 163
7.3 Conclusion 170

8 Proposition d'un modèle pour la vérification grammaticale 175
8.1 Structure du modèle 176
8.1.1 Mécanisme de lecture gauche-droite 176
8.1.2 Étiquetage morphosyntaxique 178
8.1.3 Segmentation en *chunks* 179
8.2 Des attentes aux piles 183
8.2.1 Les valences de Tesnière 183
8.2.2 Les actants de Mel'čuk 186
8.2.3 Des attentes de différents niveaux 186
8.2.4 Un traitement par piles 188
8.2.5 Contenu des piles 189
8.2.6 Portée des attentes 190
8.3 Ressources 192
8.3.1 Règles de segmentation en chunks 193
8.3.2 Ressources pour les attentes 193
8.4 Fonctionnement attendu 197
8.4.1 Exemple de détection par une attente non comblée 197
8.4.2 Exemples de détection par un échec d'unification 200
8.4.3 Rétroactions 201
8.5 Conclusion 203

Conclusion et perspectives **205**

Perspectives d'implantation du modèle **207**

1 Un système multi-agents 208

2 Limitations du modèle 211

2.1 Complexité de la détection de certaines erreurs 211

2.2 Des ressources complexes à élaborer 213

3 La question des rétroactions explicites contextuelles 214

3.1 Quel contenu ? 214

3.2 Quelle représentation ? 216

3.3 Prise en compte de la décision de l'utilisateur 217

4 Conclusion 217

Acronymes **221**

Bibliographie **223**

Annexes **239**

A Tableaux de données **241**

Liste des figures

1.1 Arbre syntagmatique 12
1.2 Arbre de dépendance 12
1.3 Exemple de structure de traits (issu de Bouillon [1998, p. 91]) 12

2.1 Structure générale en couche d'un correcteur grammatical 28
2.2 Structure générale d'un correcteur grammatical effectuant un *chunking* 33
2.3 Chevauchement de la grammaire et l'orthographe, en linguistique et en informatique. 36

3.1 Exemples de rétroactions du logiciel Antidote (énoncé (5)) 46
3.2 Exemples de rétroactions du logiciel Antidote (énoncé (6)) 47
3.3 Exemple de rétroaction de Word (énoncé (5)) 47
3.4 Exemple de rétroaction de BonPatron (énoncé (5)) 49
3.5 Exemple de rétroaction de BonPatron (énoncé (6)) 49
3.6 Exemple de rétroaction de LanguageTool (énoncé (5)) 51
3.7 Exemple de rétroaction de LanguageTool (énoncé (6)) 51
3.8 Exemple de rétroaction de Grammalecte (énoncé (5)) 52
3.9 Exemple de rétroaction de Grammalecte (énoncé (6)) 52

5.1 Structure du fichier *eXtensive Mark-up Language* (XML) 87

6.1 Typologie d'erreurs d'après Debyser *et al.* [1967] 92
6.2 Typologie d'erreurs d'après Catach *et al.* [1980] 93
6.3 Typologie d'erreurs d'après Jacquet-Pfau [2001] 93
6.4 Typologie d'erreurs d'après Lucci & Millet [1994] 94
6.5 Typologie d'erreurs d'après Granger [2007] 94
6.6 Synthèse de la première typologie d'erreurs 98
6.7 Exemple d'annotation de FRIDA-bis [Antoniadis *et al.*, 2010] 100

6.8 Exemple de balisage d'un texte . . . 101
6.9 Évolutions de la catégorie `ORTHO` . . . 103
6.10 Évolutions de la catégorie `SYNTAXE` . . . 104
6.11 Évolutions de la catégorie `ACCORD` . . . 104
6.12 Évolutions des catégories `VERBE`, `LEXIQUE` et `PONCTUATION` . . . 105
6.13 Synthèse de la typologie d'erreurs finale . . . 107
6.14 Pourcentages moyens des catégories d'erreurs dans les différents types de textes . 117

7.1 Modèle de production d'écrit selon Hayes & Flower [1980], dans sa version clarifiée [Hayes, 1996] . . . 143
7.2 Les composants de la mémoire de travail d'après Baddeley [2000]. (En gris : les composants « cristallisés » de la mémoire à long terme ; en blanc : les composants flexibles de la mémoire de travail) . . . 145
7.3 Modèle de production d'écrit d'après Kellogg [1996] (notre traduction) . . . 146
7.4 Répartition des erreurs selon les causes *Humain*, *Texte* et *Situation* . . . 154
7.5 Pourcentages d'erreurs dans les ensembles et sous-ensembles de causes . . . 155
7.6 Répartition des catégories d'erreurs selon les causes *Humain*, *Texte* et *Situation* . 157
7.7 Répartition des attentes selon les catégories d'erreurs . . . 169

8.1 Modèle pour la vérification grammaticale automatique . . . 177
8.2 Arbre syntagmatique . . . 184
8.3 Arbre de dépendance . . . 184
8.4 Exemple de structure de traits pour un nom masculin singulier . . . 191
8.5 Exemples de structure de traits pour le déterminant *des* . . . 191
8.6 Exemple d'unification qui réussit . . . 191
8.7 Exemple d'unification qui échoue . . . 192
8.8 Une entrée de DICOVALENCE (`bricoler`) . . . 196

1 Système multi-agents pour la vérification grammaticale . . . 209
2 Exemple de rétroaction graphique . . . 216

A.1 Proportion des catégories d'erreur dans les différents types de textes. Moyennes calculées à partir des effectifs totaux. . . . 244

Liste des tableaux

3.1 Tableau récapitulatif des différentes générations de correcteurs 42

5.1 Synthèse des textes du corpus 81

6.1 Première typologie d'erreurs 99
6.2 Typologie d'erreurs finale 108
6.3 Résultats de la transformation arc sinus sur les pourcentages moyens d'erreurs par catégorie dans le corpus. 110
6.4 Description du corpus 113
6.5 Pourcentages moyens d'erreurs dans chaque catégorie au sein du corpus (ET) [1] . 115
6.6 Pourcentages moyens des catégories d'erreurs dans les différents types de textes . 117
6.7 Proportions moyennes (ET) des erreurs de la catégorie `ACCORD` dans chaque type de textes 119
6.8 Proportion moyenne (ET) de chaque catégorie d'erreurs de grammaire dans chaque type de textes 123
6.9 Densité d'erreurs dans chaque catégorie et chaque type de texte 125
6.10 Proportions moyennes des erreurs à caractère homophone 128
6.11 Effectifs d'erreurs par distances en mots en fonction de la catégorie d'erreur . . . 133
6.12 Effectifs d'erreurs par distances en syntagmes en fonction de la catégorie d'erreur 133

7.1 Classement des causes possibles d'erreurs de grammaire 149
7.2 Classement des types d'attentes 166

8.1 Extrait de définition et tableau de régime de l'acception I de la lexie « Outil » [Mel'čuk, 1999] 194

A.1 Effectifs d'erreurs par (sous-)catégories et types de textes 242
A.2 Effectifs d'erreurs à caractère homophone 243
A.3 Répartition des catégories d'erreurs par type de causes 243

A.4 Proportion moyenne (ET) des catégories d'erreur dans les différents types de textes. Moyennes calculées à partir des effectifs totaux 244

A.5 Proportion moyenne (ET) des erreurs de la catégorie `LEXIQUE` dans chaque type de textes 245

A.6 Proportion moyenne (ET) des erreurs de la catégorie `SYNTAXE` dans chaque type de textes 245

A.7 Proportion moyenne (ET) des erreurs de la catégorie `VERBE` dans chaque type de textes 245

A.8 Proportion moyenne (ET) des erreurs de la catégorie `PONCTUATION` dans chaque type de textes 245

A.9 Proportion moyenne (ET) des erreurs de la catégorie `ORTHOGRAPHE` dans chaque type de textes 246

A.10 Densité d'erreurs calculées ($\frac{nb\ d'erreurs}{nb\ de\ mots}$) 247

Conventions

Dans ce manuscrit nous avons opté pour un certain nombre de conventions d'écriture et de mise en forme.

1 Traductions

Les citations dans une langue étrangère sont traduites ou glosées. Pour ne pas faire perdre de temps aux lecteurs qui n'ont que faire d'une traduction (souvent imparfaite), nous avons grisé toutes les notes de bas de page qui les concernent. Une note de bas de page sera donc indiquée ainsi [2], alors qu'une traduction sera indiquée de cette manière [3]. Quand l'auteur de la traduction n'est pas indiqué, cela signifie qu'il s'agit de notre propre traduction.

2 Exemples d'erreurs

Nous donnons de nombreux exemples d'erreurs d'orthographe ou de grammaire tout au long de ce manuscrit. La plupart de ces exemples sont issus du corpus d'erreurs que nous avons constitué. Lorsque les exemples proviennent d'autres travaux, nous le précisons explicitement.

Par ailleurs, comme il est d'usage dans la littérature de faire précéder les énoncés agrammaticaux d'un astérisque, nous procèderons de même. Ainsi, tout énoncé contenant une erreur sera précédé d'un *.

3 Orthographe utilisée

Cette thèse ne suit pas les rectifications de l'orthographe parues dans « Documents administratifs » au Journal Officiel du 6 décembre 1990.

2. Une note de bas de page.
3. Une traduction. (traduit par auteur de la traduction).

Préambule

À l'ère du numérique et des contenus textuels essentiellement tapuscrits, la vérification linguistique automatique des textes est devenue un véritable enjeu, d'autant plus que les fautes d'orthographe et de grammaire sont souvent stigmatisées et discréditent leurs auteurs. Il s'avère alors souvent indispensable de recourir à des outils pour vérifier l'orthographe, la grammaire, et parfois le style.

La vérification orthographique constitue le b.a.-ba de la vérification linguistique. Elle est aujourd'hui souvent intégrée directement aux logiciels de rédaction, voire aux systèmes d'exploitation. La grammaire est en revanche plus complexe à vérifier et constitue un défi pour les outils dédiés, souvent indépendants des logiciels de rédaction. Si les logiciels commerciaux de vérification proposés par les industriels sont aujourd'hui relativement efficaces, mais très peu documenté, les outils librement distribués sont connus pour leurs résultats limités pour la correction du français [Souque, 2007]. Leur principale limite tient au fait qu'ils se fondent sur des énumérations exhaustives d'erreurs potentielles, en comparant des portions de texte à des modèles décrits dans des règles locales. Ceci implique de construire un nombre considérable de motifs pour tenter de décrire un maximum de phénomènes possibles, avec, entre autres, le risque de d'ignorer des erreurs s'il manque une règle et de détecter de fausses erreurs si les règles sont redondantes.

L'objectif de ce travail est de proposer un modèle de vérification grammaticale alternatif pour améliorer les performances des outils libres. Il se situe dans un cadre interdisciplinaire à l'intersection de la linguistique et de l'informatique. En effet, le traitement efficace de la langue nécessite de modéliser celle-ci de façon cohérente avec les phénomènes linguistiques analysés, tout en permettant au système informatique de traiter les informations associées.

Le modèle que nous proposons réalise non pas une analyse phrase par phrase comme le font les outils libres existants, mais une analyse gauche-droite, mot par mot, au fur et à mesure de la lecture/écriture. Il s'inspire des grammaires de dépendance et repose sur un principe d'attentes qui permet la détection d'incohérences grammaticales en ayant recours à la segmentation en syntagmes minimaux et au principe d'unification.

Ce travail prend sa source dans la continuité de notre participation au projet francophone libre OpenOffice.org [4], qui fédère une large communauté d'utilisateurs et de contributeurs autour de la suite bureautique libre OpenOffice.org, dont « le but énoncé est d'offrir une alternative à la suite bureautique propriétaire Microsoft Office [5] » [Wikipédia, 2014].

Dans ce contexte, nous avons, préalablement à la thèse, développé un outil destiné au cor-

4. Depuis 2011, le projet initial s'est scindé en deux : Apache OpenOffice et LibreOffice. Cette scission ayant eu lieu après nos travaux, nous n'en tenons pas compte dans ce manuscrit et mentionnons le projet initial OpenOffice.org

5. Développée par Microsoft Corporation.

recteur orthographique d'OpenOffice.org qui permet d'extraire automatiquement les affixes des langues pour la génération automatique des ressources nécessaires à la correction orthographique [Souque, 2006]. La présente recherche, bien que prolongeant les travaux réalisés dans la communauté du logiciel libre, ne s'y inscrit pas. Nous nous centrons ici sur la correction grammaticale qui fait l'objet d'une forte demande de la part de la communauté des nombreux utilisateurs de logiciels libres.

Cette thèse s'organise en trois parties. La partie introductive (partie 1) présente notre problématique et les champs disciplinaires desquels elle s'alimente, la seconde est consacrée à l'étude d'erreurs sur corpus et la troisième à la modélisation de la détection d'erreurs.

Dans la première partie, nous nous attacherons à définir les contours de notre recherche. Après avoir défini les notions de *grammaire*, d'*orthographe*, de *faute* et d'*erreur* d'un point de vue linguistique (chapitre 1) et informatique (chapitre 2), nous ferons un état des lieux des outils de vérification grammaticale et des études disponibles sur les erreurs (chapitre 3).

La partie 2 sera consacrée à notre corpus d'erreurs tapuscrites. Dans le chapitre 4, nous aborderons la notion de *corpus* et les raisons qui nous ont conduite à faire le choix de cette approche fondée sur l'authenticité des écrits pour étudier les erreurs. Le chapitre 5 présentera de manière détaillée les différents types d'écrits que nous avons recueillis. Enfin, dans le chapitre 6, nous expliciterons la typologie que nous avons adoptée pour l'annotation du corpus et rendrons compte de l'analyse des erreurs.

Dans la partie 3, l'étude des processus cognitifs impliqués dans la production d'écrits nous permettra de comprendre et d'émettre des hypothèses sur les causes possibles d'erreurs et sur la manière de détecter les erreurs lors du processus de révision (chapitre 7). Cette étude constitue les fondements du modèle de détection d'incohérences grammaticales que nous présenterons dans le chapitre 8 et dont nous discuterons les implications dans le chapitre 9.

Première partie

Problématique

Chapitre 1

Approche linguistique de la notion d'erreur de grammaire

Sommaire

1.1 Interprétation de la notion de grammaire 7
1.1.1 La grammaire 7
1.1.2 L'orthographe 15
1.1.3 Conclusion 19
1.2 Définitions de l'erreur et de la faute 19
1.2.1 Définitions générales 20
1.2.2 Erreur et faute en didactique des langues 20
1.2.3 Précisions terminologiques 22

Notre recherche porte sur la correction automatique des erreurs de grammaire, ce qui induit nécessairement un travail préalable de définition linguistique des notions d'« erreur » et de « grammaire » qui vont jalonner cette thèse, avant d'aborder l'aspect informatique. Nous débuterons donc cette première partie avec la notion de « grammaire », telle qu'elle est généralement définie, et telle que nous l'envisageons dans la suite de notre mémoire, en contraste avec l'orthographe dont elle est difficilement dissociable. Nous nous intéresserons ensuite à la notion d'« erreur », afin de pouvoir définir et délimiter plus précisément notre objet d'étude, à savoir les erreurs de grammaire.

1.1 Interprétation de la notion de grammaire

1.1.1 La grammaire

Le mot « grammaire » évoque souvent les leçons, parfois rébarbatives, qui rythment chaque vie d'écolier. « À l'école, faire de la grammaire, c'est apprendre les principes qui régissent l'organisation de la phrase (règles de placement des mots dans une phrase interrogative ou affirmative, règles d'accord des verbes, etc.) » [Dortier, 2004, p. 285]. Le terme de « grammaire » renvoie ainsi communément à la discipline normative et pédagogique qui vise à enseigner les règles de fonctionnement de la langue. Mais la lecture de divers ouvrages révèle des définitions très variées.

La grammaire peut être présentée comme l'ensemble des règles de la langue, comme l'ouvrage qui les contient, ou de manière plus globale comme l'« étude systématique des éléments constitutifs et du fonctionnement [...] de la langue » [Grevisse, 1993, p. 4]. Elle est également parfois définie comme une discipline, ou encore comme un art. Les définitions mentionnent également les notions de linguistique, ainsi que divers types de grammaires (générale, historique, comparée, formelle, etc.). Un retour sur l'origine de la grammaire et son évolution nous permet dans les pages suivantes de clarifier ces différents points de vue, puis de nous positionner plus précisément par rapport à eux.

a) Origine de la grammaire : de l'art à la discipline

Dans la Grèce antique

L'origine historique de notre grammaire se situe dans l'Antiquité grecque, vers le Ve siècle avant J.C., et est intimement liée à la philosophie. La nécessité de rendre compte de la réalité avec des énoncés corrects a en effet conduit aux premières réflexions sur la langue.

> « Une telle tâche liait alors indissolublement la philosophie à une certaine approche du langage qui, loin de l'envisager comme l'objet d'une discipline séparée, et sans même l'envisager comme objet, visait à le normer pour constituer à partir de lui un medium discursif de notre rapport à la réalité, l'énoncé droit ou l'énoncé correct [...]. »
>
> [Ildefonse, 1997, p. 15]

Platon, un des premiers philosophes à réaliser des analyses du langage, parlait de l'« art grammatical » pour désigner l'art de l'assemblage des lettres. Il s'agit-là d'une définition très proche du sens étymologique indiqué aujourd'hui pour le mot « grammaire » [1] : « art de lire et d'écrire les lettres » [Robert, 2006]. Il ne faut cependant pas le confondre avec la calligraphie [2], l'« art de bien former les caractères d'écriture » [Robert, 2006].

L'ouvrage d'Ildefonse [1997] nous indique que Platon fut le premier à établir la distinction entre le nom et le verbe, initiant ainsi la notion de parties du discours que nous connaissons aujourd'hui, notion étoffée ensuite par son disciple Aristote. Puis l'école stoïcienne a marqué un tournant dans l'analyse du langage, dès le IIIe siècle avant J.C., en produisant de nombreux écrits sur des sujets grammaticaux, dont nous utilisons toujours une partie de la terminologie aujourd'hui. Les théories stoïciennes ont dès lors considérablement influencé les philologues alexandrins sur les questions grammaticales, ces derniers se revendiquant d'ailleurs comme des grammairiens. Ildefonse [1997] reprend l'hypothèse formulée par Di Benedetto [1958] selon laquelle le IIIe siècle avant J.C avait ainsi vu naître une « discipline autonome », la philologie. Son objet d'étude, qui consistait à l'origine essentiellement en des poèmes homériques, s'est étendu au 1er siècle avant J.C. « à l'analyse de la langue grecque et de ses éléments constitutifs », pour prendre « une tournure proprement linguistique » [Di Benedetto, 1958, p. 23]. La grammaire s'émancipe ainsi peu à peu de la philosophie puis de la philologie, pour se rapprocher de la discipline que nous connaissons aujourd'hui, plus technique. La *Technè Grammatikè*, attribuée au philologue Denys Le Thrace et datée de cette époque, serait le premier manuel de grammaire, bien que son origine et son authenticité soient contestées. Divisé en deux parties, l'ouvrage présente la grammaire comme « la connaissance empirique de ce qui se dit couramment chez les poètes et les prosateurs » [3]. Elle est en effet conçue, dans la première partie, comme « une activité appliquée, qui

1. grammaire : du latin *grammatica*, issu du grec *grammatikê* « art de lire et d'écrire les lettres ».
2. calligraphie : du grec *kállos* « beauté » et *graphein* « écrire ».
3. Denys Le Thrace, *Technè Grammatikè*, 1, trad. [Lallot, 1998, p. 43].

a pour objet le texte, en particulier le texte poétique » [Lallot, 1998, p. 73]. La seconde partie est en revanche moins empirique. Elle « s'éloigne du *texte* et se donne pour objet la *langue* » [Babu, 2007, p.24], en se focalisant sur les huit parties du discours : nom, verbe, participe, article, pronom, préposition, adverbe et conjonction.

Apollonius Dyscole, également à l'origine de travaux déterminants au IIe siècle, vient compléter ceux de Denys Le Thrace, avec la syntaxe notamment. Dans son traité *Peri suntaxeôs* (De la construction), il traite, en quatre livres, du statut syntaxique de l'article, du pronom, du verbe, de la proposition et de l'adverbe. Il y appréhende la syntaxe par les parties du discours. Pour lui, les mots respectent un ordre similaire aux autres éléments de différents niveaux du langage (sons, lettres, syllabes, mots et phrases), ce que reprend Luhtala [2005, p. 80] : « *Just as sounds cannot be joined in a random fashion to form syllables, syllables must in turn be joined appropriately to form words, and words to form sentences.* [4] ». Cette approche de la syntaxe proposée par Apollonius Dyscole « ne fait nullement intervenir la notion de *fonction* syntaxique » [Swiggers & Wouters, 2001, p. 186], ni celle de rapports entre les syntagmes.

Chez les Latins

La discipline grammaticale née dans la Grèce antique s'est ensuite transmise à Rome au début de notre ère, puis aux pays occidentaux. Le premier ouvrage notoire sur la grammaire latine est probablement *De institutione oratoria* (De l'institution oratoire) de Quintilien, rhéteur latin du 1er siècle après J.C., qui fut rédigé dans un but pédagogique, pour l'enseignement de la rhétorique aux futurs orateurs. Un chapitre du premier livre (l'oeuvre contient douze livres) est consacré à la grammaire, que l'auteur divise en deux parties : « science du dire correct et commentaire des poètes » [Ildefonse, 1997, p. 16]. La préoccupation est encore alors, comme pour les Grecs, de s'exprimer correctement et de pouvoir étudier et commenter les oeuvres littéraires classiques.

Au IVe siècle, Donat, grammairien romain, rédige *Ars grammatica*, manuel de grammaire qui fut une référence durant toute la période moyenâgeuse, aussi bien dans l'enseignement que dans l'inspiration pour d'autres manuels de grammaire latine ou parfois française [5]. Il est divisé en deux parties : l'*Ars Minor*, destiné aux débutants, ne traite que des parties du discours ; L'*Ars Major*, consacré aux étudiants avancés, est beaucoup plus complet. Il contient trois parties, qui abordent la phonétique, la morphologie et la stylistique, mais le système de construction des énoncés, à savoir la syntaxe est absent.

Grondeux [2000, p. 600] indique que Priscien est « le seul, parmi les grammairiens légués par la Rome antique au Moyen Age », à aborder la syntaxe. Il est l'auteur, à la fin de l'Antiquité, d'un ouvrage de référence dans l'enseignement de la grammaire jusqu'à la Renaissance : *Institutiones grammaticae*. Il y reprend les descriptions syntaxiques de ses prédécesseurs grecs, dont Apollonius Dyscole principalement, et prépare « la voie à une analyse qui s'intéresse aux rapports de rection et de dépendance dans la structure phrastique » [Swiggers & Wouters, 2001, p. 188], mais il ne va cependant pas plus loin qu'Apollonius Dyscole dans sa théorisation.

Finalement, depuis l'Antiquité jusqu'à la Renaissance, la grammaire constitue un des sept arts libéraux de l'enseignement et forme, avec la dialectique et la rhétorique, le *trivium*, qui représente les trois arts. Les quatre autres arts libéraux, considérés comme les quatre sciences, sont constitués de l'arithmétique, la géométrie, l'astronomie et la musique, et forment le *quadrivium*.

4. Tout comme les sons ne peuvent pas être reliés de manière aléatoire pour former les syllabes, les syllabes doivent à leur tour être reliées de manière appropriée pour former les mots, et les mots pour former les phrases.

5. La plus ancienne grammaire connue du français, publiée vers 1409, est le *Donait françois*, ouvrage de perfectionnement en français destiné aux Anglais et restreint à l'Angleterre.

À la Renaissance

Les premières véritables grammaires du français [Palsgrave, 1530 ; Meigret, 1550 ; Nicot, 1606], c'est-à-dire les premiers ouvrages rassemblant les règles de la grammaire, datent de la Renaissance, époque à laquelle le latin, langue officielle, a commencé à prendre du recul par rapport au français, langue vernaculaire. Lorsque cette dernière est devenue langue officielle en 1539 avec l'ordonnance de Villers-Cotterêts [6] , c'est le français de l'élite sociale qui a été pris comme référence du « bon usage » de la langue, c'est-à-dire l'usage correct, et qui a été décrit dans les grammaires. C'est ainsi que l'Académie Française, fondée en 1635 par Richelieu, a fixé les règles de ce bon usage du français, règles qui constituent aujourd'hui la norme, la grammaire enseignée à l'école, mais qu'elle n'a publiées que très tardivement ([Académie Française, 1932]).

Ce rapide historique nous montre que la grammaire, depuis son origine et au moins jusqu'à la Renaissance, est considérée comme un art, l'art de bien parler, bien lire, bien écrire. Nées du besoin des philosophes d'exprimer la pensée avec des énoncés corrects, les réflexions sur le langage se sont peu à peu affinées et émancipées pour finalement constituer une discipline à part entière, la grammaire, dédiée à l'étude de la langue en tant qu'objet et non plus en tant que « simple » medium pour s'exprimer correctement ou commenter les écrits littéraires classiques. La grammaire a par ailleurs dès le début fait partie intégrante de l'enseignement, avant même de devenir une discipline autonome, et occupe toujours aujourd'hui une place importante dans l'éducation. Mais si les notions de « correction » et de « bien lire » ou « bien écrire » sont toujours présentes, la grammaire avait dans les premiers temps principalement une fonction descriptive de la langue, dans le but d'étudier et commenter les oeuvres classiques. Elle est ensuite devenue de plus en plus normative, en fixant des règles pour enseigner la manière correcte de parler et écrire, et en les rassemblant dans des manuels également appelés « grammaires », par métonymie, et généralement rédigés à des fins pédagogiques.

b) Grammaire « moderne » et linguistique

Liens entre les deux disciplines

La grammaire est donc une discipline très ancienne, vouée initialement à l'étude de la langue, qui a mené à la détermination de ses éléments constitutifs, ainsi qu'à ses règles de fonctionnement, et qui a toujours constitué un des principaux enseignements à l'école. Aujourd'hui, le sens le plus courant (premier sens dans les dictionnaires) du terme « grammaire » en français désigne ainsi l'ensemble des règles du bon usage (la norme), et également de manière souvent indissociable la discipline scolaire dont il est l'objet, ainsi que le manuel qui les contient. En deuxième ou troisième sens, les dictionnaires définissent également bien la grammaire comme l'étude des structures et du fonctionnement de la langue, ce qu'était donc la grammaire à ses débuts. Mais cette acception génère aujourd'hui une confusion avec la linguistique, discipline récente, apparue au XIX^{e} siècle, dont l'objet d'étude, tout comme la grammaire, est la langue. « Linguistique » et « grammaire » sont ainsi souvent employés de manière synonyme, notamment dans l'ouvrage de Riegel *et al.* [1994]. Mais ils sont également fréquemment mis en opposition, avec par exemple le caractère traditionnel de la grammaire face à la modernité de la linguistique, mais aussi et surtout par le rôle prescriptif associé à la grammaire qui légifère sur le bon usage, contrairement à la fonction descriptive de la linguistique qui étudie tous les usages de la langue. La linguistique est en effet définie communément comme la « science qui a pour objet l'étude du langage et des langues »

6. Ordonnance de Villers-Cotterêts établie par François 1er le 15 août 1539

[Larousse, 2006], ou encore comme ayant « pour unique et véritable objet la langue envisagée en elle-même et pour elle-même » [de Saussure, 1916, p. 317]. Comme le précise Gary-Prieur [1985], elle se distingue ainsi encore de la grammaire puisque celle-ci étudie la langue non pas pour approfondir la connaissance de la langue en elle-même, mais avec des objectifs extérieurs : pédagogiques, philosophiques ou philologiques.

Il apparaît finalement que la linguistique moderne est l'héritière de la grammaire antique et de sa fonction d'étude descriptive du langage. « Il ne faut pas minimiser la filiation de l'une à l'autre : la linguistique bénéficie de tous les travaux des grammairiens sur les langues, et ne s'en coupe pas » [Gary-Prieur, 1985, p. 11]. Toutes deux ont permis (et permettent encore) de dégager entre autre les éléments constitutifs et le fonctionnement de la langue, dont ceux jugés corrects pour une langue donnée sont consignés dans les manuels de grammaire et enseignés aux écoliers via la grammaire « moderne ». La fonction de cette dernière s'est donc un peu éloignée de sa fonction d'origine, en se consacrant au bon usage de la langue, à des fins pédagogiques.

Pas une, mais des grammaires

Il est fréquent de trouver accolés au mot grammaire divers adjectifs, qui créent alors autant de locutions désignant divers courants théoriques. Nous ne mentionnons ici que les celles qui nous paraissent revenir le plus souvent dans les différents ouvrages de linguistique ou dictionnaires que nous avons consultés [Dubois *et al.*, 1994 ; Ducrot & Schaeffer, 1972 ; Fuchs, 1996 ; Larousse, 2006 ; Riegel *et al.*, 1994 ; Robert, 2006].

La « grammaire générale », par exemple, suit la théorie selon laquelle toutes les langues ont en commun des éléments et des procédés, qu'elle tente de mettre en évidence. Cette théorie se fonde sur le postulat que la pensée, exprimée par le langage, suit des schémas logiques qui sont universels. La « grammaire comparée », quant à elle, vise à dégager des affinités ou la parenté entre deux ou plusieurs langues en les comparant. Elle étudie parfois une seule et même langue dans son évolution en réalisant des comparaisons de différents stades de son histoire. Elle se confond alors avec la « grammaire historique » qui s'occupe de l'étude des langues en diachronie.

On rencontre également très fréquemment la notion de « grammaire formelle ». Sous cette dénomination sont en fait regroupées diverses grammaires dont l'objet est de décrire et formaliser les langues naturelles au moyen de modèles explicites, et qui sont notamment bien adaptées au Traitement Automatique des Langues (TAL). Parmi elles, les linguistes distinguent les « grammaires génératives » et les « grammaires de dépendances ». Les premières représentent la syntaxe d'une phrase à l'aide d'arbres syntagmatiques, alors que les secondes utilisent des arbres de dépendance. Les figures 1.1 et 1.2, page suivante, donnent un exemple de représentation syntaxique d'un même énoncé selon chacun des deux types de grammaires.

Le courant générativiste est né des travaux de Chomsky et de la « grammaire générative et transformationnelle » [Chomsky, 1965]. Les descriptions syntaxiques y sont réalisées au moyen de règles de réécriture (ou syntagmatiques), qui permettent de générer n'importe quel énoncé, à partir d'un nombre fini d'éléments. L'arbre syntagmatique en figure 1.1 permet de visualiser ces règles. Par exemple, un `Syntagme Nominal` (`SN`) est réécrit `Déterminant Nom` (`D N`)[7] avec éventuellement un `Syntagme Adjectival` (`SA`)[8]. La règle est alors formalisée de la manière suivante : `SN → D N (SA)`.

7. réalisé dans le sous-arbre `Ce film`.

8. réalisé dans le sous-arbre `une histoire très émouvante`. Le `Syntagme Adjectival` (`SA`) sera lui-même réécrit selon d'autres règles, comme celle qui a permis la formation du syntagme `très émouvante`.

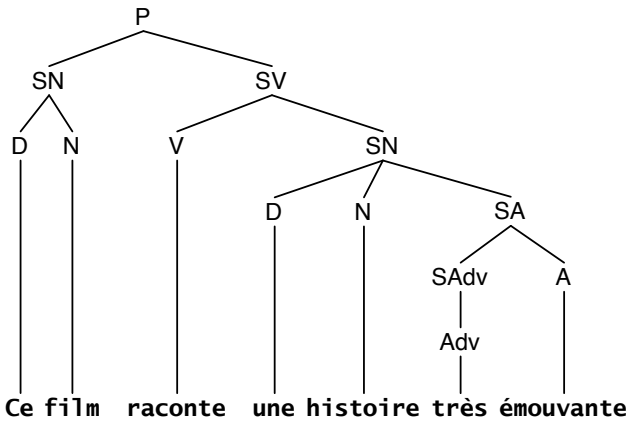


FIGURE 1.1 : Arbre syntagmatique

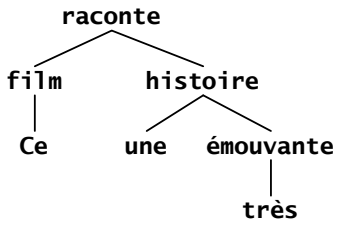


FIGURE 1.2 : Arbre de dépendance

$$\begin{bmatrix} \text{phon} = \langle \text{le garçon mange le chocolat} \rangle \\ \text{sujet} \begin{bmatrix} \text{phon} = \langle \text{le garçon} \rangle \\ \text{accord [1]} \begin{bmatrix} \text{genre} = \text{masc} \\ \text{nombre} = \text{sing} \end{bmatrix} \\ \text{DET} \begin{bmatrix} \text{phon} = \langle \text{le} \rangle \\ \text{accord [1]} \begin{bmatrix} \text{genre} = \text{masc} \\ \text{nombre} = \text{sing} \end{bmatrix} \end{bmatrix} \\ \text{N} \begin{bmatrix} \text{phon} = \langle \text{garçon} \rangle \\ \text{accord [1]} \begin{bmatrix} \text{genre} = \text{masc} \\ \text{nombre} = \text{sing} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{verbe} \begin{bmatrix} \text{phon} = \langle \text{mange} \rangle \\ \text{accord [1]} \begin{bmatrix} \text{nombre} = \text{sing} \\ \text{personne} = 3 \end{bmatrix} \end{bmatrix} \\ \text{objet} \begin{bmatrix} \text{phon} = \langle \text{le chocolat} \rangle \\ \text{accord [2]} \begin{bmatrix} \text{genre} = \text{masc} \\ \text{nombre} = \text{sing} \end{bmatrix} \\ \text{DET} \begin{bmatrix} \text{phon} = \langle \text{le} \rangle \\ \text{accord [2]} \begin{bmatrix} \text{genre} = \text{masc} \\ \text{nombre} = \text{sing} \end{bmatrix} \end{bmatrix} \\ \text{N} \begin{bmatrix} \text{phon} = \langle \text{chocolat} \rangle \\ \text{accord [2]} \begin{bmatrix} \text{genre} = \text{masc} \\ \text{nombre} = \text{sing} \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

FIGURE 1.3 : Exemple de structure de traits (issu de Bouillon [1998, p. 91])

Les grammaires génératives sont aussi appelées « grammaires syntagmatiques » (dans [Kahane, 2000], par exemple). Beaucoup sont nées de critiques du principe des transformations de

Chomsky. Parmi elles, nous pouvons mentionner la « grammaire lexicale-fonctionnelle » (LFG) [Kaplan & Bresnan, 1995], la « grammaire d'arbres adjoints » (TAG) [Joshi, 1987], ou encore la « grammaire syntagmatique guidée par les têtes » (HPSG) [Pollard & Sag, 1994]. Sans rentrer plus dans les détails de ces grammaires, nous tenons toutefois à évoquer une de leurs similitudes, celle d'utiliser des structures de traits pour décrire « les unités d'information linguistique (morphème, mot, syntagme, phrase, etc.) par l'énumération de leurs caractéristiques sous la forme d'une liste structurée et non ordonnée de paires attribut-valeur » [Bouillon, 1998, p. 89]. Nous empruntons à Bouillon [1998], en figure 1.3, un exemple simple de structure de traits décrivant les constituants d'une phrase.

Les grammaires LFG, TAG, ou encore HPSG sont ainsi toutes trois assimilées à des « grammaires d'unification », dont le principe est de comparer les valeurs que possèdent les traits communs à deux structures de traits différentes. L'unification échoue lorsque deux mêmes traits ont des valeurs contradictoires. Les grammaires d'unification sont particulièrement intéressantes pour le TAL, entre autre parce que leur formalisme leur permet d'être directement implantées dans des systèmes informatiques [Abeillé, 1993]. Mais nous ne nous attardons pas davantage ici sur ces grammaires, qui font l'objet d'un développement plus détaillé dans notre dernière partie (*cf.* § *Portée des attentes* p. 190).

Les « grammaires de dépendance » sont un autre type de grammaires formelles. Leur origine est attribuée principalement aux travaux de Tesnière [1959] dans les années 1930. Elles se fondent sur l'idée que « dans une phrase, la présence de chaque mot (sa nature et sa position) est légitimée par la présence d'un autre mot (son *gouverneur syntaxique*), à l'exception d'un mot, le mot principal associé au sommet de l'arbre syntaxique » [Kahane, 2001, p. 3]. Ces relations de dépendance, représentées par des arbres de dépendance (voir figure 1.2 p. 12), permettent de dégager l'ordre structural des phrases, différent de l'ordre linéaire qui prédomine dans les grammaires syntagmatiques. Cependant, comme le fait remarquer Kahane [2000], « grammaires syntagmatiques et grammaires de dépendance ne s'opposent guère que par le fait de mettre en avant une structure plutôt que l'autre » [Kahane, 2000, p. 3]. Il est en effet possible d'obtenir un arbre de dépendance à partir d'un arbre syntagmatique, et inversement. L'auteur rappelle également que des grammaires d'origine générativiste, comme LFG, TAG ou HPSG que nous avons mentionnées, « sans le revendiquer clairement, s'apparentent fortement à des grammaires de dépendance » [Kahane, 2000, p. 3].

Comme pour les grammaires d'unification, nous revenons de manière plus approfondie sur les relations de dépendance dans notre troisième partie (*cf.* § *Les valences de Tesnière* p. 183)

c) Notre acception de la grammaire

Nous avons tenté jusqu'ici d'apporter des éclaircissements sur la notion de grammaire, d'abord en retraçant son évolution et les différents statuts qu'elle a endossés, puis en expliquant les acceptions contemporaines qui découlent de son histoire ou dépendent de théories linguistiques diverses, avec notamment un arrêt non anodin sur la notion de grammaire formelle. Pour synthétiser ces différentes acceptions, nous pouvons reprendre les propos de Besse & Porquier [1991] qui présentent la grammaire comme « 1) un certain fonctionnement interne caractéristique d'une langue donnée, 2) l'explication plus ou moins méthodique de ce fonctionnement, 3) la méthode d'explicitation suivie » [Besse & Porquier, 1991, p. 11]. Les auteurs présentent également les livres de grammaire comme l'union de ces trois « grammaires », car ils décrivent et expliquent (sens 2) le fonctionnement d'une langue (sens 1) en suivant une certaine théorie linguistique (sens 3).

Pour compléter cet exposé au sujet de la polysémie de « grammaire », nous concluons avec une dernière, mais non moindre signification associée au terme.

La grammaire, du point de vue de la linguistique, est généralement divisée en plusieurs parties dont la phonologie, qui étudie les sons du langage du point de vue de leur fonction dans la langue, la morphologie, qui étudie les formes des mots, et la syntaxe, qui étudie la combinaison des mots pour former les phrases. Selon les théories linguistiques, sont aussi inclues comme parties de la grammaire la sémantique, qui étudie le sens porté par les formes linguistiques, la stylistique, qui étudie les effets de style, leurs procédés, leur expressivité, et la lexicologie, qui étudie les mots, notamment leurs relations avec les autres composants de la langue et leurs rapports avec les facteurs socio-culturels (voir par exemple [Grevisse, 1993 ; Dubois *et al.*, 1994 ; Riegel *et al.*, 1994]).

Notre intérêt ici porte essentiellement sur la morphologie et la syntaxe. En effet, « grammaire » est parfois employé par les linguistes comme synonyme de « morphosyntaxe », nom donné à la partie de la linguistique qui étudie à la fois la morphologie, c'est-à-dire « les formes des mots [...], par opposition à l'étude des fonctions ou syntaxe » [Dubois *et al.*, 1994, p. 311], et la syntaxe, qui s'occupe « à la fois de l'ordre des mots et des phénomènes de rections (c'est-à-dire de la façon dont certains mots imposent des variations à certains autres [..]) » [Ducrot & Schaeffer, 1972, p. 119].

Nous posons ainsi les définitions suivantes :

Définition 1.1 ***La morphologie*** *est une sous-partie de la linguistique qui étudie les formes des mots. Elle décrit les règles de combinaison des morphèmes (plus petite unité porteuse de sens) entre eux pour constituer les mots, notamment par affixation, par composition ou par dérivation. Elle étudie aussi les modifications que subissent les mots en fonction de leur distribution dans la phrase.*

Définition 1.2 ***La syntaxe*** *est une sous-partie de la linguistique qui étudie les relations entre les mots et la combinaison des mots pour former des énoncés.*

Définition 1.3 ***La morphosyntaxe*** *étudie à la fois la morphologie et la syntaxe, c'est-à-dire les règles de formation et de modification des mots en fonction des relations qu'ils entretiennent entre eux dans un énoncé, et les règles de combinaison des mots pour former les phrases.*

La morphosyntaxe est une notion essentielle dans notre travail puisque nous nous intéressons précisément aux règles qui régissent les flexions et les combinaisons des mots dans la phrase, pour la conception d'un outil de vérification grammaticale.

Dans cette thèse, nous prenons ainsi le parti de retenir, parmi les diverses acceptions du mot « grammaire » développées dans cette partie, ce dernier sens, celui pour lequel la grammaire est synonyme de morphosyntaxe. Nous restons dans le cadre de l'acception courante qui définit la grammaire comme l'« ensemble des règles à suivre pour parler et écrire correctement une langue » [Robert, 2006], mais nous en réduisons l'étendue.

Par ailleurs, nous serons également amenée à employer fréquemment le mot « grammaire » dans les appellations de différentes théories de la linguistique formelle : « grammaire formelle » bien sûr, mais aussi « grammaire d'unification » ou encore « grammaire de dépendance ».

1.1.2 L'orthographe

Puisque nous parlons de grammaire et de la « bonne façon d'écrire », nous ne pouvons faire autrement que de nous intéresser aussi à la question de l'orthographe, partie intégrante de la grammaire dont elle est indissociable et avec laquelle elle est aussi souvent confondue.

Étymologiquement, « orthographe » [9] est issu de deux mots du grec signifiant « écrire » et « droit, correct », et désigne ainsi la manière correcte d'écrire les mots. Ceci induit l'existence de normes à respecter, définies soit par les dictionnaires, auquel cas la linguistique parle d'orthographe lexicale, soit par des règles de grammaire (par ex. les règles d'accord), où il s'agit alors d'orthographe grammaticale. L'orthographe peut aussi désigner simplement la manière dont un mot est écrit, sans faire référence à une norme, et devient ainsi synonyme de graphie. Afin d'éviter toute confusion, nous ne l'emploierons pas dans cette acception, réservant à cette dernière le terme plus général de graphie. Mais avant d'aborder l'orthographe, il nous semble important de nous arrêter sur son objet d'étude, le mot, et de préciser ce que cette notion recouvre dans le cadre de nos travaux.

a) Le mot

Le « mot » est une notion complexe à définir en linguistique, bien qu'il soit, comme le dit de Saussure [1916, p.154], « une unité qui s'impose à l'esprit, quelque chose de central dans le mécanisme de la langue ».

Le mot est communément considéré de manière très simple (à l'écrit) comme une suite de lettres délimitée à droite et à gauche par des blancs ou des ponctuations. Mais se posent alors plusieurs problèmes en français, dont celui des apostrophes ou des traits d'union. Dans certains cas ils se comportent comme des blancs et séparent deux mots (`j'ai`, `allons-y`), et dans d'autres cas ils font partie d'une unité lexicale considérée comme un seul et même mot (`aujourd'hui`), appelée mot composé quand il s'agit du trait d'union (`peut-être`). Les mots composés et locutions, qui constituent une unité, peuvent aussi contenir des blancs au lieu de traits d'union (`tapis roulant`, `quelque chose`), mais ces blancs ne sont pas séparateurs. Pour tenter de pallier ces difficultés, les linguistes ont développé différents concepts pour préciser ou remplacer la notion de mot jugée trop floue.

La linguistique comparative a permis de mettre en évidence que les mots sont constitués d'unités significatives plus petites de différents types. Les comparatistes distinguent ainsi les sémantèmes, partie des mots qui portent la signification, et les morphèmes qui sont les marques grammaticales. Ces deux types correspondent à ce que la terminologie courante appelle respectivement radicaux et affixes [Ducrot & Schaeffer, 1972]. Une partie des linguistes contemporains a cependant critiqué cette classification « en faisant valoir qu'elle n'était nullement universelle, qu'elle n'avait pas de sens dans la plupart des langues non indo-européennes, et [...] [qu']elle était plus valable pour les langues anciennes que pour les langues modernes » [Picoche, 2010, p. 14]. Ils ont alors choisi d'appeler morphèmes toutes les unités significatives constituant les mots.

Martinet [1985], quant à lui, préfère utiliser, pour le français, le terme de monème à la place de morphème. Il distingue alors les monèmes lexicaux et les monèmes grammaticaux, qu'il appelle respectivement lexèmes et morphèmes. Cette décomposition du mot est très proche de la décomposition sémantème-morphème, ou encore radical-affixe. Il ajoute cependant une

9. orthographe : du latin *orthographia*, issu du grec *orthos* « droit » et *graphein* « écrire ».

caractéristique en s'inspirant du signe linguistique défini par de Saussure [1916] : selon lui, les monèmes sont des unités de choix élémentaires parmi l'ensemble des possibilités offertes par la langue au moment de la construction de l'énoncé [Ducrot & Schaeffer, 1972]. Cette notion de choix lui permet, contrairement aux autres théories, de considérer sans difficulté les expressions composées comme des monèmes. Ainsi, *pied de page* par exemple est choisi par le locuteur comme unité dans un ensemble qui comprend également des éléments comme *titre*, *colonne*, *numérotation*, etc., et forme donc un monème. Martinet donne le nom de synthème à ce type de monèmes particuliers constitués de plusieurs monèmes lexicaux. D'autres linguistes, selon leur théorie, appellent ces unités complexes de fonctionnement « synapsie » (E. Benveniste), « unités phraséologiques » (C. Bally) ou encore « lexie » (B. Pottier). Ce dernier terme est le plus utilisé en France aujourd'hui en linguistique, mais la terminologie courante parle de mots composés et de locutions pour désigner ces expressions complexes. Traditionnellement, seulement « les termes dont les composants sont graphiquement soudés (*portefeuille*) ou reliés par un trait d'union (*chou-fleur*) » [Dubois *et al.*, 1994] sont appelés composés. Lorsque les composants sont séparés par des blancs, on parle plutôt de locution.

Finalement, les différentes théories ont en commun de prendre en considération les unités graphiques (délimitées par des blancs ou des ponctuations), qu'ils décomposent en unités significatives minimales dont les noms varient, mais conservant la notion de morphème. Ces unités graphiques sont appelées « mots » par certains linguistes, ce qui rejoint la représentation courante que l'on se fait de ce terme, et que nous utiliserons également dans cette acception.

En revanche, les expressions composées de plusieurs unités graphiques sont rarement appelées « mot », même si elles fonctionnent généralement comme un mot graphique unique. Elles sont en effet choisies en tant qu'unité sémantique et lexicale dans un ensemble d'autres formes (simples ou composées) avec lesquelles elles peuvent commuter syntaxiquement, et elles répondent généralement au critère d'inséparabilité, c'est-à-dire qu'il est impossible d'y intercaler un morphème supplémentaire. Ce second critère n'est cependant pas toujours vrai pour les locutions. Par exemple, la locution verbale *faire appel à* peut être aisément divisée par l'ajout d'un adverbe : *faire souvent appel à*. Mais surtout, le fait que les constituants des mots composés ou des locutions soient séparés, que ce soit par des blancs ou des traits d'union, conduit à les distinguer des formes graphiques simples traditionnellement appelées « mot ».

Dans cette thèse portant sur la vérification grammaticale, notre approche des phrases et de leurs constituants est bien plus morphologique que sémantique, et consacrée à l'écrit, donc essentiellement graphique. Par ailleurs, d'un point de vue plus technique, la prise en charge du niveau sémantique par les technologies du TAL en est encore à ses débuts, alors que l'identification de suites de caractères est déjà relativement bien maitrisée. Il nous semble alors pertinent, dans l'objectif de modélisation des erreurs tapuscrites et d'un outil informatique pour les détecter, de définir ce que recouvre la notion de mot en utilisant des critères graphiques plutôt que sémantiques.

Nous réservons donc aux formes graphiques simples (suite de lettres entourées de blancs ou de ponctuations, par ex. *magasin*) l'appellation de « mot », en y incluant les mots composés dont les éléments sont graphiquement soudés et qui sont donc graphiquement simples (par ex. *bienfaisance*). Nous utilisons le terme « mot composé » pour désigner les unités lexicales formées de deux ou plusieurs unités graphiques simples, liées soit par des traits d'union (par ex. *au-delà*), soit par des blancs (par ex. *tapis roulant*), et qui par ailleurs satisfont le critère d'inséparabilité. Nous réservons le terme de « locution » plutôt aux expressions qui ne sont pas totalement figées (par ex. *avoir peur*). Enfin, nous employons le terme « morphème » pour parler des unités

constitutives des mots, en distinguant les morphèmes lexicaux porteurs du sens du mot, et les morphèmes grammaticaux.

Face à la diversité des définitions dans la littérature, nous posons les définitions sur lesquelles nous nous appuierons dans cette thèse :

Définition 1.4 ***Le mot*** *est une unité graphiquement simple, c'est-à-dire délimitée à droite et à gauche par des blancs ou des ponctuations, et ne contenant ni blanc ni trait d'union. Par ex.* `sac`, `infrastructure`, *etc.*

Définition 1.5 ***Le morphème*** *est l'unité minimale porteuse de sens constituant le mot. Un mot peut souvent être décomposé en plusieurs morphèmes. Nous distinguons :*

- *les morphèmes lexicaux : ils correspondent d'une part aux radicaux, d'autre part aux affixes dérivationnels permettant de former de nouveaux mots. Ils constituent des ensembles ouverts et nombreux. Par ex.* `bouton`, `juge`-`ment`, *etc.*
- *les morphèmes grammaticaux : ils correspondent aux affixes flexionnels et ont une signification grammaticale. Ils ne forment pas de mots nouveaux, mais ajoutent des informations de genre, de nombre, de temps, de personne ou de mode aux mots en fonction de leur rôle syntaxique. Ils constituent des ensembles restreints et fermés. Par ex.* `aim`-`ons`, `grand`-`e`-`s`, *etc.*

Définition 1.6 ***Le mot composé*** *est une unité lexicale composée d'au moins deux formes graphiques simples liées par un blanc ou un tait d'union, qui fonctionne comme un mot simple. Par ex.* `tapis roulant`, `peut-être`, *etc.*

Définition 1.7 ***La locution*** *est une expression composée d'au moins deux formes graphiques simples, qui fait l'objet d'un choix unique dans un inventaire des possibilités de la langue, mais qui n'est pas totalement figée et donc syntaxiquement divisible. Par ex.* `faire appel à`, `avoir peur`, *etc.*

b) L'orthographe lexicale

L'orthographe lexicale, appelée aussi orthographe d'usage, « a pour objet les mots pris en eux-mêmes, tels que les donne le dictionnaire, sans égard à leur rôle dans la phrase » [Grevisse, 1993, p. 91]. Elle fixe la graphie de ce que nous appelons morphèmes lexicaux (*cf.* définition 1.5 p. 17), par opposition aux morphèmes grammaticaux qui sont laissés aux soins de l'orthographe grammaticale. Les morphèmes lexicaux sont essentiellement constitués des unités de base qui forment le lexique (noms, adjectifs, verbes et adverbes), à savoir les radicaux et les affixes dérivationnels.

L'orthographe lexicale s'occupe donc de la graphie de ces morphèmes, mais en dehors de tout contexte syntaxique. Elle n'est ainsi pas concernée par les modifications que peuvent subir les mots variables selon leur fonction syntaxique. Elle a donc pour objet la graphie :

- des radicaux : elle définit par exemple l'orthographe du radical `froid`, mais pas des modifications désinentielles dont il fait l'objet au féminin et/ou au pluriel, avec l'adjonction d'un ou plusieurs morphèmes grammaticaux (`froid`-`e`, `froid`-`e`-`s`, `froid`-`s`) ;

- des affixes dérivationnels : ces morphèmes lexicaux sont par exemple *sur-*, *-tion*, *-ment*, etc. et permettent la création de mots nouveaux lorsqu'ils sont accolés à d'autres morphèmes (par ex. juste*-ment*) ;
- des formes lemmatisées qui constituent les entrées des dictionnaires : elle définit par exemple l'orthographe des verbes à l'infinitif, tel que *rédiger*, mais pas des différentes désinences qu'ils prennent une fois conjugués (*rédige*-ons, *rédiger*-ez, *rédige*-ant, etc.) ;
- des mots invariables : elle définit par exemple l'orthographe des morphèmes lexicaux tels que *plusieurs* ou *discours* dont la forme ne varie jamais.

Chaque unité lexicale a donc une graphie définie, qui ne dépend pas de la grammaire. Pour connaître celle qui correspond au mot que l'on souhaite écrire, il est nécessaire de recourir à un dictionnaire. Beaucoup de phonèmes ont d'ailleurs une transcription variable d'un mot à l'autre en français, tel le phonème [o] dans les mots *numéro*, *niveau*, *faux*, *côté*, ou encore le phonème [s] dans les mots *scène*, *six*, *notion*, *basse*. Ces diverses formes graphiques possibles pour un même son ont pour origine la transcription du français à l'époque médiévale. Cette transcription fut réalisée à l'aide de l'alphabet latin qui n'était pas assez riche pour représenter les phonèmes du français, plus nombreux que ceux de la langue latine [Burney, 1970]. Divers subterfuges furent alors utilisés par les scribes de l'époque pour pallier ces lacunes, conduisant à des graphies variées pour un même mot. Ainsi, aux XVI^{e} et XVII^{e} siècles, plusieurs manières d'écrire cohabitent : celle des imprimeurs d'un côté, qui favorise les graphies simplifiées et proches de la prononciation, et d'autres plus étymologiques influencées par le latin. Il n'y a alors pas encore de norme unifiée, et donc pas une, mais des orthographes. Ce n'est qu'au XVIII^{e} siècle que la notion de norme orthographique apparaît, avec la fixation des graphies dans les dictionnaires, et notamment dans le Dictionnaire de l'Académie Française [10] . Ces graphies ne sont pas influencées par la grammaire. Elles ont en revanche, selon Burney [1970], un fort caractère étymologique, dû à la volonté de leur faire conserver des traces de leurs origines prestigieuses latine ou grecque. Elles ont également pour certaines un aspect idéographique guidé par la nécessité de distinguer des homonymes. Ces graphies ainsi fixées, elles constituent l'orthographe lexicale, qui n'a quasiment plus été modifiée depuis, les réformes successives, dont la dernière date de 1990, peinant à s'imposer.

c) L'orthographe grammaticale

Conjointement à l'orthographe lexicale, il existe l'orthographe grammaticale, appelée aussi orthographe de règle, ou d'accord. Elle « concerne les modifications grammaticales des mots, celles qu'ils subissent pour jouer leur rôle dans la phrase. » [Grevisse, 1993, p. 91]. Elle régit principalement les accords entre les mots (déterminant et nom, sujet et verbe, etc.) et détermine par exemple les marques du pluriel ou du féminin, ou les désinences de conjugaison. Elle implique ainsi de savoir identifier les rapports qu'entretiennent les mots d'une phrase afin de pouvoir les accorder, et par conséquent de savoir appliquer les règles normatives de la grammaire.

Contrairement à l'orthographe lexicale, elle traite des mots en contexte, dans la phrase, et s'intéresse uniquement aux parties variables de ces mots. Ces dernières sont formées des affixes flexionnels, qui sont des morphèmes grammaticaux « porteurs d'une signification proprement grammaticale et qui ne créent pas des mots nouveaux, mais des formes différentes d'un même mot » [Riegel *et al.*, 1994, p. 537]. Par ailleurs, alors que les normes de l'orthographe lexicale figurent dans les dictionnaires, les règles de l'orthographe grammaticale se trouvent dans les manuels de grammaire, d'où probablement la confusion commune entre grammaire et orthographe.

10. Dictionnaire de l'Académie Française : première édition en 1694.

Outre les marques morphologiques de flexion, l'orthographe grammaticale s'intéresse également aux homophones grammaticaux, qui sont fréquemment confondus et conduisent alors à des erreurs syntaxiques. L'orthographe grammaticale régit donc :

- les marques de genre et de nombre, tels par exemple le suffixe -e du féminin (*grand*-e), ou le suffixe -s du pluriel (*grand*-s) ;
- les marques de personne, de temps et de mode, tels par exemple le suffixe -ez de la deuxième personne du pluriel (*rédig*-ez), ou les suffixes du futur (*rédig*-erez) ;
- les homophones grammaticaux, comme par exemple a et à, sait et c'est, quand et quant, etc.

1.1.3 Conclusion

Après avoir présenté les notions de grammaire et d'orthographe dans leurs diverses acceptions, nous avons retenu dans le cadre de ce travail, la définition de grammaire qui la présente comme synonyme de morphosyntaxe, puisque notre travail a pour objet d'étude précisément l'organisation des mots dans la phrase, et les modifications morphologiques qu'ils subissent en fonction de leur rôle syntaxique.

La question de l'orthographe, indissociable de la grammaire, a également été abordée, avec une distinction entre l'orthographe lexicale et l'orthographe grammaticale. La première fixe la graphie des mots en eux-mêmes, en dehors de tout contexte grammatical. Elle nous intéresse peu dans le cadre de cette thèse dans la mesure où elle est, à l'heure actuelle, très bien prise en charge par les outils informatiques. Doll & Coulombe [2004, p. 35] expliquent en effet que « l'algorithme de base [d'un correcteur orthographique] s'écrit en moins d'une journée » et que « le plus gros du travail est de créer le lexique de référence ». Nous concentrons donc notre travail sur la seconde, l'orthographe grammaticale, qui régit la graphie des parties variables des mots en fonction de leur contexte syntaxique.

Nous avons également précisé ce que recouvre pour nous la notion de mot, à savoir une forme graphiquement simple, et nous avons mentionné la morphologie (*cf.* définition 1.1 p. 14), qu'il est difficile de ne pas rapprocher de l'orthographe (grammaticale et lexicale), puisque toutes deux focalisent leur attention sur les mots. Mais alors que la morphologie traite de la structure et de la formation des mots (aussi bien à l'oral qu'à l'écrit), par combinaison de morphèmes par exemple, l'orthographe définit les graphies qui sont associées à chaque forme. À un autre niveau, l'orthographe grammaticale et la morphosyntaxe sont également intimement liées, dans la mesure, par exemple, où les variations morphologiques des mots (flexions) sont fortement tributaires de la syntaxe [Riegel *et al.*, 1994].

Ainsi, pour résumer, l'orthographe permet et régit la représentation graphique des objets d'étude de la morphologie, et l'orthographe grammaticale plus précisément se concentre sur la représentation graphique des variations induites par les règles de la morphosyntaxe (ou grammaire dans notre acception).

1.2 Définitions de l'erreur et de la faute

Nous avons maintenant une idée plus précise sur les notions de grammaire et d'orthographe telles que nous les envisageons. Il s'agit maintenant de préciser ce que nous entendons par fautes et erreurs de grammaire.

1.2.1 Définitions générales

Les mots « erreur » et « faute » sont communément employés comme des synonymes pour qualifier un acte (ou son résultat) qui transgresse ou ne se conforme pas à une règle ou une norme. Une connotation négative plus lourde est cependant associée à la faute, ce que nous confirment par exemple des dictionnaires de la langue française.

De son étymologie latine, « erreur » [11] garde l'idée de « se tromper », d'un « acte de l'esprit qui tient pour vrai ce qui est faux et inversement » [Robert, 2006], d'où il découle également l'idée d'un écart par rapport à une norme, une vérité, une valeur exacte : « action non prévue par rapport à une norme » [Robert, 2006].

Le mot « faute » [12] quant à lui a les mêmes racines que le verbe « faillir » avec lequel il partage l'idée de manquement à quelque chose. Il est connoté bien plus négativement qu'« erreur » car généralement associé aux notions de morale, de religion ou de loi, dont le manquement à l'une des règles est souvent fortement stigmatisé. Ainsi, lorsque « erreur » renvoie à un écart par rapport à une norme ou une règle, sans exprimer de jugement de valeur, « faute » est défini comme un manquement à cette norme ou règle, et porte ainsi une valeur dépréciative et moralisante [Robert, 2006]. Dans le cadre de cette thèse, les notions d'erreur et de faute sont indissociables des notions de grammaire et d'orthographe puisque nous nous intéressons à ce qui est communément appelé « faute de grammaire » et « faute d'orthographe ». Nous pouvons alors restreindre au domaine de la langue la portée des définitions que nous venons d'énoncer. Dans ce contexte, la norme évoquée par les définitions génériques n'est autre que la norme grammaticale que nous présentions dans le point 1.1 p. 7. Ce sont ainsi tous les écarts affectant les règles de grammaire et d'orthographe qui constituent les erreurs, ou les fautes, auxquelles nous nous intéressons.

1.2.2 Erreur et faute en didactique des langues

Nous évoquions en début de ce chapitre (*cf.* § *Origine de la grammaire : de l'art à la discipline* p. 8) la place privilégiée occupée dans l'enseignement par la grammaire depuis son origine. Aujourd'hui encore, les questions de grammaire et d'orthographe sont une préoccupation importante dans le domaine de la pédagogie, et plus généralement dans celui de la didactique des langues qui l'englobe. La pédagogie « s'intéresse avant tout à la relation entre l'élève et l'enseignant, quel que soit l'objet d'étude » [Simard, 1997, p. 2], alors que la didactique des langues s'intéresse aux processus d'enseignement (relations entre enseignant et apprenant) et d'apprentissage (relations entre apprenant et contenu) des langues, étrangères ou maternelles, ainsi qu'aux contenus enseignés. L'élaboration de ces contenus nécessitant « des descriptions et théorisations des langues en présence » [Chiss, 2009, p. 130], la didactique des langues entretient aussi des relations privilégiées avec les sciences du langage. Cuq [2010] la présente d'ailleurs comme cousine de la linguistique, toutes deux constituant des branches des sciences du langage.

La didactique des langues se focalisant sur « [l]es conditions et [l]es modalités d'enseignement et d'appropriation des langues en milieu non naturel [13] » [Cuq & Gruca, 2005, p. 25], les notions de faute et d'erreur y tiennent une place non négligeable et sont l'objet de nombreuses publications depuis le début du XX^{e}. Les plus anciennes ne distinguent généralement pas les

11. erreur : issu du latin *errare* « aller ça et là, faire fausse route, s'égarer ».
12. faute : issu du latin *fallere* « faire trébucher, tromper ».
13. par opposition à *l'acquisition* naturelle des langues qui, selon Cuq & Gruca [2005, p. 52], ne relève pas de la didactique des langues mais se situe « exactement à la charnière entre la didactique et la linguistique ».

deux termes, voire privilégient celui de faute, telles *La grammaire des fautes* de Frei [1928], la *Grille de classement typologique des fautes* de Debyser *et al.* [1967] ou encore le *Dictionnaire de didactique des langues* de Galisson & Coste [1976]. Cette préférence pour « faute » s'accompagne alors de la valeur dépréciative et moralisante couramment associée au terme. Mais au fil des ans, « erreur » a progressivement pris le dessus sur « faute » d'une part, et gagné une place primordiale dans l'enseignement d'autre part, dans lequel elle est désormais généralement non plus bannie et sanctionnée, mais plutôt considérée comme « une étape primordiale dans l'acquisition des connaissances langagières » [Demirtaş & Gümüş, 2009, p. 133]. L'erreur est alors placée au centre de la démarche pédagogique et acquiert le statut « d'indicateur et d'analyseur des processus intellectuels en jeu [...]. Au lieu d'une fixation (un peu névrotique ?) sur l'écart à la norme, il s'agit plutôt de décortiquer la « logique de l'erreur » et d'en tirer parti pour améliorer les apprentissages. » [Astolfi, 1997, p 17]. Nous sommes alors dans une pédagogie de l'erreur.

L'erreur a donc peu à peu acquis un nouveau statut en didactique, au détriment de la faute. Mais cette dernière n'a cependant pas disparu de la littérature du domaine. Les auteurs s'efforcent dans ce cas de la distinguer explicitement de l'erreur. Ainsi, dans Larruy [2003], les fautes sont définies comme les erreurs ayant pour origine l'inattention ou la fatigue et pouvant être corrigées par l'apprenant lui-même. Lorsqu'au contraire c'est la méconnaissance d'une règle de la langue qui entre en jeu et que l'autocorrection n'est pas possible, il s'agit là de « simples » erreurs. Cette différenciation se fonde sur la distinction établie par Chomsky [1965] entre compétence et performance.

> « *Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech-community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance. [...] We thus make a fundamental distinction between* competence *(the speaker-hearer's knowledge of his language) and* performance *(the actual use of language in concrete situations)*[14] »
>
> [Chomsky, 1965, p. 3-4]

Ainsi, selon la théorie linguistique de Chomsky, nous possédons une connaissance intrinsèque de la langue (la compétence linguistique), sous-jacente à l'utilisation réelle que nous en faisons (la performance linguistique) en fonction de divers facteurs (environnementaux, physiologiques, psychiques, etc.).

Corder [1980] reprend cette opposition compétence/performance pour distinguer les erreurs des apprenants :

> « Il nous faut alors distinguer les erreurs qui sont dues au hasard des circonstances de celles qui reflètent à un moment donné sa [l'apprenant] connaissance sous-jacente [...]. Aussi sera-t-il commode désormais d'appeler « fautes » les erreurs de performance, en réservant le terme d'« erreur » aux erreurs systématiques des apprenants, celles

14. « L'objet premier de la théorie linguistique est un locuteur-auditeur idéal, appartenant à une communauté linguistique complètement homogène, qui connaît parfaitement sa langue et qui, lorsqu'il applique en une performance effective sa connaissance de la langue, n'est pas affecté par des conditions grammaticalement non pertinentes, telles que limitation de mémoire, distractions, déplacements d'intérêt ou d'attention, erreurs (fortuites ou caractéristiques). [...] Nous établissons donc une distinction fondamentale entre la *compétence* (la connaissance que le locuteur-auditeur a de sa langue) et la *performance* (l'emploi effectif de la langue dans des situations concrètes). » (traduit par [Chomsky, 1971], p. 12-13).

> qui nous permettent de reconstruire leur connaissance temporaire de la langue, c'est-à-dire leur compétence transitoire. »
>
> [Corder, 1980, p. 13]

Cette dichotomie est bien adaptée à l'interprétation informatique des erreurs d'écriture. En effet, l'utilisation du clavier, par exemple, peut conduire à diverses coquilles (par ex. oubli, inversion, doublement de lettres) qui sont typiquement des erreurs de performance, contrairement à d'autres erreurs grammaticales systématiques dues à la compétence du scripteur, comme un accord erroné entre un verbe et son sujet.

1.2.3 Précisions terminologiques

Nous avons parlé jusqu'ici d'erreur, de faute, d'écart par rapport à une norme, mais nous n'avons pas encore mentionné une terminologie utilisée en linguistique pour caractériser les phrases grammaticalement correctes, incorrectes, ou douteuses. Il s'agit des notions de « grammaticalité » et d'« acceptabilité ». Lorsqu'un énoncé ne respecte pas les règles de la grammaire d'une langue, et donc qu'il contient une ou des erreurs morphosyntaxiques, il est dit agrammatical, ou inacceptable. Il arrive également que certains énoncés soient douteux, c'est-à-dire qu'en fonction du contexte, de leur interprétation, etc., ils peuvent être jugés soit acceptables, soit agrammaticaux. Il existe ainsi différents degrés de jugement de grammaticalité possibles des énoncés, allant de l'énoncé grammaticalement correct à l'énoncé agrammatical. Il est d'usage dans la littérature en linguistique d'identifier les énoncés agrammaticaux en les faisant précéder d'un astérisque (par ex. *`C'est qui a disparu?`), et les énoncés jugés d'acceptabilité douteuse avec un point d'interrogation (par ex. ?`Cela ne m'étonne.`). Nous nous conformerons à ces conventions dans les pages suivantes.

Si nous reprenons à présent les précisions que nous avons apportées sur la grammaire, l'orthographe et l'erreur tout au long de cette première partie de la problématique, nous nous attachons donc à étudier les écarts de graphie et de syntaxe par rapport aux règles de l'orthographe et de la morphosyntaxe. Pour désigner ces écarts, nous privilégierons le terme d'« erreur » (ou d'« agrammaticalité » pour les énoncés complets). Nous utiliserons également ponctuellement le terme « faute », mais en faisant référence à la dichotomie compétence/performance. Nous lui associerons ainsi systématiquement l'idée d'erreur de performance, afin d'éviter toute connotation négative.

Pour réutiliser les termes communs de grammaire et orthographe, nous faisons donc les distinctions suivantes :

Définition 1.8 ***Les erreurs d'orthographe*** *sont constituées des seules erreurs de l'orthographe lexicale (*`plusieur`*). Nous les appelons également « erreurs de graphie ». Nous ne nous attarderons pas sur ces erreurs qui sont déjà bien prises en charge par les outils de correction actuels.*

Définition 1.9 ***Les erreurs de grammaire*** *comprennent à la fois les erreurs d'orthographe grammaticale (*`tu insère`*) et les erreurs syntaxiques (*`c'est pas grave`*). Nous les appelons également « erreurs morphosyntaxiques » puisqu'elles concernent la morphologie et la syntaxe. Ce sont sur les erreurs de ce type que nous focalisons notre travail.*

Définition 1.10 ***Les fautes*** *correspondent aux erreurs manifestement dues à un problème de performance (faute de frappe, d'inattention, etc.), et non de compétence. Il peut s'agir de fautes au niveau orthographique (*`dasn` pour `dans`) ou grammatical (`lange` pour `langue`).*

Définition 1.11 ***Les énoncés agrammaticaux*** *désignent les énoncés qui contiennent des erreurs morphosyntaxiques (*`C'est qui a disparu?`).*

Enfin, dans le cadre de nos travaux, sans aller jusqu'à concevoir un outil destiné à l'apprentissage de la langue, nous nous positionnons en faveur de la place progressivement donnée à l'erreur dans la démarche pédagogique. La vérification grammaticale automatique telle que nous la concevons ne doit pas, en effet, se contenter de signaler les erreurs présumées, comme c'est parfois le cas avec certains correcteurs grammaticaux, mais doit expliquer leur nature et le moyen de les corriger. Elle se doit d'avoir une dimension pédagogique, au moyen des rétroactions destinées au scripteur. Celui-ci doit pouvoir bénéficier d'une explication sur l'origine de l'incohérence grammaticale qui a été détectée, et disposer également d'aides pour la rectifier si nécessaire.

Chapitre 2

Interprétation informatique de l'erreur de grammaire

Sommaire

2.1 Mécanismes de gestion des erreurs tapuscrites 25
2.1.1 La vérification orthographique 26
2.1.2 La vérification grammaticale 27
2.2 De la grammaire académique à la grammaire en bureautique . 35
2.2.1 La grammaire des outils bureautiques 35
2.2.2 Les types d'erreurs 36
2.2.3 Conclusion 37

Nous avons présenté dans le chapitre précédent ce que recouvrait la notion d'erreur de grammaire d'un point de vue linguistique, et la portée que nous lui donnions dans cette thèse. L'objet de notre recherche étant orienté vers l'informatique, avec l'étude des erreurs commises dans les tapuscrits et l'implantation d'un modèle de ces erreurs dans un outil de correction, il est indispensable à présent de définir la notion d'erreur de grammaire d'un point de vue informatique, ou plus exactement d'un point de vue bureautique. Nous allons voir que cette définition se fonde non pas sur des critères linguistiques, mais sur des critères techniques. Pour cette raison, il nous semble pertinent de commencer par présenter « comment » et « par quoi » sont prises en charge les erreurs, avant d'aborder la question de la grammaire (et de l'orthographe) en bureautique.

2.1 Mécanismes de gestion des erreurs tapuscrites

Avec l'essor de l'informatique et la quantité grandissante de textes électroniques, les besoins en outils bureautiques d'aide à la rédaction ont conduit au développement de nombreux logiciels de correction, appelés parfois « correcticiels » [1]. Ils sont généralement spécialisés dans l'orthographe et la grammaire. Les plus complets, comme Antidote [2] ou Cordial [3], proposent également quelques

1. terme utilisé notamment par la communauté québécoise, composé à partir de « correction » et « logiciel », sur le modèle de « didacticiel ». Le terme n'est cependant pas référencé par les bases de données terminologiques de la langue française
2. Antidote RX, Druide informatique, 2011
3. Cordial, Synapse Développement, 1994-2011

fonctionnalités de vérification relatives à la sémantique et la stylistique, mais elles sont encore relativement limitées et donc peu utilisées. En effet, la stylistique est un domaine qui laisse une grande liberté à l'expression et qui est difficilement interprétable par un correcteur. Ceux-ci se bornent alors à vérifier la longueur des phrases, l'absence de répétitions, d'anglicismes, de barbarismes, etc. En ce qui concerne l'aspect sémantique, les recherches en TAL ne permettent pour le moment pas d'effectuer de représentation du sens des phrases. Les outils de correction proposent alors simplement des synonymes, des définitions ou des analogies.

Nous laissons donc de côté la correction encore balbutiante de la sémantique et de la stylistique, pour nous focaliser sur la correction de la grammaire, en passant par celle de l'orthographe. Pour ce faire, nous expliquerons comment la machine détermine l'erreur, prémisse à toute correction.

2.1.1 La vérification orthographique

Les correcteurs orthographiques, comme leur nom l'indique, s'occupent de l'orthographe. Plus précisément, leur fonction consiste à vérifier que chaque mot du texte appartient bien à la langue. Pour cela, ils parcourent la chaîne de caractères que constitue le texte afin de délimiter chaque mot, dont ils vérifient ensuite que la graphie est connue, en la comparant à une liste de mots correctement orthographiés. Les correcteurs se fondent pour cela sur des lexiques constitués non seulement des lemmes tels qu'ils apparaissent dans les entrées des dictionnaires, mais également de toutes les formes fléchies possibles pour ces lemmes. Ces lexiques aspirent à être aussi exhaustifs que possible dans leur énumération de tous les mots, mais ils ne peuvent en pratique jamais atteindre l'exhaustivité, la langue évoluant sans cesse.

Ainsi, lorsqu'un correcteur orthographique rencontre dans un texte un mot dont la graphie ne correspond à aucune de celles présentes dans son lexique, il signale aussitôt que le mot est mal orthographié. La plupart des outils de ce type est de plus capable aujourd'hui de suggérer une ou des graphie(s) alternative(s) au mot détecté comme incorrect. Il s'agit de mots proches graphiquement et/ou phonétiquement du mot incriminé. Ainsi, pour le mot inconnu *`choisient`, le correcteur orthographique intégré à notre logiciel de traitement de texte nous propose comme correction, dans l'ordre : `choisirent`, `choisie`, `choisisseuse`, `choisi` et `chosifient`, et ce quel que soit le contexte.

La vérification de l'orthographe peut se faire à la demande de l'utilisateur à tout moment de la rédaction, sur l'ensemble de ce qui a déjà été rédigé, mais la plupart des correcteurs orthographiques permettent également la détection des erreurs au fur et à mesure de la saisie du texte. Dès qu'un mot est écrit, il est vérifié et mis en évidence (par ex. souligné en rouge) s'il n'est pas reconnu par le système, ce qui permet à l'utilisateur de le rectifier dans la foulée si nécessaire.

Les correcteurs orthographiques vérifient donc qu'ils connaissent la graphie des mots du texte, sans égard à leur contexte syntaxique, et suggèrent des corrections si nécessaire, sans toutefois les effectuer automatiquement pour la plupart. Certains outils disposent cependant de dictionnaires d'erreurs très fréquentes qui peuvent alors être corrigées de manière automatique. Le correcteur d'OpenOffice.org [4] par exemple, qui utilise le moteur de correction Hunspell [5], peut être paramétré pour rectifier automatiquement des erreurs comme *`dnas` ou *`domage`. Lorsqu'un document

4. OpenOffice.org : suite bureautique *open-source* (http ://fr.openoffice.org/)
5. Hunspell : correcteur orthographique *open-source* (http ://hunspell.sourceforge.net)

contient plusieurs occurrences d'une même erreur, il est également possible de les corriger toutes en même temps. Mais de manière générale, l'intervention de l'utilisateur est nécessaire pour valider ou modifier les suggestions de correction proposées par l'outil. Il serait alors plus approprié de le nommer « vérificateur » plutôt que « correcteur », comme le fait l'anglais qui utilise le terme « *spellchecker* », soit littéralement « vérificateur d'épellation ». Le terme « épellation » est d'ailleurs également moins ambigu qu'« orthographe », ce dernier étant communément très lié à la grammaire, alors qu'il ne s'agit pour l'outil de vérification que de traiter des graphies, en dehors de toute notion grammaticale.

2.1.2 La vérification grammaticale

Les correcteurs grammaticaux ont vocation à détecter les erreurs pour lesquelles les correcteurs orthographiques ne sont pas compétents. Ils sont ainsi chargés d'une part des erreurs concernant l'organisation de la phrase (par ex. ordre des mots) et les relations entre les mots (par ex. accords), et d'autre part des erreurs de graphie qui conduisent à un mot existant mais incorrect dans le contexte de la phrase (par ex. confusion d'homophones, mauvais accord, etc.).

Leurs fonctionnements sont beaucoup plus complexes que les vérificateurs d'orthographe. Ils se décomposent classiquement en plusieurs étapes successives de traitement (voir figure 2.1 p. 28), effectuées généralement après la rédaction, rarement à la volée pendant la construction de la phrase comme c'est le cas en correction orthographique. En effet, les différentes étapes réalisent soit des analyses ascendantes des phrases (approche *bottom-up*), en partant des mots graphiques pour arriver aux syntagmes par exemple, soit au contraire des analyses descendantes (approche *top-down*), allant de la phrase vers le mot. Dans un cas comme dans l'autre, il est alors nécessaire que les phrases soient terminées pour que les différents traitements soient fonctionnels.

Nous ne présentons ici que les principales étapes de traitement, que nous avons pu observer au sein de logiciels libres [6].

a) Tokenisation

La première étape de traitement consiste à découper la chaîne de caractères en phrases et en mots, ou plus exactement en *tokens*. C'est la segmentation, ou tokenisation. Nous appelons *tokens* les séquences de caractères du texte délimitées par des blancs ou des ponctuations, ce qui correspond à la définition que nous avons donnée du mot (*cf.* définition 1.4 p. 17). C'est pourquoi pour simplifier nous parlerons parfois de mot à la place de *token*, mais nous privilégierons tout de même le second terme. *Token* englobe en effet d'autres éléments graphiques que les mots, comme les ponctuations (porteuses de sens syntaxiques), les nombres, les sigles, etc. Ainsi, l'énoncé

```
Cependant, le TAL n'est pas fiable à 100%.
```

contient 11 *tokens* :

```
| Cependant | , | le | TAL | n' | est | pas | fiable | à | 100% | . |
```

La segmentation en *tokens* n'est pas aussi triviale qu'elle peut sembler l'être à première vue, mais c'est sur elle que reposent par la suite tous les traitements et analyses du texte. Elle doit

6. Compte tenu du secret industriel, il ne nous a pas été possible d'étudier la structure des logiciels commerciaux. Nous décrivons donc ici la constitution classique des correcteurs libres (dont le code source est librement accessible), qui n'est pas nécessairement identique à celle des correcteurs propriétaires.

donc être réalisée le mieux possible, malgré les diverses difficultés auxquelles elle se heurte. La principale d'entre elles rappelle directement celles rencontrées dans la définition de la notion de mot. Nous évoquions en effet, dans le paragraphe 1.1.2 a) p.15 sur le mot, les problèmes posés par les apostrophes, les traits d'union et plus généralement par les mots composés.

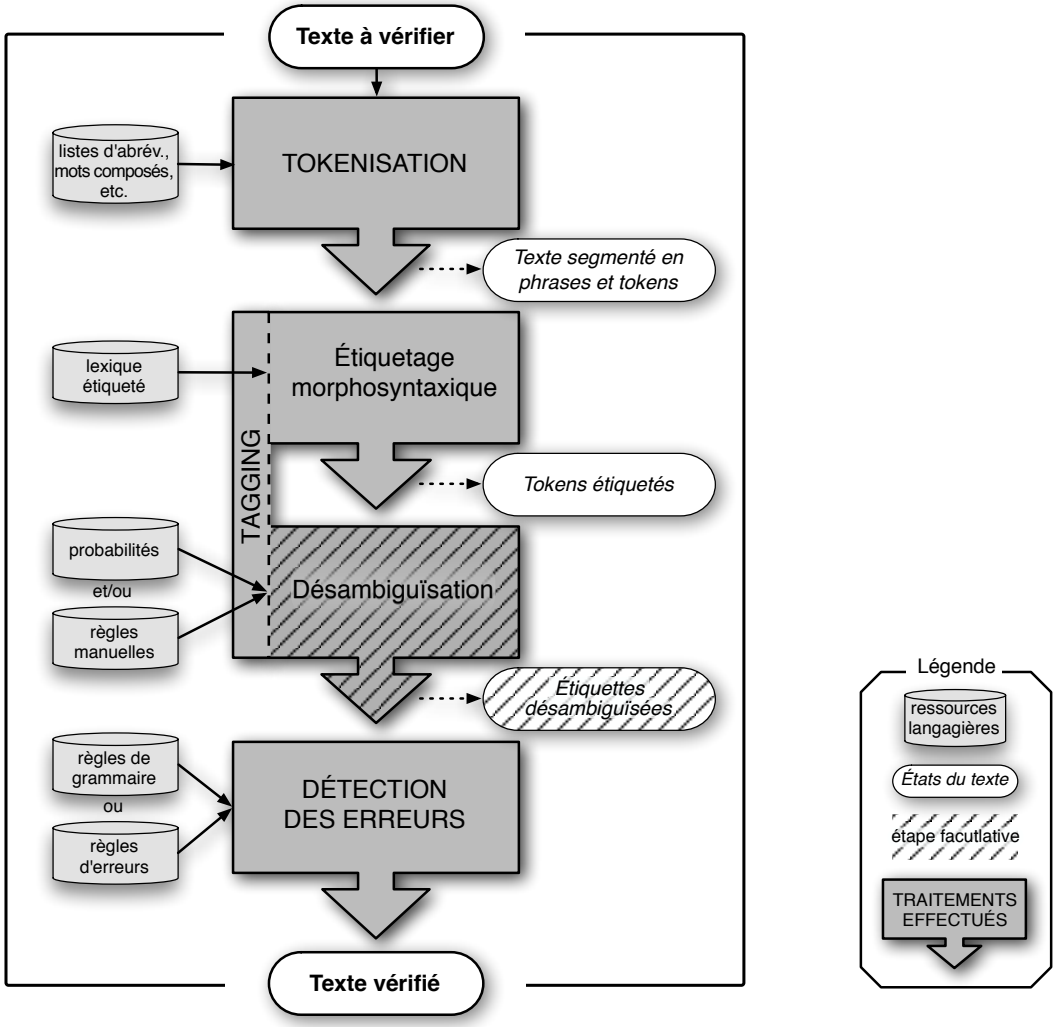


FIGURE 2.1 : Structure générale en couche d'un correcteur grammatical

Fontenelle [2005b], dans son article, prend l'exemple de l'apostrophe en français et présente les avantages et les inconvénients à la considérer comme séparateur ou comme non séparateur. Dans le premier cas, toutes les formes élidées, qui constituent la majorité des occurrences d'apostrophes en français, sont correctement séparées du mot qu'elles précèdent (`|l'|élection|`, `|lorsqu'|elle|`, `|m'|en|`, etc.). Il faut en revanche dans ce cas considérer et traiter spécifiquement un certain nombre d'exceptions pour lesquelles il ne doit pas y avoir de scission. La liste, *a priori* fermée et connue, contient des mots tels que `aujourd'hui`, `presqu'île`, `s'entr'aimer`,

etc., mais d'autres formes imprévisibles peuvent venir se rajouter à ces exceptions, tels certains noms propres (par ex. `O'Brian`), ou encore, comme le remarque Habert [1998], des noms de conférences accolés à un millésime (par ex. `JEP'10` désigne les Journées d'Études de la Parole qui se sont déroulées en 2010). Si l'apostrophe n'est au contraire pas considérée comme séparateur, les exceptions gardent leur unité par défaut, mais il faut en revanche mettre en place des règles spécifiques pour que les formes comprenant une apostrophe d'élision soient à juste titre scindées en deux mots distincts.

Le trait d'union est tout autant problématique. S'il est considéré comme séparateur, les formes composées à valeur grammaticale (verbe-pronom : *est-elle* ; pronom-même : *lui-même* ; etc.) sont séparées adéquatement, mais les mots composés sont divisés à tort, et il est alors nécessaire pour les réunifier de disposer d'un lexique qui les liste. Cela n'aurait pas de sens, par exemple, de considérer comme deux mots distincts un mot tel que *plate-forme*. Si au contraire le trait d'union n'est pas inclus aux caractères séparateurs, alors les mots composés gardent leur unité et ne posent pas de problème, mais il est en revanche nécessaire de pouvoir distinguer les cas où des mots sont liés dans une relation grammaticale, afin de pouvoir effectuer leur séparation. Une expression composée comme *vais-je* ne peut pas être analysée correctement si ses différents constituants ne sont pas isolés.

Nous pouvons citer également certains mots composés, certes rares, qui contiennent à la fois une apostrophe et un trait d'union : `chef-d'oeuvre` par exemple, selon les décisions prises concernant les caractères séparateurs, peut se voir divisé en 3 *tokens* (`|chef-|d'|oeuvre|`), en 2 *tokens* (`|chef-d'|oeuvre|` ou `|chef-|d'oeuvre|`), ou bien rester unifié (`|chef-d'oeuvre|`). Il va de soi qu'un tel mot ne doit pas être scindé par les analyseurs, mais cela montre à quel point la tokenisation peut s'avérer complexe.

La segmentation en phrases présente également ses difficultés. Elle s'appuie principalement sur les marqueurs de début et de fin de phrases que sont les majuscules et les ponctuations finales (point, point d'interrogation, d'exclamation, etc.). Mais il faut que l'outil de segmentation soit en mesure de discerner avec précision les points finaux des points d'abréviations ou de sigles (`i.e.`, `H.L.M.`), ou encore des points de formes numériques (*2.3*). Les nombres décimaux, dans la notation française, contiennent d'ailleurs généralement une virgule plutôt qu'un point, virgule qui dans ce cas exceptionnellement ne doit pas être considérée comme un caractère séparateur. Même chose pour les points d'exclamation ou d'interrogation, qui, comme dans des exemples donnés par Dister [1997, p.439] et tirés d'un corpus journalistique, ne marquent pas la fin d'une phrase :

« `Hélas ! aucun modèle économétrique ne permet de le calculer.` »

ou encore :

« `Comment Israël peut-il demander aux Palestiniens de faire la paix tout en confisquant leurs terres ? se demandait le patriarcat latin [...]` ».

Les majuscules ne sont pas non plus un indice infaillible de début de phrase lorsqu'elles sont précédées d'un point. Friburger *et al.* [2000] distinguent ainsi trois configurations dans lesquelles on peut trouver un point suivi d'une majuscule sans que cela ne marque une frontière entre deux phrases : les motifs anthroponymiques contenant une abréviation (par ex. `MM. Dupont et Dupond`, `J. Dupont`, etc.), les sigles dans leur ancienne notation (la nouvelle notation privilégie l'absence des points), les abréviations (par ex. `éd. Gallimard`, `cf. France-Italie en juin 2000`, etc.). Les auteurs citent un quatrième cas où au contraire la configuration peut amener le

tokeniseur à ne pas effectuer de segmentation alors qu'elle serait nécessaire. Lorsqu'une phrase se termine par une lettre majuscule, suivie donc par un point final, puis par une nouvelle majuscule de début de phrase, le segmenteur entraîné à reconnaître des motifs anthroponymiques comme ci-dessus est facilement induit en erreur. L'énoncé donné en exemple par Friburger *et al.* [2000, p. 184],
« `Ces aliments contiennent de la vitamine A, B et C. Durant me l'a confirmé.` »
doit bien être segmenté entre le point et la majuscule, et `C. Durant` ne doit donc pas être confondu avec l'initiale d'un prénom suivie d'un nom propre.

La tokenisation n'est donc pas un traitement aisé, il n'est pas anodin non plus, dans la mesure où toutes les erreurs de segmentation commises lors de ce traitement, aussi bien au niveau de la phrase que du *token*, se répercutent sur les étapes d'analyses suivantes, qui génèrent alors inévitablement à leur tour des résultats erronés. Rien qu'au niveau de la phrase, « avec une bonne liste d'abréviations et de noms propres, un algorithme à base de règles ou d'automates à états finis sera capable de découper un texte en phrases avec une efficacité entre 95% et 97% » [Doll & Coulombe, 2004, p. 40], ce qui laisse donc de manière non négligeable jusqu'à 5% d'erreurs de segmentation.

b) Étiquetage morphosyntaxique

L'étape qui suit la tokenisation consiste en l'étiquetage morphosyntaxique (ou *tagging*). L'étiqueteur attribue à chaque mot une ou plusieurs étiquettes (*tag*), contenant des informations sur sa catégorie grammaticale (verbe, nom, pronom, etc.), ainsi que sur ses traits de sous-catégorisation (genre, nombre, temps, personne, etc.), appelés aussi traits morphosyntaxiques. Ces *tags* proviennent d'un lexique de formes fléchies étiquetées. Il s'agit de lexiques qui contiennent, comme pour les vérificateurs orthographiques, « toutes » les formes fléchies de la langue, mais complétées par leurs caractéristiques morphosyntaxiques. Beaucoup de mots, dits « ambigus », reçoivent plusieurs *tags*. Ce sont les mots ayant des homographes avec des informations morphosyntaxiques différentes, comme le nom féminin singulier `bête`, qui a pour homographe l'adjectif singulier `bête` qui peut lui même être soit masculin soit féminin. La forme `bête` a donc les trois étiquettes suivantes :

`cat`	N
`genre`	f
`nombre`	s
`base`	bête

`cat`	A
`genre`	m
`nombre`	s
`base`	bête

`cat`	A
`genre`	f
`nombre`	s
`base`	bête

La forme `sommes`, qui peut désigner un nom masculin ou féminin, ou deux verbes conjugués, est un exemple ayant de très nombreuses étiquettes :

`cat`	N
`genre`	m
`nombre`	p
`base`	somme

`cat`	N
`genre`	f
`nombre`	p
`base`	somme

`cat`	V
`pers`	1
`nombre`	p
`temps`	prés
`mode`	ind
`base`	être

`cat`	V
`pers`	2
`nombre`	s
`temps`	prés
`mode`	ind
`base`	sommer

`cat`	V
`pers`	2
`nombre`	2
`temps`	prés
`mode`	subj
`base`	sommer

La forme `condition` en revanche n'est pas ambiguë et n'a donc qu'une seule étiquette :

`cat`	N
`genre`	f
`nombre`	s
`base`	condition

Cette multitude de *tags* pour un même mot peut facilement conduire à une mauvaise analyse morphosyntaxique du texte. Une désambiguïsation est donc parfois effectuée pour limiter le nombre d'étiquettes de ces mots et améliorer par la suite la détection des erreurs de grammaire. Une première approche pour désambiguïser les mots est l'approche probabiliste. Elle nécessite un corpus d'apprentissage sans erreur, étiqueté avec les informations morphosyntaxiques. Des calculs sont alors effectués. Il s'agit de la probabilité pour chaque mot d'avoir tel ou tel *tag*, ou encore la probabilité qu'un mot ait un certain *tag* en fonction des *tags* des mots qui l'entourent. Lors de l'étiquetage, ces probabilités sont appliquées à chaque mot du texte analysé, et chacun reçoit alors l'étiquette qui correspond à la plus forte probabilité. L'algorithme de Brill [1997] permet de générer des règles de désambiguïsation établies statistiquement.

Une autre approche consiste à utiliser des règles manuelles de désambiguïsation [Vergne & Giguet, 1998], sous forme d'expressions régulières et fondées sur le contexte immédiat. Chaque règle consiste en un modèle (ou *pattern*) d'un contexte en présence duquel tel mot prend tel *tag*.

Certains outils de vérification grammaticale utilisent l'approche probabiliste (par ex. GRAC de Biais [2005]), d'autres les règles manuelles (par ex. An Gramadóir de Scannell [2003]), certains combinent les deux approches (par ex. LanguageTool de Naber [2003a] dans sa première version), enfin d'autres ne font pas du tout de désambiguïsation (par ex. LanguageTool après avoir été réécrit dans un nouveau langage de programmation, et au début de la prise en charge du français).

c) Détection d'erreurs

La dernière étape concerne la vérification de la grammaire. Les correcteurs que nous avons analysés procèdent sur le principe du *pattern-matching*, c'est-à-dire sur la correspondance (*matching*) exacte entre un segment du texte et un modèle, ou motif (*pattern*) décrit dans une règle de correction. Il existe des règles de deux types : certains correcteurs utilisent des règles de grammaire, d'autres au contraire utilisent des règles d'erreurs.

GRAC fait partie des vérificateurs qui utilisent des règles de grammaire. Ces règles décrivent des modèles, des combinaisons de mots grammaticalement correctes. Les mots y sont représentés sous leur forme graphique (par ex. : `Quant + à`) et/ou de manière plus générique sous forme de catégories grammaticales et de traits morphosyntaxiques (par ex. : `Dét fém. sing. + Nom fém. sing.`). Chaque segment du texte à vérifier est comparé aux modèles contenus dans les règles de la base. S'il est agrammatical, aucune correspondance exacte ne sera trouvée entre lui et les règles de grammaire, puisqu'elles répertorient uniquement des constructions syntaxiques correctes. Une erreur sera alors signalée par l'outil.

D'autres correcteurs, comme LanguageTool, utilisent au contraire des règles d'erreurs. Le principe est le même que pour les règles de grammaire, à la différence près que ce sont cette fois les combinaisons de mots agrammaticales qui sont décrites dans les règles. Une erreur de grammaire est alors signalée par le correcteur lorsqu'un segment du texte et le modèle d'une règle coïncident.

Ce système de détection d'erreurs à base de règles présente des limites importantes, dont la principale tient au fait qu'il s'agit d'un système rigide qui impose que les segments de texte analysés coïncident parfaitement avec les *patterns* des règles. Ceci a pour conséquence que la détection d'une erreur échoue à partir du moment où il y a une différence, même minime, entre le segment de texte traité et le modèle dans la règle correspondante. Il peut s'agir d'un mot mal orthographié, mal étiqueté, ou encore d'un mot en plus ou en moins.

Ainsi, avec le système utilisant les règles de grammaire, des phrases grammaticales peuvent être à tort déclarées agrammaticales simplement parce qu'elles contiennent une combinaison de mots qui n'est pas décrite de manière exactement identique dans les règles, mais qui n'en est pas moins correcte. Ce phénomène de fausses alarmes est communément appelé « bruit » dans la détection des erreurs. À l'inverse, avec le système utilisant les règles d'erreurs, des phrases agrammaticales peuvent ne pas être repérées, et considérées comme correctes, si les structures erronées qu'elles contiennent ne sont définies dans aucune règle. Il s'agit dans ce cas de « silence ». Pour que la détection d'erreurs soit optimale, que ce soit avec les règles de grammaire ou avec les règles d'erreurs, il est donc nécessaire, mais cependant impossible, de répertorier absolument toutes les constructions syntaxiques possibles, soit correctes (pour les règles de grammaire), soit erronées (pour les règles d'erreurs).

Par ailleurs, le principe du *pattern-matching* restreint l'analyse aux mots et à leur contexte immédiat, ce qui rend impossible toute détection des erreurs qui impliquent des mots distants. Les correcteurs utilisant cette méthode font partie des outils de deuxième génération. La première génération correspond aux vérificateurs d'orthographe, qui ne considèrent que le mot. Aujourd'hui, les outils sont entrés dans la troisième génération, celle de la phrase, et non plus simplement du mot ou du groupe de mots. Clément *et al.* [2009] expérimentent par exemple un système fondé sur « une analyse profonde et complète de la phrase, par opposition aux grammaires superficielles et locales fréquemment utilisées », comme les règles de correction et le *pattern-matching* que nous venons d'aborder. Le principe présenté par les auteurs n'utilise donc pas de règles de ce type, mais construit à la place une forêt d'analyses de la phrase, et calcule pour chaque analyse le coût minimal de correction. Un coût nul suppose l'absence d'erreur, puis « les coûts de correction des alternatives permettent d'ordonner les propositions de correction par ordre de plausibilité » [Clément *et al.*, 2009].

Les logiciels commerciaux appartiennent également plutôt à la troisième génération de correcteurs. C'est le cas par exemple du Correcteur 101 [Doll & Coulombe, 2004], qui s'appuie notamment sur les grammaires de dépendance et la théorie Sens-Texte de Mel'čuk [1988a].

La détection des erreurs fondée sur le *pattern-matching* n'est donc pas une solution optimale, mais même si d'autres méthodes existent, c'est celle qui est encore utilisée par les correcteurs libres, et donc la seule que nous avons pu observer.

d) Traitement facultatif : le *chunking*

Dans le processus d'analyse des textes pour la recherche d'erreurs, certains correcteurs effectuent un traitement supplémentaire par rapport à ceux que nous venons de présenter. Suite à l'étiquetage morphosyntaxique (voir figure 2.2 p. 33), ces outils opèrent une segmentation à un niveau intermédiaire entre la phrase et le mot, avec la délimitation de groupes de mots appelés *chunks* [Abney, 1991], ou syntagmes minimaux [Giguet, 1998], ou encore syntagmes non récursifs (SNR) [Vergne & Giguet, 1998]. En s'inspirant d'Abney [1991], Lebarbé [2002, p. 32] donne la

définition suivante : « Le chunk est constitué d'un mot lexical (au sens restreint : nominal ou verbal) entouré d'une constellation de mots fonctionnels (au sens large : déterminants, pronoms, adverbes, adjectifs...) ».

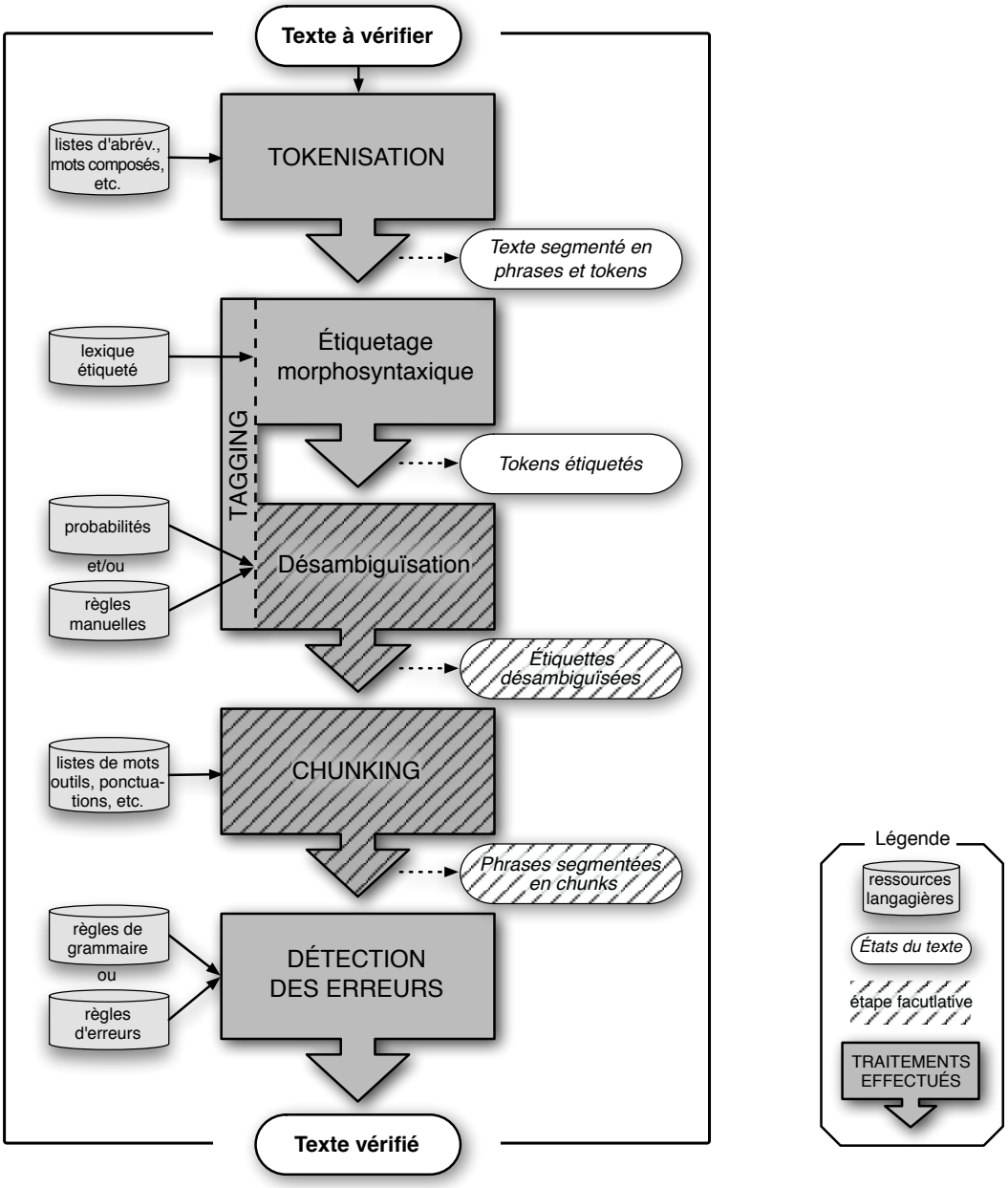


FIGURE 2.2 : Structure générale d'un correcteur grammatical effectuant un *chunking*

Il s'agit d'une unité, que nous nommerons indifféremment *chunk* ou syntagme, définie non pas en fonction de son contenu, mais en fonction de ses frontières, principalement des mots grammaticaux (conjonctions, prépositions, pronoms, déterminants, etc.), des marques morphologiques (flexions majoritairement) et des ponctuations. Les syntagmes sont ainsi souvent aisés à repérer

et délimiter. De plus, ils ne peuvent ni se chevaucher ni être imbriqués, et par conséquent le début d'un *chunk* induit la fin du précédent. Nous donnons ci-dessous un exemple de découpage d'un énoncé en syntagmes (nominaux, verbal et prépositionnels) :

$[\texttt{Le TAL}]_{Nom.}\,[\texttt{connait}]_{Ver.}\,[\texttt{un grand nombre}]_{Nom.}\,[\texttt{de domaines}]_{Prp.}\,[\texttt{d'applications}]_{Prp.}\,[\texttt{.}]_{Pct.}$

Au sein d'un syntagme, les mots fonctionnels sont dépendants de la tête lexicale, ce qui fait de la structure interne des *chunks* une structure relativement figée, dans laquelle les contraintes d'accords sont fortes [Giguet, 1998]. L'intérêt d'une segmentation à ce niveau de granularité apparaît alors comme évident pour un outil de vérification grammaticale, dont une des tâches principales est de s'assurer que les accords, très nombreux en français, sont correctement effectués.

La segmentation en *chunks* peut ainsi être très utile dans le TAL [Abney, 1991], et nous intéresse particulièrement dans le cadre de nos travaux, pour l'outil que nous souhaitons développer. Elle est cependant encore peu utilisée par les correcteurs libres que nous avons étudiés. An Gramadóir par exemple n'y a pas recours. Quant à LanguageTool, il effectuait un *chunking* assez basique dans sa première version, mais cette fonctionnalité a disparu dans la version suivante, car elle avait peu d'intérêt pour les langues traitées par l'outil (l'anglais principalement) et générait des erreurs d'analyse. Il subsiste une forme de *chunks* implicites, qui sont définis non pas par leurs frontières, mais par leur contenu, en se conformant à des modèles prédéfinis. Leur intérêt s'en retrouve fortement amoindri.

Les *chunks* étant un des principes du TAL sur lesquels s'appuient notre travail, nous y revenons de manière détaillée dans la troisième partie de cet exposé (*cf.* section 8.1.3 p. 179) et ne développons donc pas davantage cette notion dans le présent chapitre.

e) Correction vs. vérification

Pour les correcteurs grammaticaux, comme pour les correcteurs d'orthographe, la dénomination est maladroite. Le mot « correcteur » nous laisse à penser que l'outil va effectuer des corrections tout seul, de manière automatique. Ce n'est pourtant pas le cas. En effet, il vérifie le texte et recherche les erreurs, il explique éventuellement le type de problème rencontré dans le texte afin d'aider l'utilisateur à le rectifier, mais en aucun cas il ne corrige automatiquement les erreurs. La dénomination anglaise est beaucoup plus adaptée avec *grammar checker*, c'est-à-dire « vérificateur de grammaire ». Il nous semble en effet bien plus adéquat pour désigner les correcteurs grammaticaux d'utiliser les termes « détecteurs » ou « vérificateurs », qu'il faudrait associer au terme d'« incohérence », plutôt qu'« erreur ». En effet, ces outils repèrent des éléments qui ne sont pas cohérents avec ce qu'ils attendent normalement dans une phrase grammaticale, mais il ne s'agit pas forcément d'erreurs. Il peut s'agir simplement d'une formation syntaxique non prévue par les règles de correction, mais qui n'est pas erronée pour autant. Comme l'indique Jacquet-Pfau [2001, p. 85], dans le traitement de la langue par les outils informatiques, « est considéré comme erreur tout ce qui n'est pas reconnu par le système ».

Jacquet-Pfau [2001] et Doll & Coulombe [2004] utilisent également l'expression « Correction Assistée par Ordinateur » (soit CAO), ce qui correspond bien à l'utilisation qui est faite des outils de vérification de la langue, mais reste trop générique en ne précisant pas le type de correction effectuée. Nous préférons privilégier désormais l'expression « vérificateur grammatical », ou « détecteur d'incohérences grammaticales », pour nommer ce qui est appelé « correcteur grammatical » dans l'usage courant.

2.2 De la grammaire académique à la grammaire en bureautique

Les vérificateurs orthographiques ont été initialement conçus pour les besoins des professionnels de la bureautique (secrétaires, journalistes, etc.). Ce domaine de l'informatique regroupe « l'ensemble des techniques et des moyens tendant à automatiser les activités de bureau et principalement le traitement et la communication de la parole, de l'écrit et de l'image » [7]. Ceci inclut, entre autre, les outils pour la production de documents, et donc pour l'aide à leur rédaction. Les vérificateurs linguistiques sont ainsi considérés comme des outils bureautiques. C'est pour cette raison que nous parlons de grammaire en bureautique, plutôt que de manière plus générale en informatique. Cependant, ces outils ne sont aujourd'hui plus limités aux seuls professionnels, et couvrent un large champ d'applications permettant la saisie de texte plus ou moins élaboré (traitement de texte, mail, web, graphisme, etc.). Nous engloberons donc tous ces types d'applications lorsque nous emploierons le terme de bureautique dans les pages suivantes.

La présentation que nous venons de faire des outils chargés de traiter les erreurs d'écriture tapuscrites nous a permis de constater qu'il y a, dans ce que nous appelons bureautique, une distinction très nette entre le domaine de l'orthographe et celui de la grammaire, et que ces domaines sont délimités en fonction de la catégorie de l'outil compétent pour traiter les phénomènes linguistiques qui leur appartiennent. Mais ces domaines ne coïncident pas tout à fait avec la définition que nous en avons faite dans le chapitre précédent. Pour parler de la grammaire du côté de l'ordinateur, nous revenons donc sur la dichotomie orthographe/grammaire, telle qu'elle est perçue en bureautique, en regard de la perception linguistique. Nous nous concentrerons ensuite sur le seul côté grammatical, objet principal de nos travaux.

2.2.1 La grammaire des outils bureautiques

En informatique bureautique, nous venons de voir (*cf.* § 2.1 p. 25) que l'orthographe est vérifiée par des outils qui ne s'occupent que de la graphie. Ils ne font, contrairement à la linguistique, aucune distinction entre les erreurs sur les radicaux, sur les flexions ou sur les mots invariables. Ils se contentent de repérer les mots qui leurs sont inconnus et de les signaler comme erronés, tels les exemples *`fâce` ou *`choisient`.

Les erreurs qui ne sont pas détectables par un vérificateur d'orthographe, comme par exemple *`tu insère` qui ne contient aucun mot inconnu, sont regroupées dans la catégorie grammaire. Cette catégorie englobe les erreurs qui portent sur la syntaxe (ordre, omission, répétition de mots), et les erreurs qui portent sur les parties variables des mots, à condition qu'elles ne génèrent pas un mot inconnu, auquel cas elles deviennent du ressort du vérificateur d'orthographe.

Si nous prenons par exemple l'extrait *`les américains ont déjà choisient`, pour l'ordinateur il n'y a aucun doute sur le fait que le mot *`choisient` constitue une erreur orthographique. Cette graphie ne correspond en effet à aucun mot des lexiques. Pour le linguiste, il s'agit-là d'un problème d'orthographe grammaticale. Nous sommes en effet en présence d'un mot auquel une flexion erronée a été appliquée. De même, dans *`Les effets spécials` l'ordinateur décèle une erreur d'orthographe, alors que le linguiste identifie une erreur grammaticale due à la méconnaissance de la flexion du mot au masculin pluriel. Selon le point de vue adopté, ces erreurs d'écriture sont donc à cheval sur l'orthographe et la grammaire.

Ainsi, en dehors des erreurs de construction syntaxique qui restent dans la catégorie gram-

7. Définition du Journal officiel de la République française du 17 janvier 1982.

maticale quel que soit le point de vue, une même erreur d'écriture peut ne pas être catégorisée de la même façon selon le point de vue linguistique, où les erreurs sont plutôt classées en fonction de l'explication de leur origine, et le point de vue informatique bureautique, où les erreurs sont classées en fonction du traitement pour les détecter.

Il en découle un chevauchement des niveaux orthographique et grammatical dans ces deux disciplines, que nous représentons dans la figure 2.3. Nous y voyons en blanc des exemples d'erreurs qui sont toujours de l'ordre de l'orthographe lexical (erreurs de graphie), en gris des erreurs toujours d'ordre grammatical, et au milieu, en hachuré, des erreurs à cheval sur les deux catégories.

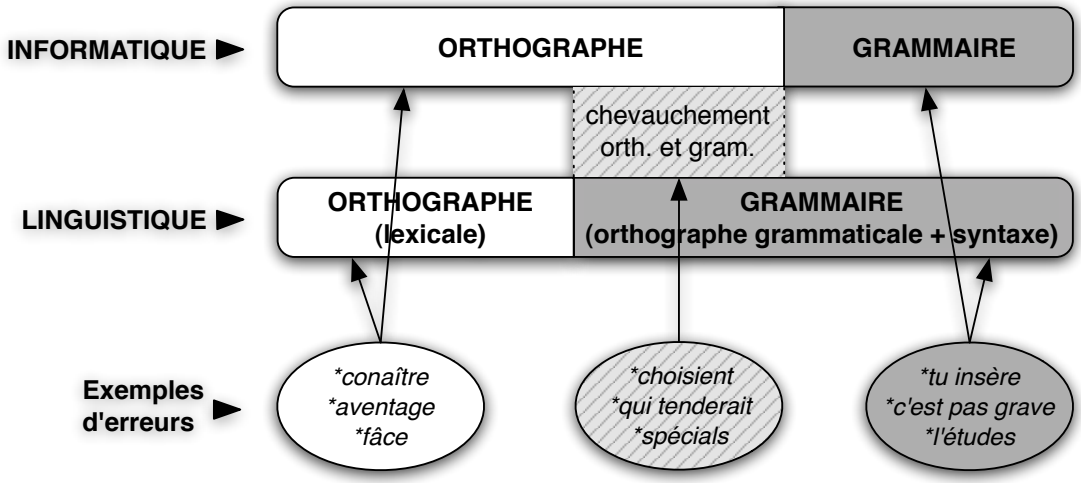


FIGURE 2.3 : Chevauchement de la grammaire et l'orthographe, en linguistique et en informatique.

2.2.2 Les types d'erreurs

Les erreurs de grammaire, pour un vérificateur informatique, sont donc les erreurs morpho-syntaxiques (*cf.* définition 1.9 p. 22), à l'exception de celles qui génèrent un mot inconnu. Nous ne citons pas ici toutes celles qui existent, car comme nous l'expliquions précédemment (*cf.* § *Détection d'erreurs* p. 31) il n'est pas possible de toutes les inventorier. Nous ne pouvons pas non plus énumérer toutes celles qui sont prises en charge par les outils actuels, en dehors des logiciels libres dont les règles sont accessibles, car aucune liste exhaustive n'est publiée pour les outils propriétaires. Nous mentionnons simplement les erreurs que nous avons vues détectées par le plus grand nombre de vérificateurs grammaticaux, ainsi que celles dont il est souvent question dans la littérature à cause des problèmes qu'elles posent.

Toutes les erreurs qui sont normalement du ressort des vérificateurs grammaticaux ne sont donc pas prises en charge par tous les outils. Les erreurs les plus communément traitées sont celles concernant les accords. La plupart des outils est capable d'en détecter un certain nombre, dont les erreurs d'accord en genre et en nombre dans les syntagmes nominaux (déterminant-nom, nom-adjectif), les erreurs d'accord des attributs ou des verbes avec le sujet, des participes passés avec le sujet ou avec l'objet, les erreurs d'accord des adjectifs de couleur, des numéraux, etc. Le traitement d'une partie de ces erreurs reste toutefois limité en présence de certaines constructions

syntaxiques qui nécessiteraient une analyse sémantique pour la désambiguïsation des accords (par ex. avec le rattachement de l'adjectif dans *`le langage des signes américains`).

Les outils détectent également les erreurs fréquentes dues à l'homophonie ou la paronymie. Les confusions les plus souvent prises en charge concernent les mots grammaticaux homophones (`on`/`ont`, `sa`/`ça`, `quand`/`quant`, etc.) et la terminaison des infinitifs et des participes passés (*`va retrouvé`/`va retrouver`), ou de manière plus générale les flexions verbales qui se prononcent /e/ (-er, -é, -ez) ou /ɛ/ (-ait, -ai, etc.).

Certains vérificateurs grammaticaux, comme Antidote, sont capables aussi de repérer des erreurs concernant les modifications graphiques des mots selon leur place dans la phrase, comme par exemple l'euphonie (*`si il`/`s'il`, *`ce outil`/`cet outil`), ou encore le tiret des formes composées à valeur grammaticale (*`est elle`/`est-elle`). Certains aussi sont capables de détecter des erreurs de ponctuation, comme l'appariement des parenthèses ou des guillemets, l'utilisation des espaces, les points finaux non suivis d'une majuscule, etc.

Beaucoup d'autres types d'erreurs sont parfois traités par les vérificateurs les plus performants (généralement les outils commerciaux de troisième génération, tel Antidote, Cordial, etc.), comme par exemple l'utilisation d'un mauvais pronom relatif, d'un mauvais mode de conjugaison, d'un mauvais auxiliaire, ou encore comme la construction erronée d'une négation, d'une collocation, etc.

En revanche, un grand nombre d'erreurs grammaticales n'est quasiment jamais pris en charge par les outils de vérification, ou bien de manière très incomplète. Tisserand [2004] soulève par exemple les difficultés posées par les mots composés et leurs accords. « Selon le sens que l'on veut donner à [une] suite de mots, les accords ne seront pas les mêmes » [Tisserand, 2004, p. 175]. L'auteur illustre ses propos avec l'exemple `des armoires à linge(s) sale(s)`, qui peut avoir différentes significations (des armoires sales qui contiennent du linge, des armoires qui contiennent du linge sale, etc.) selon que `linge` et/ou `sale` sont ou non au pluriel.

Une autre grande difficulté de la vérification grammaticale concerne les erreurs qui impliquent des éléments distants, dans une même phrase ou de manière plus complexe dans des phrases distinctes, et qui nécessitent la résolution des anaphores par l'analyseur morphosyntaxique pour être détectables. Si des travaux sont en cours sur cette question [Boudreau & Kittredge, 2005], il est encore très difficile en TAL d'identifier les référents de mots éloignés. De plus, il faudrait des outils de vérification grammaticale dont le modèle de la langue considère le texte dans son intégralité pour pouvoir résoudre les anaphores dans des énoncés différents, mais les outils actuels se limitent encore à la phrase [Doll & Coulombe, 2004].

2.2.3 Conclusion

Nous avons ainsi vu dans ce chapitre comment sont traitées, par les outils informatiques de vérification de la langue, les erreurs que nous commettons à l'écrit. Communément appelés correcteurs, nous avons vu qu'en fait aucun de ces outils n'effectue de réelle correction automatique, raison pour laquelle nous préférons les désigner par le terme de vérificateurs.

Certains d'entre eux sont spécialisés dans la vérification de l'orthographe, et d'autres dans celle de la grammaire, mais ces deux domaines ne doivent pas être délimités comme ils le sont dans l'approche linguistique (*cf.* chapitre *Approche linguistique de la notion d'erreur de grammaire* p. 7). En effet, sous le terme « orthographe » associé aux vérificateurs orthographiques, sont en

fait réunis à la fois l'orthographe lexicale et une partie de la grammaire de l'approche linguistique. Pour ces outils, une orthographe correcte est une graphie qui existe dans leur lexique, peu importe qu'elle soit, ou non, adaptée au contexte syntaxique dans lequel elle apparaît. Un accord mal réalisé qui conduit à l'écriture d'un mot inconnu du lexique (par ex. *`Les effets spécials`) est une erreur d'orthographe pour un vérificateur orthographique, même si le linguiste la considère plutôt comme d'origine grammaticale.

Mais ce ne sont pas ces outils qui nous intéressent dans le cadre de cette thèse. Notre travail se focalise au contraire sur les erreurs de grammaire qu'ils ne traitent pas, ainsi que sur les vérificateurs grammaticaux destinés à prendre en charge ces erreurs. Les premiers outils de ce type réalisent des analyses au niveau des groupes de mots, qu'ils comparent à des modèles dans des règles de correction afin de détecter les séquences erronées. Cette méthode permet d'identifier efficacement certains types d'erreurs morphosyntaxiques, dont les erreurs d'accords dans les contextes locaux principalement. Elle marque le début de la vérification grammaticale et est encore utilisée aujourd'hui dans les logiciels libres, mais elle est très limitée, notamment pour le français. Notre objectif est donc de proposer un modèle de détection d'incohérences grammaticales fondé sur des méthodes mieux adaptées au traitement du français. Une nouvelle génération de vérificateurs existe déjà, qui utilise des modèles de traitement améliorant sensiblement les performances de la vérification grammaticale, en effectuant des analyses de la phrase complète notamment, et non plus seulement de groupes de mots. Mais ce sont pour le moment principalement des logiciels commerciaux dont le fonctionnement est opaque. Nous nous plaçons au contraire dans une optique d'outil libre en proposant un modèle de détection d'erreurs documenté et librement accessible, qui tendra également à combler un manque dans le domaine.

Chapitre 3

Etat des lieux des outils et études sur les erreurs

Sommaire

3.1 Documentation et fonctionnement des vérificateurs existants . 39

3.1.1 Des outils très peu documentés 40

3.1.2 Un fonctionnement limité 41

3.1.3 Les utilisateurs livrés à eux-mêmes 45

3.2 Panorama des études sur les erreurs tapuscrites 53

3.2.1 Les études existantes 54

3.2.2 Spécificité des tapuscrits 57

Dans le chapitre précédent, nous avons présenté le chevauchement de la grammaire et de l'orthographe en informatique et linguistique. Il nous confirme que l'analyse et le traitement d'une même erreur seront souvent différents selon qu'ils seront effectués par l'humain ou par l'ordinateur. Ce qui nous intéresse dans cette thèse, ce sont justement l'analyse et le traitement par l'ordinateur des erreurs tapuscrites, et plus précisément la détection et l'explication des incohérences grammaticales. Il nous faudra pour cela contribuer à pallier des faiblesses dans le domaine, tels les outils de vérification encore incomplets et imparfaits, ou encore l'absence d'étude spécifique des erreurs tapuscrites à détecter.

3.1 Documentation et fonctionnement des vérificateurs existants

Le début de nos travaux sur la vérification grammaticale automatique a été l'occasion de nous intéresser aux divers outils existants, que nous divisons en deux catégories : les logiciels libres, documentés mais aux résultats assez limités, et les logiciels propriétaires, assez performants mais non documentés. Les deux nécessiteraient une amélioration plus ou moins importante des explications qu'ils renvoient à l'utilisateur au sujet de ses erreurs.

3.1.1 Des outils très peu documentés

Plusieurs auteurs ([Doll & Coulombe, 2004], [Vienney, 2004], [Piolat, 2007]) s'accordent à dire que, depuis une vingtaine d'années, la vérification automatique de textes semble avoir migré des laboratoires de recherche vers le terrain industriel et commercial. Elle est « une des technologies du TAL les plus utilisées du grand public » [Clément *et al.*, 2009], mais semble en effet négligée par la recherche publique. Les auteurs avancent plusieurs hypothèses quant à l'origine de ce constat.

Clément *et al.* [2009] évoquent par exemple l'importance pour un vérificateur linguistique de pouvoir être intégré à un traitement de texte pour accroître son utilité pour le public. Or, ceci n'est vraiment possible que depuis l'arrivée de la suite libre OpenOffice.org au début des années 2000. Jusque-là, le quasi monopole [1] de la suite commerciale Office de Microsoft sur ce type de logiciels ne permettait pas de réaliser une telle intégration.

Par ailleurs, selon Doll & Coulombe [2004], les performances atteintes aujourd'hui par certains outils de vérification commerciaux sont considérées comme satisfaisantes, et de coûteux efforts supplémentaires n'apporteraient pas d'amélioration significative. D'autant que la recherche publique ne dispose pas des mêmes moyens que des firmes comme Microsoft, qui « continue d'investir des sommes considérables dans la recherche et le développement de tels outils » [Fontenelle, 2005a, p. 119].

La recherche privée prédomine ainsi sur le sujet de la vérification automatique, qui est finalement perçue par les chercheurs comme le domaine réservé de grands industriels [Doll & Coulombe, 2004], avec pour conséquence un certain désintérêt pour le sujet dans les laboratoires. Ceci se remarque notamment au niveau des publications sur la question de la vérification automatique, qui sont peu nombreuses. Leur faible quantité ne peut par ailleurs pas compenser les publications sur les logiciels commerciaux, qui sont également peu nombreuses, et en général délibérément évasives, pour des raisons évidentes de secret industriel.

Les documents des industriels ne présentent jamais, ou alors de manière très superficielle, les techniques ou les ressources langagières utilisées par l'outil dont ils traitent. Ils n'exposent généralement que les performances, les erreurs linguistiques prises en charges, ou encore le mode d'emploi. L'article de Brunelle [2004] au sujet du logiciel Antidote en est un exemple. Il apporte des données chiffrées très précises, il mentionne les langages de programmation utilisés pour coder les diverses couches du programme, mais ne donne aucune indication algorithmique. Il signale simplement que l'analyseur utilise une grammaire de dépendance qui génère des arbres. Nous apprenons également dans l'article que le logiciel utilise environ 2000 règles syntaxiques, mais rien n'est dit sur la manière dont elles sont formalisées, ni sur l'approche linguistique de l'erreur. Même constat pour les mots du dictionnaire, dont il est simplement dit qu'ils ont « fait l'objet d'une description formelle poussée, faite entièrement à la main par un linguiste, et vérifiée par un compilateur » [Brunelle, 2004, p. 28].

Fontenelle est un peu plus loquace dans quelques articles au sujet de la vérification de textes dans la suite Office de Microsoft [2]. Dans [Fontenelle, 2006] par exemple, il explique que les erreurs, et notamment leur fréquence, sont étudiées à partir de corpus de textes authentiques, mais qu'il est parfois fait usage, dans des cas particuliers de tests, de logiciels du type « pourrisseur

1. Les concurrents à l'époque étaient également des solutions propriétaires : StarOffice de Sun Microsystems, ou encore Lotus Symphony de Lotus Development Corporation.
2. Microsoft Corporation

de textes » [Véronis, 2005]. Ces outils permettent d'ajouter automatiquement dans les textes des erreurs qui peuvent être ciblées (par ex. des erreurs de confusion entre les homophones grammaticaux), et facilitent ainsi l'évaluation du moteur de correction sur des types d'erreurs donnés. Un corpus de textes authentiques ne contient en effet pas forcément une fréquence suffisante de certaines erreurs pour tester leur détection de manière fiable.

Fontenelle [2006] aborde également la question des caractères séparateurs pour le tokeniseur, mais s'il dit clairement que le trait d'union en fait partie, nous devinons, sans certitude, que ce n'est pas le cas de l'apostrophe. De manière générale, l'auteur ne s'attarde pas sur les techniques employées pour l'analyse et la détection des erreurs. Il donne presque exclusivement des indications sur les ressources langagières utilisées par l'outil, comme des exemples de codes grammaticaux, ou d'informations lexicales ou sémantiques, associés aux entrées des dictionnaires électroniques, dans [Fontenelle, 2004, 2005a].

Au contraire, la thèse de Naber [2003b] décrit avec précision toute la structure et les ressources de LanguageTool, l'outil libre de vérification orthographique et grammaticale de l'anglais qu'il a conçu. L'auteur détaille l'architecture logicielle, les outils existants qu'il a utilisés, comme par exemple le module de vérification orthographique, ou encore l'étiqueteur morphosyntaxique, qui s'appuie sur un *tagger* effectuant une désambiguïsation statistique, qu'il a complétée par des règles manuelles. L'auteur présente aussi la manière dont la catégorie des mots inconnus est inférée à partir notamment de la terminaison des mots, souvent caractéristique d'une catégorie en anglais (fin en `-ly` = adverbe ; fin en `-ise`, `-ize`, `-ate` ou `-fy` = verbe ; etc.). De même, la structure et le fonctionnement des règles de correction, l'interface utilisateur, les évaluations sur corpus, etc., n'échappent pas aux descriptions et aux explications précises de l'auteur.

Dans [Clément *et al.*, 2009] également, même s'il ne s'agit que d'un outil prototypique, les auteurs expliquent de façon assez détaillée le principe de fonctionnement, fondé sur une grammaire profonde qui construit une forêt d'analyses de la phrase. Les auteurs décrivent notamment comment sont effectués différents calculs, dont celui de proximité entre deux analyses, afin de calculer les coûts minimaux de correction de chaque phrase. Ils donnent également un exemple de leur lexique dans le formalisme qu'ils utilisent.

Mais ces documents assez complets et précis sur la vérification grammaticale, qui concernent généralement soit des outils libres, soit des recherches universitaires, sont trop peu nombreux. Ceci ne contribue pas à accélérer l'évolution et l'amélioration des outils en dehors du secteur commercial. Si certains logiciels propriétaires ont atteint un niveau de performance satisfaisant, il n'en va pas de même pour les outils libres.

3.1.2 Un fonctionnement limité

D'importantes améliorations ont été apportées aux vérificateurs de textes depuis leur apparition. Comme nous l'évoquions dans le chapitre précédent (*cf.* § c) p. 31) et comme l'expliquent Doll & Coulombe [2004], la toute première génération de ces outils prenait le mot comme unité de traitement pour en vérifier la graphie, et permettait la détection d'environ 40% des erreurs des textes. La seconde génération a accru de manière notable le taux de détection d'erreurs (environ 60%) en prenant en considération les mots avec leur contexte immédiat et en utilisant des règles d'analyses syntaxiques locales. Les vérificateurs de la troisième génération sont les plus performants. En procédant généralement à des analyses complètes de la phrase, au lieu d'analyses locales, ils sont proches des 80% d'erreurs détectées, toujours selon Doll & Coulombe [2004].

Nous synthétisons ces trois générations d'outils dans le tableau 3.1 ci-après.

<table>
<tr><th></th><th>1re génération</th><th>2e génération</th><th>3e génération</th></tr>
<tr><td>Niveau d'analyse</td><td>mot</td><td>mot + contexte local</td><td>phrase</td></tr>
<tr><td>Méthode de détection</td><td>lexique de formes fléchies</td><td>lexique étiqueté + règles d'analyse locales</td><td>lexique étiqueté + règles locales + analyse profonde de la phrase</td></tr>
<tr><td>Principales erreurs détectées</td><td>orthographe lexicale</td><td>orthographe, accords (surtout genre et nombre)</td><td>orthographe, accords (SN, verbes, part. passés), conjugaison, homophonie, syntaxe, etc.</td></tr>
<tr><td>Taux d'erreurs détectées</td><td>≃ 40%</td><td>≃ 60%</td><td>≃ 80%</td></tr>
<tr><td rowspan="2">Exemples de logiciels</td><td colspan="2">vérificateurs intégrés aux suites bureautiques, etc.</td><td rowspan="2">Antidote, Cordial, Correcteur 101, etc.</td></tr>
<tr><td>vérificateurs d'orthographe intégrés aux systèmes d'exploitation, etc.</td><td>LanguageTool, An Gramadóir, etc.</td></tr>
</table>

Tableau 3.1 : Tableau récapitulatif des différentes générations de correcteurs

Sans connaître précisément les techniques utilisées par les logiciels commerciaux, nous savons toutefois que ces derniers appartiennent pour la plupart à cette troisième génération d'outils. Leurs concepteurs le mentionnent parfois, de manière explicite pour le Correcteur 101 par exemple [Doll & Coulombe, 2004], un peu moins clairement pour Antidote [Brunelle, 2004]. Il existe encore beaucoup d'erreurs que ces vérificateurs ne peuvent pas détecter. Nous évoquions par exemple dans le paragraphe 2.2.2 p. 37 les problèmes posés par la coréférence et la résolution des anaphores. Mais les logiciels de vérification actuels se heurtent au problème du sens plus généralement. « [I]ls se trouvent rapidement limités face à des erreurs ou ambiguïtés dont la résolution nécessite la prise en compte d'éléments sémantiques » [Thouet, 2004, p. 155]. Or, il s'agit-là d'un pan de la linguistique dans lequel le TAL est encore balbutiant. Les performances de ces vérificateurs commerciaux de troisième génération sont considérées aujourd'hui comme satisfaisantes, en dépit des difficultés qu'ils rencontrent encore. Ce n'est en revanche pas le cas des outils qui nous intéressent plus précisément dans cette thèse, à savoir les vérificateurs libres, qui font l'objet d'une forte attente de la part des utilisateurs. Ceux que nous avons analysés, principalement LanguageTool et An Gramadóir, font partie des outils de seconde génération, qui travaillent au niveau des groupes de mots. Comme nous l'avons indiqué plus haut (*cf.* § *Détection d'erreurs* p. 31), ils présentent des limites importantes. Ces limites trouvent leur origine majoritairement dans l'utilisation d'une grammaire de surface, par opposition aux grammaires profondes des vérificateurs de troisième génération. En effet, Clément *et al.* [2009] expliquent que, pour réduire la propension au bruit et au silence des analyses superficielles dans la détection d'erreurs, les grammaires de surface se bornent à des règles locales. C'est précisément l'utilisation de règles de ce type qui limite les performances des outils libres, pour diverses raisons que nous

avons abordées dans [Souque, 2007, 2008], et que nous présentons ici.

a) Un nombre considérable de règles...

Tout d'abord, l'utilisation de règles locales s'appuie sur une comparaison entre les suites de mots du texte et des suites de mots décrites dans les règles. C'est ce que nous appelons le *pattern-matching*. Pour qu'une erreur soit détectée avec des outils comme LanguageTool ou An Gramadóir qui utilisent des règles d'erreurs, celle-ci doit être décrite précisément dans une règle. Il est alors nécessaire de créer autant de règles que d'erreurs potentielles, et leur nombre devient vite considérable. Prenons pour exemple un groupe nominal constitué d'un déterminant, d'un nom et d'un adjectif. Les possibilités de genre et de nombre pour chacun sont multiples, sachant qu'ils peuvent être masculin, féminin ou épicène, et singulier, pluriel ou invariable. Il y a ainsi théoriquement neuf combinaisons genre-nombre pour chacun, ce qui conduit à 729 (soit 9^3) possibilités pour la suite déterminant + nom + adjectif de notre groupe nominal, et 1458 (soit 729*2) si nous tenons compte de l'inversion possible du nom et de l'adjectif.

Au fur et à mesure de l'ajout d'un ou plusieurs adjectifs, qui peuvent prendre des places différentes dans le syntagme nominal, le nombre de combinaisons incorrectes s'accroit exponentiellement. An Gramadóir dispose par exemple d'une liste de 447 règles uniquement pour les syntagmes nominaux, et il ne prend pas en compte dans ces règles les mots épicènes ou invariables. Ils sont pourtant étiquetés comme tels dans le lexique, de façon pas toujours très cohérente par ailleurs. Le nom `élève` fait ainsi l'objet de deux entrées dans le lexique : la première comme nom épicène singulier, la seconde, redondante, comme nom féminin singulier. Les mots du lexique, ou plus précisément les combinaisons genre-nombre ne sont ainsi pas toutes incluses aux règles de détection d'erreurs, dont le nombre atteindrait plusieurs centaines de milliers si c'était le cas. Un fichier contenant un nombre de règles aussi important est bien trop lourd à traiter pour un système informatique. Toutes les erreurs ne peuvent alors pas être détectées.

b) ... qui provoque des détections redondantes...

De cette grande quantité de règles, associée à leur application locale dans le texte, découle un problème de détections multiples et redondantes d'erreurs pour un même mot ou un même groupe de mots. L'exemple que nous donnons dans [Souque, 2007] illustre ce point. Avec les règles de LanguageTool en 2007, l'extrait (1) ci-dessous déclenchait trois alertes pour une seule véritable erreur sur *`ont`* :

(1) **`ont a tendance à croire...`*

Alerte 1 : *`ont`* ⇒ « utiliser le pronom personnel `on` au lieu de l'auxiliaire `ont` ».
règle : *ont + verbe conjugué*

Alerte 2 : `a` ⇒ « il faut un participe passé après l'auxiliaire `ont` ».
règle : *auxiliaire + verbe conjugué*

Alerte 3 : `a` ⇒ « utiliser la préposition `à` au lieu de l'auxiliaire `a` ».
règle : *verbe + a*

Seule l'application de la première règle est justifiée ici. Le deux autres, toutes deux concernant `a`, constituent du bruit.

c) ...et ne peut pas décrire l'infinité des erreurs possibles

La localité des règles a également pour conséquence que seules les erreurs impliquant des éléments très proches peuvent être recherchées. Tout ce qui concerne des relations distantes est hors de portée des vérificateurs qui utilisent cette technique. Dans un énoncé comme en (2) ci-dessous, où le sujet et le verbe sont très éloignés, une erreur sur le verbe (ici `a provoqué`) n'est pas détectable.

(2) *`Les effluves d'une décoquetion de piment, que le cuisinier d'un restaurant thaïlandais de Londres préparait, a provoqué une alerte terrotiste en plein centre de la capitale britannique, a rapporté mercredi le quotidien Times.`

Nous avons également indiqué que l'utilisation de règles locales, associée au *pattern-matching*, oblige à prévoir toutes les erreurs possibles et à les décrire dans les motifs de règles, pour être en mesure de les détecter. Mais, comme le dit une expression de Chomsky [1969, p. 33], reprise au linguiste allemand von Humboldt [1836]), « le locuteur fait un usage infini de moyens finis ». Une infinité de suites de mots peuvent être produites, ce qui rend irréalisable une telle tâche de description exhaustive des combinaisons erronées. D'autant plus que la langue est en constante évolution et que des mots nouveaux apparaissent en permanence. Victor Hugo l'évoque dans sa préface de Cromwell :

> « [...] la langue française n'est pas fixée et ne se fixera point. Une langue ne se fixe pas. [...] Toute époque a ses idées propres, il faut qu'elle ait aussi les mots propres à ces idées. Les langues sont comme la mer : elles oscillent sans cesse »
>
> [Hugo, 1827, p. 73-74]

Un texte peut toujours contenir un mot nouveau que le vérificateur ne connaîtra pas et qui le fera échouer dans sa recherche d'erreurs. Les mots inconnus sont un problème, mais pas tout à fait au même titre que les mots mal orthographiés. Ces derniers sont considérés comme inconnus tant qu'ils n'ont pas fait l'objet d'une vérification et d'une correction orthographique, mais deviennent par la suite des mots connus et analysables. Le problème le plus évident posé par les mots inconnus (néologismes, emprunts, noms propres, mots spécialisés, etc.) est qu'ils ne figurent dans aucune règle. Toute erreur les impliquant est irrémédiablement passée sous silence. L'énoncé (3) ci-dessous contient par exemple le mot `monosémique`, qui est inconnu des dictionnaires utilisés par les correcteurs libres LanguageTool et Grammalecte [Ronez, 2011] :

(3) *`Les termes techniques utilisés sont monosémique.`

Il en résulte qu'aucun des deux vérificateurs grammaticaux ne détecte l'erreur d'accord, alors qu'une erreur similaire sur un mot connu (par ex. `long` au lieu de `monosémique`) serait détectée par Grammalecte. Pour ce qui est de LanguageTool, la longueur du groupe nominal sujet empêche toute détection d'erreur sur l'attribut du sujet, même si tous les mots sont connus. La règle correspondante ne prévoit en effet que les cas où le groupe nominal sujet n'est constitué que d'un déterminant et un nom.

d) Des erreurs dans le texte qui génèrent des erreurs d'analyse...

De manière plus générale, à cause de la structure en couche des vérificateurs de deuxième génération (voir figure 2.1 p. 28), tout élément du texte qui est erroné ou inconnu peut avoir des conséquences sur l'ensemble de la chaîne d'analyse et fausser la détection d'erreurs. En effet,

chaque étape du traitement transmet ses résultats à l'étape suivante, mais si elle commet une erreur ou ne sait pas traiter un élément, ceci se répercute sur toutes les étapes suivantes et il n'est pas possible de revenir en arrière. Ainsi, un élément inattendu (mot inconnu, erroné, omis ou ajouté) dans le texte peut être à l'origine d'une erreur d'étiquetage, ou de désambiguïsation, qui empêchera à son tour l'application des règles qui auraient permis de détecter une erreur impliquant cet élément inattendu.

Par exemple, puisque l'attribution des étiquettes morphosyntaxiques est effectuée en fonction de la graphie des mots, si l'un d'eux est erroné (mauvais accord, confusion avec un homophone, etc.), le *tag* qui lui est affecté contient les informations correspondant à la graphie erronée, et non pas à celle attendue en fonction du contexte. Dans l'extrait (4) ci-après, le mot `calcule`, écrit par erreur d'homophonie à la place du substantif `calcul`, est sans ambiguïté étiqueté `verbe conjugué`, alors que dans le contexte de la phrase c'est une étiquette `nom masculin` qui serait attendue.

(4) *`Ce calcule se base sur un calcul mathématique[...]`*

Cette confusion conduit par exemple le vérificateur LanguageTool à signaler à tort une erreur sur `Ce`, qu'il propose de remplacer par `Se`.

Un *tag* inadapté de ce type altère ensuite la désambiguïsation des mots de son environnement immédiat, et indirectement de la totalité de la phrase. En effet, les règles de désambiguïsation, qu'elles soient fondées sur une méthode statistique ou qu'elles soient manuelles (*cf.* § b) p. 31), s'appuient toujours sur le contexte immédiat d'un mot pour déterminer l'étiquette adéquate parmi l'ensemble possible. Si ce contexte contient une erreur, il ne permet pas d'effectuer une désambiguïsation correcte. Par ricochet, les erreurs d'étiquetage qui en découlent nuisent à l'application des règles de détection d'erreurs, qui peuvent alors générer du silence ou du bruit.

e) ...d'où un cercle vicieux de la vérification grammaticale

Finalement, la rigidité du principe du *pattern-matching*, en requérant une correspondance parfaite entre les suites de mots du texte et celles des motifs des règles, limite la détection aux seules erreurs attendues. Toutes celles qui ne le sont pas perturbent le bon déroulement de cette détection. La vérification grammaticale se heurte ainsi à un cercle vicieux : les erreurs dans les textes sont à l'origine d'erreurs d'analyses morphosyntaxiques, qui sont à leur tour à l'origine d'une mauvaise détection des erreurs du texte. Autrement dit, de manière tout à fait contradictoire, il est nécessaire d'avoir un texte sans erreur pour que les règles de détection d'erreurs fonctionnent de manière optimale.

3.1.3 Les utilisateurs livrés à eux-mêmes

Analyser le texte et attirer l'attention de l'utilisateur sur les incohérences grammaticales détectées est une première étape dans la correction. Nous venons de voir que cette étape est sujette à de nombreuses difficultés, principalement dans les outils de seconde génération, dont font partie les outils libres. La détection d'erreur est ainsi loin d'être toujours fiable. Pour que l'utilisateur puisse alors s'assurer qu'il n'est pas en présence d'une fausse alarme issue d'une mauvaise analyse, puis comprendre pourquoi l'ordinateur a signalé une erreur, et enfin comment effectuer les modifications éventuellement nécessaires, il faut qu'il ait à disposition suffisamment d'informations. Il n'est cependant pas toujours aidé comme il le faudrait dans cette tâche. Les

rétroactions sur les incohérences détectées sont en effet très inégales selon les outils. Nous en donnons ci-après des exemples issus de différents vérificateurs, à partir de deux extraits du corpus d'erreurs que nous avons constitué (*cf.* chapitres 4 et 5 p. 63 à 87) :

(5) **Une formation effectué par l'éducateur canin de la ville.*

(6) **De nombreux chercheurs se sont montrés septiques quand à la capacité de Washoe à communiquer avec des humains.*

a) Exemples de logiciels commerciaux

Dans la famille des correcticiels commerciaux, Antidote propose des rétroactions de différents niveaux de détail. Un clic sur un mot détecté comme erroné donne accès à une première rétroaction très succincte (*cf.* figure 3.1, fenêtre 1), qui indique simplement la correction et un mot-clé d'explication. Dans l'énoncé (5) en figure 3.1, l'erreur (soulignée en rouge) porte sur ***effectué*** qui doit être écrit au féminin. En cliquant sur le mot « Féminin », nous accédons à un deuxième niveau de rétroaction, avec des précisions supplémentaires intégrant les éléments concernés par l'incohérence et une brève explication (fenêtre 2). Le logiciel indique pour notre énoncé (5) que le mot ***effectué*** doit être au féminin pour avoir le même genre et le même nombre que ***formation***, mais il n'explique pas pourquoi. Pour avoir cette explication, il faut cliquer sur la petite icône livre qui amène à un troisième niveau de rétroaction (fenêtre 3). Elle ouvre le « guide » contenant les règles de français, directement sur la règle transgressée par l'erreur détectée. La règle est détaillée et illustrée de plusieurs exemples.

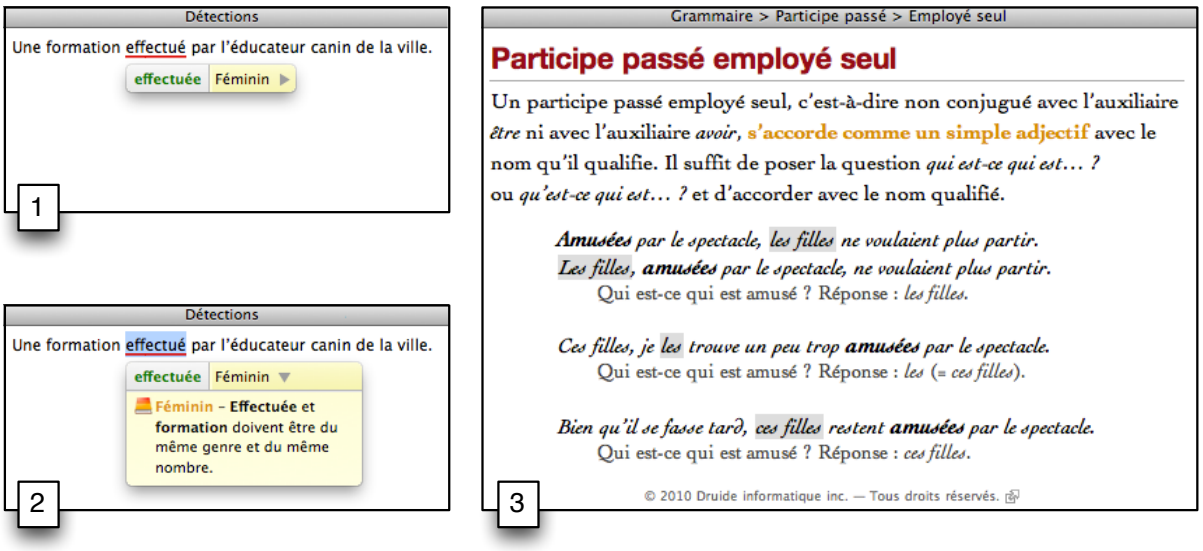


FIGURE 3.1 : Exemples de rétroactions du logiciel Antidote (énoncé (5))

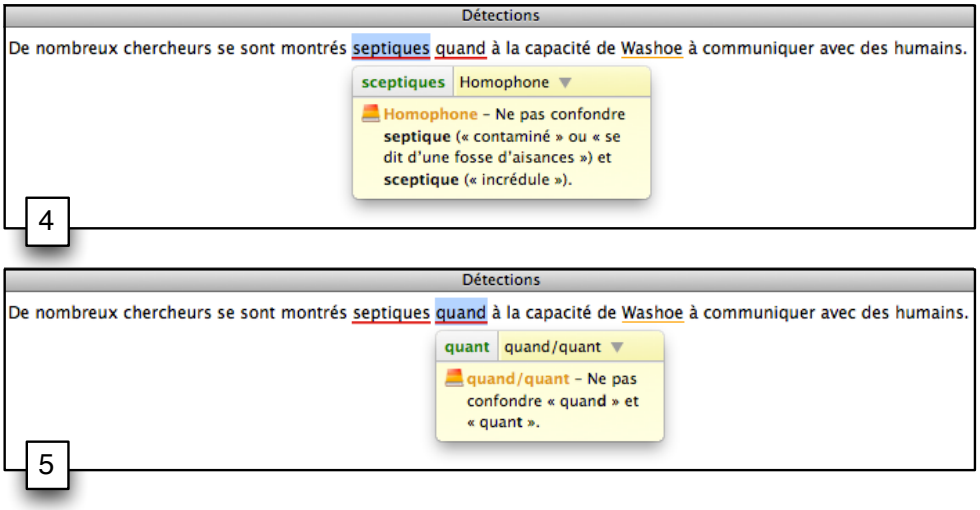


FIGURE 3.2 : Exemples de rétroactions du logiciel Antidote (énoncé (6))

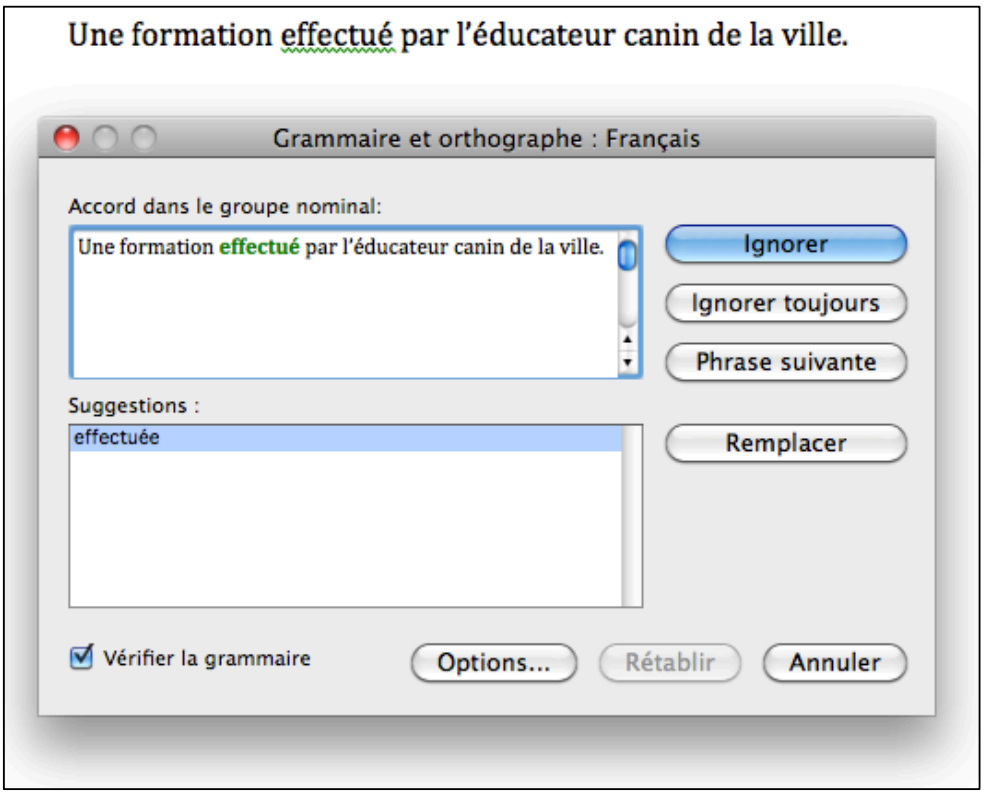


FIGURE 3.3 : Exemple de rétroaction de Word (énoncé (5))

Dans le second exemple, avec l'énoncé (6) qui contient deux confusions d'homophones, les rétroactions ne se valent pas. Nous voyons dans la fenêtre 4 de la figure 3.2 que les deux mots confondus, `septiques` et `sceptiques`, sont tous deux définis brièvement de manière à ce que l'utilisateur sache laquelle des deux graphies correspond à celle qui convient dans son énoncé.

Curieusement, il n'en est pas de même pour la confusion entre `quant` et `quand` dans la fenêtre 5. Le logiciel dit juste de ne pas les confondre, mais ne donne aucune indication sur le contexte d'utilisation qui convient à chacun. Pour accéder à cette information, il faut aller consulter le « guide » des règles de grammaire.

Dans Word [3], le vérificateur de grammaire ne donne qu'une rétroaction très sommaire sur les erreurs qu'il détecte (soulignées en vert). Dans notre exemple en figure 3.3, sur l'énoncé (5), il indique simplement le type d'erreur, à savoir une erreur d'« Accord dans le groupe nominal » et suggère une correction. Il ne précise ni quels sont les mots à accorder, ni selon quel critère (genre et/ou nombre). Il ne propose pas non plus d'explication plus détaillée sur les accords, ni même d'exemple. Concernant l'énoncé (6), aucune erreur n'est détectée.

b) Exemple d'un logiciel gratuit

Il existe également des outils gratuits, mais dont les ressources ne sont pas libres et donc pas accessibles. Parmi eux, BonPatron [4] [Nadasdi & Sinclair, 2001], disponible uniquement en ligne, propose des rétroactions qui indiquent comment corriger l'erreur, avec un exemple en illustration. Ces rétroactions sont à visée didactique puisque l'outil est principalement destiné à des utilisateurs non francophones natifs, mais beaucoup ne sont cependant pas suffisamment claires à notre avis. Par exemple, si nous reprenons l'énoncé erroné (5), la rétroaction (figure 3.4) peut être difficile à interpréter. Tout d'abord, elle n'est pas contextualisée et ne reprend donc pas les termes impliqués dans l'erreur. Ensuite, dans la rétroaction sur notre erreur précisément, l'exception citée (« si vous utilisez un participe présent *-ant* ») associée à l'adjectif donné en exemple (« *importante* ») peut prêter à confusion, surtout pour un non francophone. En effet, au masculin, cet adjectif est identique au participe présent du verbe `importer`.

Le second exemple en figure 3.5, avec l'énoncé (6), montre une rétroaction moins ambiguë sur la confusion des deux homophones `quand` et `quant`, avec une explication très claire sur la façon de choisir le bon terme. Notons dans ce second exemple qu'une erreur est suspectée à tort sur la fin de la phrase, à cause sans doute de la présence d'un nom propre inconnu, et qu'*a contrario* la confusion sur `septique` n'est pas détectée. Notons également que les erreurs dans nos deux exemples sont surlignées en jaune, ce qui signifie que le programme n'est pas certain qu'il y ait une erreur et qu'il invite à vérifier la séquence en question. Lorsqu'il n'a aucun doute, l'erreur est surlignée en rouge et il est demandé de corriger, et non plus simplement de vérifier.

3. Microsoft®Word, Microsoft Corporation, 2007
4. BonPatron existe également en version « pro » accessible via un abonnement payant.

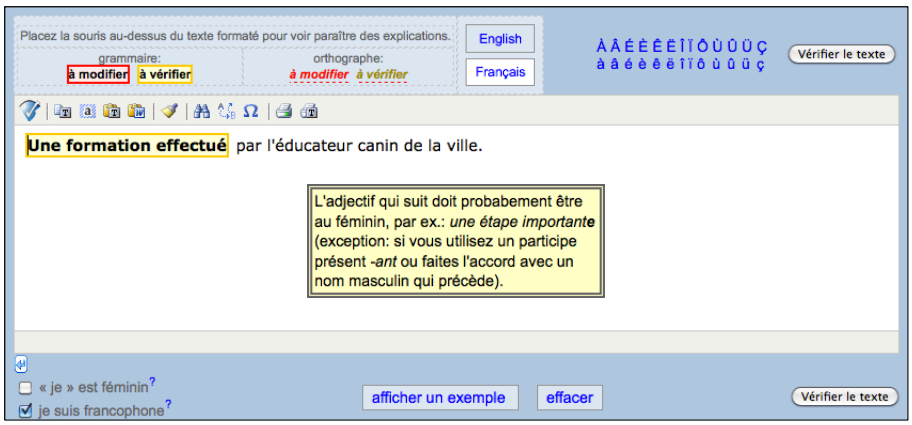


FIGURE 3.4 : Exemple de rétroaction de BonPatron (énoncé (5))

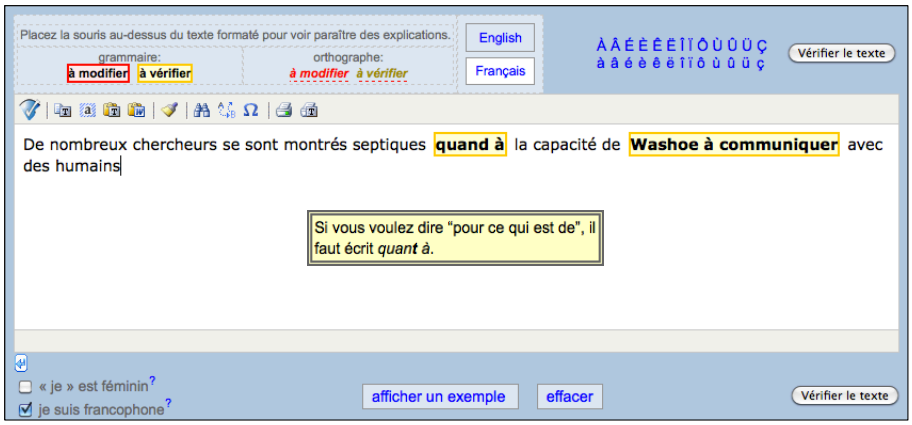


FIGURE 3.5 : Exemple de rétroaction de BonPatron (énoncé (6))

c) Exemples de logiciels libres

Du côté des logiciels libres, il existe principalement à l'heure actuelle deux outils qui sont capables d'effectuer quelques détections d'erreurs sur le français : LanguageTool de Naber [2003a] et Grammalecte de Ronez [2011]. Le premier a été initialement créé pour l'anglais, puis adapté au français [Souque, 2007]. Le second est dérivé de Lightproof, vérificateur grammatical conçu pour le hongrois par László [2009]. Tous deux constituent des extensions pour OpenOffice.org, mais LanguageTool existe également en version autonome.

Les rétroactions de LanguageTool reprennent, comme Antidote, les éléments incriminés dans les incohérences détectées, accompagnés de l'explication sur le type d'erreur et parfois d'exemples, mais il n'y a pas d'accès à des règles de grammaire. Sur la figure 3.6, nous voyons la vérification en cours sur l'énoncé (5), dans la version de LanguageTool intégrée au traitement de texte Writer d'OpenOffice.org. Le segment de texte qui déclenche une alerte est souligné en bleu dans la page, et repris en bleu dans la fenêtre de vérification. Le bouton « Expliquer... » permet d'accéder à la rétroaction. Dans notre exemple, LanguageTool indique qu'il y a un problème d'accord en genre entre `formation` et `effectué`, mais ne précise pas s'il faudrait un masculin ou un féminin, et ne propose pas non plus d'exemples d'énoncés corrects ou incorrects. En fait, ces derniers sont

disponibles dans les règles, mais ils ne sont pas affichés à l'utilisateur. Nous donnons ci-dessous une version simplifiée de la règle appliquée par l'outil avec notre énoncé précisément, afin de comprendre comment sont construites les rétroactions de LanguageTool :

```
<rule name="nom féminin suivi du masculin">
   <pattern>
      <token postag="N f .*" postag_regexp="yes" skip="1"/>
      <token postag="J m .*" postag_regexp="yes"/>
   </pattern>
   <message>« \1 » et « \2 » ne semblent pas bien accordés en genre</message>
   <example type="incorrect">Une <marker>forêt tropical</marker>.</example>
   <example type="correct">Une <marker>forêt tropicale</marker>.</example>
</rule>
```

Les règles de LanguageTool sont formalisées dans le langage de représentation XML. Une règle est délimitée par la balise `<rule></rule>`. Elle contient un `pattern` qui décrit le modèle de l'erreur, un `message` personnalisé affiché à l'utilisateur, et des exemples (`example`) d'énoncés corrects ou incorrects sur le point grammatical traité par la règle. Dans notre exemple, le `pattern` est constitué de deux `tokens` : un nom féminin (`N f .`) suivi d'un adjectif masculin (`J m .`). Si ce patron est reconnu dans le texte, le `message` personnalisé s'affiche. Il reprend en arguments `\1` et `\2` le premier `token` (le nom féminin) et le second (l'adjectif masculin), ce qui donne avec notre exemple en figure 3.6 la rétroaction « « formation » et « effectué » ne semblent pas bien accordés en genre ».

Dans le second exemple (figure 3.7), nous avons utilisé la version autonome de LanguageTool pour vérifier l'énoncé (6). La partie supérieure de la fenêtre contient le texte, et la partie inférieure les rétroactions. Pour les erreurs de confusion d'homophones que contient l'énoncé, le logiciel met les mots erronés en rouge au sein de leur contexte d'apparition, et nous voyons également qu'une correction est proposée. Ce n'est le cas que pour quelques types d'erreurs, dont les erreurs d'homophonie comme ici, ou encore les anglicismes. En revanche, contrairement à BonPatron, ou encore à Grammalecte comme nous le verrons juste après, il n'explique pas comment différencier les deux homophones `quand` et `quant`. Il indique simplement de remplacer l'un par l'autre.

L'interface de vérification grammaticale de Grammalecte est très similaire à celle de LanguageTool intégrée à OpenOffice.org, car il s'agit également d'une extension de la suite bureautique. Au niveau des rétroactions, il ne donne pas d'exemple, ni de règle de grammaire, mais fait parfois des suggestions de correction, comme en figure 3.9. L'explication concernant l'incohérence détectée est en revanche un peu plus précise que dans LanguageTool. Avec notre énoncé (5) dans la figure 3.8, nous voyons que la rétroaction, en plus d'indiquer que l'erreur porte sur l'accord de genre, donne le genre respectif des deux mots impliqués dans l'accord. Dans l'énoncé (6) de la figure 3.9, il ne détecte que l'erreur sur `quant`, mais il propose à l'utilisateur une brève formulation synonyme de chacun des homophones `quand` et `quant` pour l'aider à sélectionner la forme adéquate.

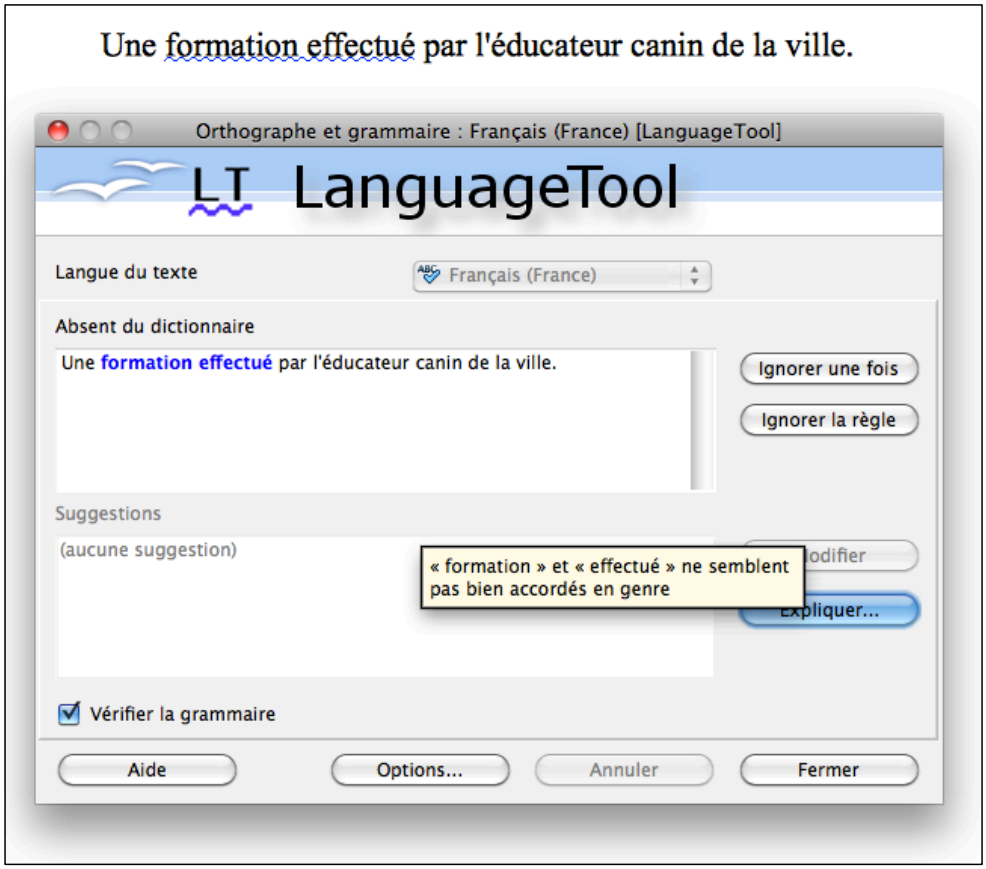


FIGURE 3.6 : Exemple de rétroaction de LanguageTool (énoncé (5))

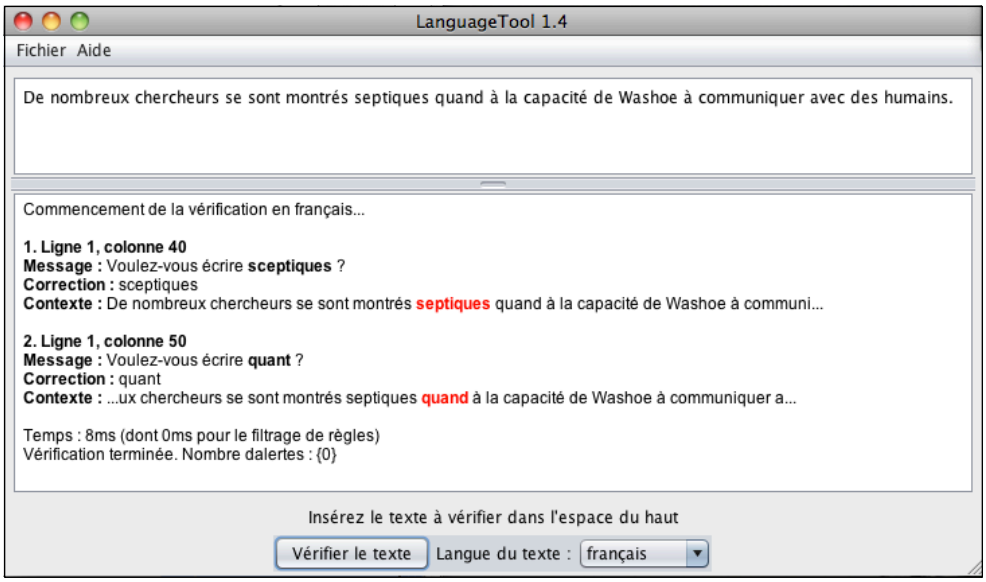


FIGURE 3.7 : Exemple de rétroaction de LanguageTool (énoncé (6))

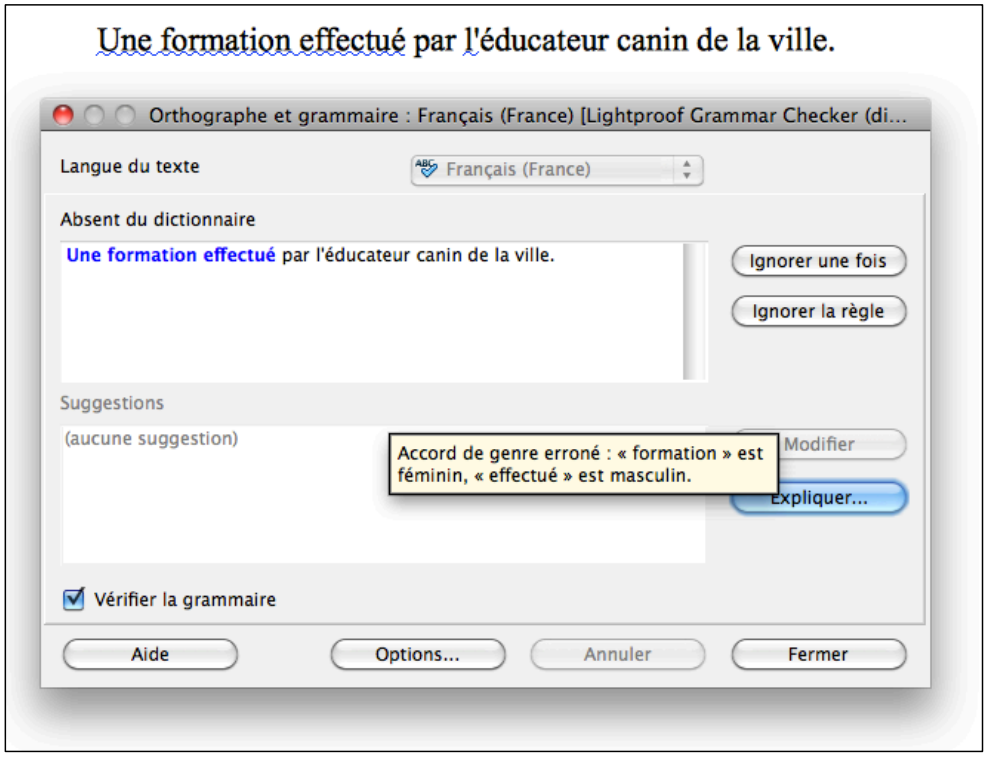


FIGURE 3.8 : Exemple de rétroaction de Grammalecte (énoncé (5))

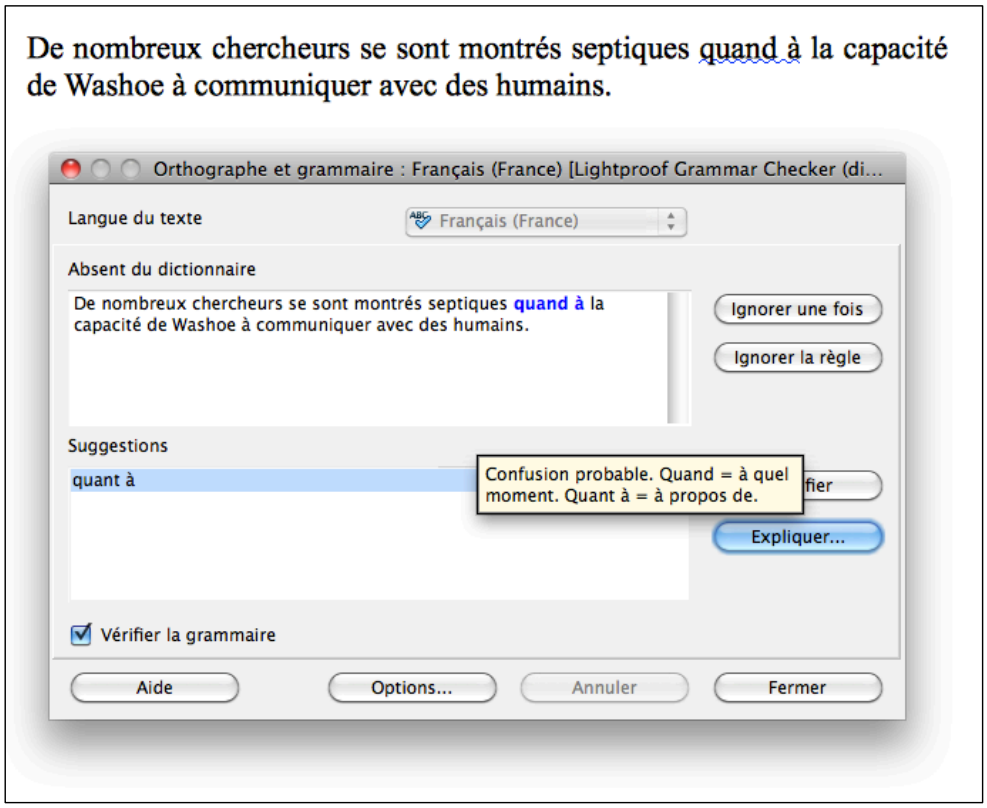


FIGURE 3.9 : Exemple de rétroaction de Grammalecte (énoncé (6))

d) Des rétroactions inégales

Ces quelques exemples nous montrent que les rétroactions sont très inégales selon les outils, et parfois au sein-même des outils. Elles devraient selon nous être claires et avoir une portée didactique, en expliquant à l'utilisateur l'origine probable de son erreur et comment la rectifier, pour l'accompagner au mieux dans la résolution des incohérences grammaticales détectées. Ce n'est pas le cas par exemple avec Word, dans la figure 3.3 p.47. Elle nous montre une rétroaction qui indique le type d'erreur, à savoir un problème sur l'accord dans le groupe nominal, mais il faut alors que l'utilisateur sache ce qu'est et comment identifier un groupe nominal pour comprendre et corriger son erreur, ce qui n'est pas nécessairement une évidence pour tous.

Les rétroactions sont aussi souvent prescriptives, en demandant à l'utilisateur, sans explication, d'effectuer telle ou telle modification. LanguageTool (*cf.* figure 3.7 p.51) suggère ainsi de remplacer `quand` par `quant`, sans la moindre justification. L'utilisateur n'a alors d'autre choix que de faire une confiance aveugle au vérificateur, dont nous avons fait remarqué qu'il est encore loin d'être infaillible.

Les outils de vérification grammaticale proposent trop souvent des rétroactions insuffisantes, laissant alors l'utilisateur livré à lui-même. Mais des outils trop verbeux ne sont pas non plus forcément souhaitables. Comme le fait remarquer Madec [2004, p. 98], « Quand on rédige un article ou un rapport, est-on prêt à recevoir une demi-heure de leçon de grammaire et d'explications diverses sur le fonctionnement des accords des verbes pronominaux parce qu'on vient de se tromper sur un cas un peu subtil d'accord de participe passé ? ». Le système de rétroactions à trois niveaux d'Antidote (*cf.* figure 3.7 p.51) est peut-être ainsi un bon compromis. Le premier niveau, qui ne donne que la correction et l'information clé sur l'incohérence relevée, est sans doute suffisant dans la plupart des situations pour les plus pressés ou aguerris. Le simple soulignage d'une erreur potentielle permet généralement à une personne possédant de bonnes compétences en grammaire, ou à un utilisateur ayant conscience des possibilités et des faiblesses de son logiciel en matière de correction, de décider s'il faut corriger ou pas. Le second niveau peut venir renforcer l'explication si nécessaire, et enfin le troisième niveau, avec le guide des règles grammaticales, apporte l'explication approfondie dont peut avoir besoin l'utilisateur peu à l'aise avec certains phénomènes linguistiques.

De manière plus générale, une étude des erreurs, de leur fréquence et du profil d'utilisateur plus enclin à les produire, permettrait sans doute de réaliser des rétroactions plus adaptées à chaque type d'erreur.

3.2 Panorama des études sur les erreurs tapuscrites

Nous avons jusqu'ici beaucoup parlé des vérificateurs grammaticaux et de leur fonctionnement, mais nous n'avons que brièvement abordé la question de ce qui constitue tout de même la raison d'exister et le grain à moudre de ces outils : les erreurs grammaticales, et plus précisément les erreurs tapuscrites. Écrire avec un crayon dans un cahier, ou avec un clavier sur un ordinateur, cela ne change rien quant à la norme grammaticale à respecter. Mais les erreurs que nous commettons sur papier sont-elles les mêmes que celles commises sur clavier ? Les études existantes nous renseignent peu sur cette question, mais quelques erreurs spécifiquement tapuscrites peuvent tout de même être dégagées.

3.2.1 Les études existantes

C'est sans conteste dans le domaine de la didactique des langues que les études sur l'erreur sont les plus nombreuses, mais il en existe également qui se focalisent sur les non-apprenants, sur des catégories particulières de scripteurs ou encore sur des types spécifiques d'écrits.Cependant, rares sont celles qui s'appuient exclusivement sur des textes tapuscrits.

a) Études des erreurs d'apprenants

Les travaux sur l'analyse des erreurs foisonnent en didactique des langues, surtout depuis les années 70.

> « D'une part, [l'analyse des erreurs] sert à décrire, expliquer et corriger les erreurs (orientation didactique) ; d'autre part, elle aide à mieux comprendre les processus et les stratégies d'apprentissage des langues étrangères (orientation psycholinguistique). »
>
> [Porquier, 1977, p. 23]

De nombreuses typologies d'erreurs ont vu le jour. Celle de Debyser *et al.* [1967] « a certainement comblé dans une certaine mesure un manque d'outil d'analyse qui se faisait ressentir auprès des enseignants à cette période-là, face aux besoins didactiques du moment » [Luste-Chaa, 2009, p. 143]. À partir de l'analyse d'un corpus d'écrits d'élèves africains francophones de 1er cycle, et dans le but de concevoir un manuel de rattrapage en français pour les élèves d'Afrique en classe de 6e, les chercheurs du BELC (Bureau pour l'Enseignement de la Langue et de la Civilisation françaises à l'étranger) ont élaboré la « grille de classement typologique des fautes ». Ils établissent notamment deux distinctions principales : la première entre « faute relative » et « faute absolue », et la seconde entre « faute graphique » et « faute orale ». Une faute est absolue quand elle conduit à une forme inexistante ; elle est relative quand elle créé une forme existante mais incorrecte dans son contexte. Nous remarquons ici la similitude avec la conception de l'orthographe et de la grammaire dans les outils de vérification de textes. Les erreurs traitées par les vérificateurs orthographiques s'apparentent aux erreurs absolues, puisqu'il s'agit de détecter les formes inexistantes dans le lexique, et celles prises en charge par les vérificateurs grammaticaux sont au contraire des erreurs relatives. La seconde distinction établie par Debyser *et al.* [1967], entre « faute graphique » et « faute orale », fait intervenir la prononciation. Une forme erronée à l'écrit et dont l'oralisation serait également altérée, est considérée comme une erreur orale (par ex. *`distination` pour `destination`). Si la prononciation de la forme écrite erronée n'est pas touchée, alors l'erreur est graphique (par ex. *`riveaux` pour `rivaux`).

La grille du BELC a été conçue de manière à couvrir toutes les erreurs possibles, et à pouvoir être utilisée auprès de n'importe quelle population d'élèves. Mais si elle constitue une avancée notable dans l'analyse des erreurs et est reconnue comme telle, elle n'en a pas moins été aussi vivement critiquée, par Porquier [1977] notamment. Il lui est reproché entre autre la distinction entre fautes absolues et relatives qui, si elle est pertinente au niveau du mot, est difficilement applicable à niveau syntagmatique. Les critiques ont cependant été constructives et ont ouvert la voie à d'autres travaux.

La typologie de Catach *et al.* [1980] a également rencontré un succès important dans le milieu scolaire. Elle permet un relevé et une description fine linguistiquement des erreurs d'orthographe, en suivant les mêmes ordres que l'analyse du langage : phonique, grammatical, lexical et sémantique. La classification des erreurs se veut exhaustive et applicable à différents types

d'enseignements et d'écrits du milieu scolaire [Luste-Chaa, 2009]. Ses concepteurs distinguent les erreurs [5] à dominante :

- phonétique, pouvant être causées par une mauvaise audition ou prononciation des sons (*`maitenant`),
- phonogrammique, lorsqu'un son connu est mal transcrit (*`pingoin`),
- morphogrammique, lorsque la graphie des éléments non phonétiques (non marqués à l'oral) n'est pas respectée (*`Il crit`),
- logogrammique, lorsque des homophones sont confondus (`chant-champ`),
- idéogrammique, pour ce qui touche aux ponctuations, majuscules, etc.,
- non fonctionnelle, comme les lettres étymologiques (*`sculter`), les finales particulières (*`abrit`), les consonnes simples ou doubles difficilement justifiables (*`combatif`), etc.

Notons que ce classement ne tient pas compte des erreurs syntaxiques, comme l'ordre des mots, ou encore des erreurs lexicales, comme l'utilisation d'un mot pour un autre. Comme son nom l'indique, il se restreint à l'orthographe.

Cette typologie constitue un outil précieux pour aider les enseignants à pointer et comprendre les erreurs de leurs élèves et mettre en place une progression pédagogique. Elle a toutefois ses limites. En visant une couverture exhaustive des erreurs orthographiques avec un niveau poussé de détail dans leur description, son utilisation est rendue complexe et il n'est pas rare que des erreurs appartiennent simultanément à plusieurs cases différentes de la grille de classement. Enfin, elle se consacre exclusivement aux erreurs d'orthographe et ne permet donc pas l'analyse ou l'évaluation complète d'un écrit.

Nous n'avons présenté ici que les travaux de Debyser *et al.* [1967] et ceux de Catach *et al.* [1980], qui s'intéressent tous deux aux erreurs d'élèves francophones et ont connu un important succès, mais il en existe beaucoup d'autres (citons par exemple les travaux de François [1974], ou encore de Ducard *et al.* [1995]). Les typologies et autres analyses d'erreurs d'apprenants du français sont nombreuses, mais beaucoup d'autres études se focalisent sur les erreurs d'apprenants du Français Langue Étrangère (FLE), telle par exemple celle réalisée dans le cadre du projet européen FreeText [Granger *et al.*, 2001 ; Granger, 2007], dans le but de concevoir un programme d'Apprentissage des Langues Assisté par Ordinateur (ALAO) pour le FLE. Les chercheurs ont notamment constitué et annoté un corpus d'erreurs (FRIDA, *French Interlanguage Database*) à partir d'écrits rédigés par des apprenants du français.

Dans un contexte tout autre, Rabadi & Odeh [2010] ont réalisé une étude des erreurs récurrentes commises par des apprenants arabophones (jordaniens et bahreïniens) en FLE, à partir d'un corpus de copies d'examens, dans le but d'améliorer le développement du français au Moyen-Orient. Ils s'intéressent notamment aux diverses interférences (morphosyntaxiques, sémantiques, phonétiques, etc.) entre la langue maternelle et le français, qui créent un système intermédiaire entre les deux langues, appelé « interlangue » [Selinker, 1972]. C'est un phénomène particulièrement étudié par les chercheurs en didactique des langues.

Nous ne nous attardons pas davantage sur ces travaux, aussi intéressants qu'ils soient, car s'ils constituent bien des études des erreurs, ils s'appuient presque exclusivement sur des documents manuscrits. Par ailleurs, ils se focalisent sur les écrits d'apprenants, qui constituent un type trop spécifique pour l'analyse des erreurs commises par un scripteur *lambda*, même s'il n'est cependant pas souhaitable de les éliminer totalement, puisqu'ils font partie des utilisateurs potentiels de vérificateurs grammaticaux et des bénéficiaires des rétroactions.

5. Les exemples d'erreurs cités dans ce paragraphe sont extraits de Catach *et al.* [1980, p.13-15]

b) Études des erreurs d'adultes francophones

Si les erreurs d'apprenants ont fait l'objet de nombreuses études, les erreurs de francophones adultes sont bien moins connues. Parmi les études les concernant, nous pouvons évoquer « La grammaire des fautes » de Frei [1928], un des tous premiers ouvrages analysant les erreurs du français, dans lequel l'auteur a étudié des lettres manuscrites (adressées à des prisonniers de guerre) afin d'y « rechercher en quoi les fautes sont conditionnées par le fonctionnement du langage et comment elles le reflètent »[Frei, 1928, p. 9]. Il s'agit-là bien sûr d'un travail reposant exclusivement sur des documents manuscrits.

Plus récemment, et dans un cadre plus sociolinguistique, Lucci & Millet [1994] ont étudié les erreurs d'orthographe commises par les scripteurs francophones adultes (au-delà des classes de 2nde et 1re) afin d'obtenir « une photographie objective de l'orthographe dans ses usages réels ». Ils ont à cette fin rassemblé un corpus de documents variés : des manuscrits (lettres de demande d'emploi, lettres privées, cahiers de liaison, etc.) de scripteurs ordinaires et de futurs professionnels de l'écrit (futurs professeurs ou secrétaires) d'une part, et des journaux d'autre part. L'élaboration d'une grille d'analyse, dont la nomenclature n'est pas sans rappeler celle utilisée par Catach *et al.* [1980], a permis aux auteurs d'essayer de comprendre et d'expliquer les « variations orthographiques » (terme employé par les auteurs pour rendre compte à la fois des formes erronées et des formes variées conséquentes aux réformes de l'orthographe) relevées dans le corpus. Les variations sont réparties en trois grandes catégories : phonogrammes, idéogrammes (divisés en morphogrammes et logogrammes) et mutogrammes, et peuvent être le résultat d'une omission, d'une adjonction, d'une substitution ou encore d'une sélection (choix entre deux formes tolérées pour un même mot). Une seconde grille, très similaire à la première, a été conçue spécifiquement pour les signes diacritiques et auxiliaires d'écriture (traits d'union, blanc).

Le travail de Lucci & Millet [1994], tout comme celui de Catach *et al.* [1980], ne considère que l'orthographe, et ne permet donc pas le relevé et l'étude des erreurs de syntaxe par exemple. Par ailleurs, comme pour tous les autres travaux mentionnés précédemment, il s'appuie majoritairement sur des documents manuscrits et ne permet donc pas l'analyse des erreurs tapuscrites. Nous avons trouvé quelques études qui sont spécifiques aux erreurs commises sur clavier, mais pas nécessairement par des scripteurs « ordinaires ».

Jacquet-Pfau [2001], par exemple, propose une typologie qui traite exclusivement des erreurs tapuscrites détectées par les correcteurs, fondée sur les classes d'erreurs de Sabah [1989], elles mêmes issues de la dichotomie compétence/performance (*cf.* § 1.2.2 p. 21) de Chomsky [1965]. La typologie ainsi élaborée distingue les erreurs de performance du scripteur (inattention, saisie, erreur intentionnelle comme les néologismes par exemple), les erreurs de compétence du scripteur (orthographe, flexion, segmentation) et les erreurs du système (lexique lacunaire : néologismes, variantes orthographiques, mots étrangers, mots spécialisés, etc.). Cependant, cette typologie ne semble pas s'appuyer sur l'analyse d'un corpus d'erreurs effectives, et ne constitue donc pas une étude des erreurs tapuscrites à proprement parler.

Dans une autre étude, menée en neuropsycholinguistique par Bouraoui *et al.* [2009], l'attention est portée spécifiquement sur les erreurs tapuscrites, mais commises par des scripteurs infirmes moteurs cérébraux. À partir de l'analyse de leurs écrits et de la description précise des incohérences orthographiques et grammaticales relevées, les auteurs ont constitué un modèle général de ces erreurs. Leur étude a notamment pour but d'améliorer la détection par les vérificateurs automatiques des erreurs commises par ces scripteurs particuliers, erreurs dont ils ont montré qu'elles sont parfois spécifiques au handicap, mais également parfois à l'utilisation d'un clavier.

Ce dernier point nous intéresse particulièrement étant donné que nous souhaitons connaître les éventuelles erreurs d'écriture propres à la saisie sur clavier. Il est cependant difficile de prendre en compte les résultats des travaux de Bouraoui *et al.* [2009], dont la portée est restreinte à une catégorie très particulière de scripteurs handicapés, et de les généraliser aux autres scripteurs.

Ce sont finalement des travaux publiés très récemment qui s'approchent le plus du type d'étude que nous recherchons. Wisniewski *et al.* [2010] ont constitué un large corpus d'erreurs, extrait automatiquement à partir des historiques d'édition de l'encyclopédie collective Wikipédia. Les auteurs ont réuni un peu plus de 146 000 erreurs grammaticales et orthographiques avec leurs corrections, en sélectionnant de manière automatique, parmi toutes les modifications de textes, celles répondant à des critères donnés (par ex. modifications sur un seul mot, ne comportant pas de ponctuation, pas de chiffre, pas plus d'une majuscule, etc.). En dépit de l'intérêt certain d'un corpus de cet envergure, nous pouvons nous demander si les auteurs de ces erreurs et de leurs corrections sont des scripteurs ordinaires, et s'ils ne sont pas, comme le précisent Wisniewski *et al.* [2010], « globalement plus éduqués, plus familiers des nouvelles technologies, etc. que l'ensemble des scripteurs du français ». Par ailleurs, pour limiter le bruit dans le corpus, un nombre important de filtres a été appliqué aux modifications extraites. La diversité des erreurs finalement obtenue est alors nécessairement réduite.

Il est difficile de trouver des études générales sur les erreurs tapuscrites en français. Cette lacune trouve peut-être une explication dans la popularisation toute récente d'Internet et des ordinateurs (66,6% des foyers français équipés en 2009 [INSEE, 2009]), avec jusque-là l'absence d'utilisation grand public des machines à écrire, généralement confinées dans les bureaux et inutilisées en didactique. Il paraît difficile à présent de tenir les documents tapuscrits à l'écart des études sur l'erreur, tant ils sont omniprésents dans notre quotidien. Il leur sera sans doute accordé une place grandissante dans les études à venir, par rapport aux documents manuscrits, bien que ces derniers restent pour l'heure prédominants dans la conception des correcticiels qui analysent nos textes tapuscrits.

3.2.2 Spécificité des tapuscrits

Bien que le phénomène soit encore peu étudié, l'utilisation d'un clavier pour la saisie d'un texte est de toute évidence source d'erreurs spécifiques que l'on ne rencontre pas dans les documents manuscrits et qui ont d'ailleurs leur appellation propre : « erreur typographique », « faute de frappe », « coquille », etc. La souplesse des touches, la vitesse de frappe et surtout la configuration spatiale du clavier sont autant de paramètres qui peuvent conduire à :

- la substitution d'un caractère par son voisin sur le clavier (**avaluer* pour *évaluer*) ;
- l'inversion de deux caractères (**venderdi* pour *vendredi*) ;
- l'ajout d'un caractère, en appuyant par erreur simultanément sur deux touches contiguës (**onbjectif* pour *objectif*, ou par répétition d'un caractère (**paarc* pour *parc*) ;
- l'omission d'un caractère (**rster* pour *rester*) ;
- la substitution d'un caractère par celui situé à la même place sur un clavier de configuration différente (par exemple le `A` et le `Q` sont inversés sur un clavier AZERTY et un clavier QWERTY ce qui peut conduire une personne changeant de clavier à saisir l'un à la place de l'autre).

Ce type d'erreurs particulier est classé par Sabah [1989] dans les erreurs de performance (dues à l'inattention), en reprenant la théorie de Chomsky [1965] (*cf.* § 1.2.2 p. 21), telle que nous

l'évoquons dans le paragraphe 3.2.1 b) p.56, par opposition aux erreurs de compétence dues à une méconnaissance de la langue :

> « Bien que les causes soient souvent indiscernables, on peut dire que les fautes typographiques, par exemple dues à une mauvaise frappe sur le clavier, sont des erreurs de performance, alors que les erreurs phono-graphiques (comme écrire *ipotainuse* pour *hypoténuse*) sont généralement des erreurs de compétence. »
>
> [Sabah, 1989, p. 155]

Cette classification en deux catégories est un peu floue chez l'auteur, qui reconnaît en outre qu'il n'est pas toujours facile de distinguer les erreurs de performance et les erreurs de compétence du scripteur. Sabah [1989, p. 155] indique par exemple que « les erreurs de performance de l'utilisateur correspondent à deux types de phénomènes : les fautes d'attention et les modifications intentionnelles de la langue (création de néologismes, par exemple). » Cependant, un peu plus loin, il classe les néologismes dans les erreurs de compétence, de même qu'un exemple dont nous pensons qu'il devrait plutôt relever de la performance, si nous nous en tenons à ses propos. Il s'agit de « `médoukipudonctan` » [Queneau, 1959], donné en illustration de la suppression de blancs dans les erreurs de segmentation, et qui est visiblement une modification intentionnelle.

La typologie de Jacquet-Pfau [2001], comme la classification proposée par Sabah [1989] sur laquelle elle s'appuie, est parfois ambigüe. Elle donne, comme son prédécesseur, les erreurs de flexion et les erreurs de segmentation comme erreurs de compétence, en plus des erreurs phonographiques (substitution de graphèmes de même prononciation). A propos des erreurs de segmentation, l'auteur les décrit comme « résultant très souvent d'une erreur de frappe » [Jacquet-Pfau, 2001, p.87], ce qui vient en contradiction avec leur classement dans les erreurs de compétence.

À propos des erreurs de flexion, et des erreurs d'accord plus particulièrement, Sabah [1989, p. 161] signale « qu'il peut également s'agir d'erreurs de performance pour les locuteurs maîtrisant le français ». Ces erreurs existent dans l'écriture manuscrite, mais elles peuvent aussi être causées par la manipulation du texte rédigé. En effet, la rédaction sur ordinateur, avec les fonctions d'édition (notamment le « copier-coller ») qui facilitent l'ajout, la suppression, le remplacement ou le déplacement de fragments du texte, conduit parfois à des erreurs morphosyntaxiques. Elles ne sont alors pas dues à une méconnaissance de la langue, mais à une construction initialement correcte rendue fautive non intentionnellement par la modification d'un élément du texte. Nous en avons fait l'expérience en rédigeant ce chapitre, avec par exemple `Les travaux existants` qui est devenu, après reformulation, *`Les études existants`.

Il apparaît ainsi difficile d'effectuer un classement des erreurs des écrits dactylographiés en se fondant sur la dichotomie compétence/performance. Une des raisons se trouve sans doute dans les erreurs de frappe qui sont souvent délicates à identifier comme telles, et donc comme erreurs de performance. Comment savoir si l'absence d'un -s du pluriel est due à une méconnaissance du scripteur (compétence), à un problème d'attention (performance) ou encore à une pression trop faible sur la touche [S] pour que la lettre soit effectivement saisie (performance) ?

Notons que les classements d'erreurs proposés par Sabah [1989], et dans une moindre mesure par Jacquet-Pfau [2001], ne reflètent peut-être pas l'écriture tapuscrite telle qu'elle est pratiquée aujourd'hui. En effet, l'utilisation d'un clavier était alors beaucoup moins répandue. « Le développement de l'équipement des ménages en micro-ordinateur ne commence que vers la fin des années 1990, 15% des ménages en possèdent en 1995, ils sont 27% en 2000 et 62% au début de l'année 2008. » [Lacroix, 2009, p. 2]. Les classements présentés ont toutefois la particularité, par rapport aux classements d'erreurs manuscrites, de prendre en compte le système comme source

potentielle d'erreurs.

Plus récemment, Boissière et son équipe ont effectué une étude dans laquelle ils considèrent spécifiquement les erreurs dues à l'utilisation du clavier, et plus précisément dues à la proximité des touches sur le clavier. Mais ils ne classent dans cette catégorie que « les erreurs d'addition, ou de substitution entre une lettre cible et un périmètre situé immédiatement autour », et certaines erreurs sur les diacritiques (omissions et substitutions) [Boissière *et al.*, 2007, p. 7]. Même si l'étude est centrée sur des écrits de personnes handicapés, et si elle ne caractérise comme des erreurs de saisie qu'un type d'erreurs très précis, elle montre la réalité des erreurs proprement tapuscrites.

En revanche, il s'agit dans l'étude de Boissière *et al.* [2007] d'exemples purement orthographiques. Ce sont là des erreurs qui sont pour nous moins intéressantes à étudier que les erreurs d'accords par exemple, puisqu'elles ne relèvent généralement pas de la vérification grammaticale. Nous voyons cependant que les erreurs observables en contexte spécifiquement tapuscrit peuvent différer en partie de celles rencontrées dans les manuscrits, et leur étude est incontournable si nous voulons les détecter efficacement.

Deuxième partie

Caractérisation des erreurs tapuscrites

Chapitre 4

Choix d'une approche corpus

Sommaire

4.1 Justification d'une approche corpus 64
4.1.1 Définition de la notion de corpus 64
4.1.2 Les corpus disponibles 67
4.2 Méthodologie de constitution du corpus 68
4.2.1 Caractéristiques communes des données 68
4.2.2 Variété des scripteurs 69
4.2.3 Variété des situations de scription 70
4.2.4 Variété des types de documents 71
4.3 Caractérisation du corpus de l'étude 72
4.3.1 Écueils de la collecte des textes 72
4.3.2 Contenu du corpus et représentativité 73
4.3.3 Positionnement de notre corpus 75

> « Le choix d'un champ d'étude composé essentiellement de textes proprement littéraires [...] correspond aussi au désir, ressenti par de nombreux linguistes, d'échapper aux pièges d'une syntagmatique purement locale, génératrice de contre-exemples artificiels.
>
> Une telle approche nous semble aussi naturelle que celle du zoologiste préférant l'étude des espèces "réelles" à celle des animaux fabuleux ou mythologiques. »
>
> [Braffort, 1981, p. 120]

Pour être efficace, un vérificateur grammatical doit être capable de détecter un maximum des erreurs de grammaire que commettent les scripteurs. Il doit donc être conçu de manière à pouvoir reconnaître ces écarts d'écriture. Il est ainsi nécessaire, pour sa conception, de disposer d'une typologie et d'un modèle des erreurs possibles, afin de lui apporter les moyens de les détecter. Une analyse linguistique préalable des erreurs est donc indispensable. De plus, l'outil étant destiné à analyser des textes authentiques, contenant des erreurs non artificielles, son moteur de détection d'incohérences doit s'appuyer sur des données réelles. La nécessité de travailler sur des données attestées pour une analyse linguistique mène généralement à l'utilisation de corpus, solution que nous avons adoptée.

Ce chapitre est donc l'occasion d'une part, de présenter la notion de corpus et notre choix de nous tourner vers une telle approche, et d'autre part, d'apporter un regard critique sur le corpus que nous avons constitué, en abordant notamment les questions des besoins et de sa représentativité.

4.1 Justification d'une approche corpus

4.1.1 Définition de la notion de corpus

Les corpus permettent de rendre compte des fonctionnements de la langue en se fondant sur des « textes réels, c'est-à-dire produits pour des raisons de communication entre êtres humains et non [sur] des productions artificielles produites par l'introspection des linguistes » [Williams, 2005, p. 13]. La métaphore utilisée par Sinclair [1991, p. 6] illustre bien ces propos : « *One does not study all of botany by making artificial flowers.* [1] »

Les corpus permettent aussi d'effectuer des analyses statistiques sur les faits de langue observés, ce qui n'est pas le cas avec la linguistique introspective, mise en opposition à la linguistique de corpus par Jacques [2002], et qui s'appuie en effet sur des intuitions et non sur des faits attestés. Il est toutefois difficile de définir précisément la notion de corpus, dont les approches sont diverses et pour laquelle il n'existe pas de définition unifiée. Nous rendons compte de la diversité de ces approches dans les sections suivantes.

a) Linguistique(s) de/sur corpus

Les corpus ont pris leur essor en linguistique anglophone au début des années 1990, au sein du courant « *corpus linguistics* » qui, comme le note Péry-Woodley [1995, p. 214], « se trouve en quelque sorte officialisé [...] par une floraison d'articles visant explicitement à faire le point sur ce nouveau domaine, témoignant ainsi de sa vitalité sinon de sa maturité ». Ce courant « corpus » est aujourd'hui également très en vogue dans la linguistique francophone, mais il n'a en revanche pas de dénomination unifiée comme dans la linguistique anglophone. Pour Péry-Woodley [1995, p. 214], « le fait que n'existe pas en français un terme unificateur a pour conséquence que rien ne vient cacher la diversité des objectifs et des méthodes des différents utilisateurs de corpus ».

Ainsi, pour certains linguistes, les corpus appartiennent à « la linguistique de corpus », qu'ils considèrent comme une discipline en soi. Dans l'introduction de son ouvrage, Williams [2005] insiste sur ce terme de « discipline ». Habert *et al.* [1997] ne vont pas aussi loin et parlent « des linguistiques de corpus ». Ce pluriel « rend bien compte de la diversité des champs et des approches proposées » [Cori & David, 2008, p. 112]. Pour Mayaffre [2005], « cette pluralité trahit d'importantes différences dans les visées et les pratiques de linguistes venus d'horizons différents (phonologie, syntaxe, sémantique, etc.) ». Nous ne nous aventurons pas dans la présentation de cette diversité [2]. Nous préférons nous focaliser sur les approches et définitions qui correspondent au mieux à la manière dont nous envisageons notre corpus.

Par exemple, Mayaffre [2005] et Williams [2005] distinguent deux approches : déductives et inductives, appelées aussi respectivement *corpus-based* et *corpus-driven*. La première s'appuie

1. On n'étudie pas toute la botanique en fabriquant des fleurs artificielles.
2. Le lecteur intéressé pourra se référer notamment à Péry-Woodley [1995] au sujet de la diversité des approches, à Atkins *et al.* [1992] ou Sinclair [1996] pour une typologie des corpus.

sur les corpus, les utilise comme « support » pour valider ou invalider des hypothèses formulées *a priori*, et ne leur donne alors parfois « qu'une fonction validante, voire illustrative » [Mayaffre, 2005]. Cette approche est qualifiée par Mayaffre de « linguistique **sur** corpus », par opposition à « linguistique **de** corpus » qu'il emploie pour désigner la seconde approche, celle dans laquelle il se positionne. Celle-ci en effet utilise les corpus comme « apport » qui permet « de décrire puis d'élaborer des modèles *a posteriori* » [Mayaffre, 2005]. L'auteur ne nie pas que « des hypothèses de travail "théoriques" auront présidé à la constitution du corpus » dans cette seconde approche, mais ces hypothèses ne sont ensuite pas prises en compte dans la description interprétative du contenu.

Par ailleurs, si Mayaffre [2005] parle de « frontière » entre les deux approches, il mentionne également l'existence incontestable « d'un va-et-vient incessant entre la théorie et l'empirie, entre procédés déductifs et procédés inductifs ». Selon lui, ce va-et-vient implique nécessairement un point de départ dans l'une ou l'autre des deux approches, point de départ qui détermine la posture du linguiste. Dans notre cas, les objectifs attribués au corpus que nous avons constitué conduisent à un va-et-vient de la sorte. D'une part, il doit servir à la modélisation des erreurs morphosyntaxiques qu'il contient, revêtant ainsi un rôle d'apport et nous situant dans une approche inductive. D'autre part, il va nous servir de support pour étudier les erreurs, pour valider notre modèle, pour tester un futur prototype de vérificateur grammatical, ainsi que pour nous fournir des exemples d'erreurs, ce qui nous place cette fois dans l'approche déductive. Nous alternons ainsi entre les deux approches, mais notre « point de départ » se situe plutôt dans l'induction, notre objectif premier étant la modélisation des erreurs. Si nous suivons Mayaffre [2005], nous pouvons donc nous revendiquer de la linguistique de corpus. Cependant, l'usage majoritaire que nous ferons de notre corpus en tant que support nous incite à nous positionner préférablement dans la linguistique sur corpus.

b) Définition générale

De la diversité de ces approches découlent également différents types de corpus et donc des définitions variées de cette notion. Les points communs sont néanmoins fréquents entre les définitions. Pour Biber *et al.* [1998] par exemple,

> « *A corpus is not simply a collection of texts. Rather, a corpus seeks to represent a language or some part of a language. The appropriate design for a corpus therefore depends upon what it is meant to represent* [3] »
>
> Biber *et al.* [1998, p. 246]

Comme dans cette définition, les corpus sont souvent qualifiés de « collections de textes », mais avec généralement d'autres caractéristiques en complément, qui permettent alors de les distinguer de « simples » collections de textes. La définition de Biber *et al.* [1998] mentionne également la visée représentative d'une (partie d'une) langue qu'a le corpus, ce qui influe aussi sur sa conception. Celle-ci doit ainsi être effectuée en fonction de ce que les chercheurs souhaitent étudier, et donc de ce que le corpus doit représenter. Ainsi, selon Habert [1998, p. 35], « un corpus résulte d'un regroupement raisonné, conduit par une hypothèse de recherche explicite ».

Cette idée de « regroupement raisonné », initiée par Sinclair, est reprise dans les définitions de Williams [2005] ou de Sinclair [2005], qui évoquent également la notion de stockage électronique des textes :

3. Un corpus n'est pas simplement une collection de textes. Il cherche plutôt à représenter une langue ou une partie d'une langue. Ainsi, son contenu dépend de ce qu'il est censé représenter.

> « Les corpus dépassent le simple texte pour constituer des ensembles de textes choisis et ordonnés selon des critères précis. [...] La notion de corpus doit pouvoir englober [...] des grands corpus constitués, mais aussi des petits corpus d'études, des corpus littéraires, des corpus monosource, le point commun étant qu'il s'agit de textes stockés électroniquement sous un format texte. »
>
> [Williams, 2005, p. 13-14]

La définition de Sinclair [2005] précise en plus que les critères de choix et d'ordonnancement des textes sont « externes » :

> « *A* corpus *is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.* [4] »
>
> [Sinclair, 2005]

Les « critères externes » mentionnés dans cette définition sont des critères essentiellement non linguistiques (sociaux et contextuels) [Clear, 1992]. Ils sont variables selon les objectifs du corpus à constituer, mais concernent communément, pour Sinclair [2005], le media de production initiale du texte (oral, écrit, ou encore électronique), son type (livre, journal, lettre, etc.), son domaine (sciences, art, etc.), les langues ou variétés de langue, la localisation (français de France, du Québec, de Belgique, etc.) et la date de production. L'auteur insiste sur l'importance de ces critères extralinguistiques pour la constitution d'un corpus : « *Corpora should be designed and constructed exclusively on external criteria* [5] ».

Les corpus ne doivent donc pas être confondus avec de simples collections de textes, dans lesquelles les documents sont réunis simplement parce qu'ils sont faciles d'accès, sans qu'il y ait réellement de lien entre eux, au lieu d'être sélectionnés et regroupés selon des critères précis répondant à un objectif de recherche [Habert, 1998]. Ils doivent par ailleurs être constitués uniquement de données langagières authentiques, ce qui peut sembler suffisamment évident pour ne pas toujours être mentionné dans les définitions. Il nous paraît pourtant important de le préciser, comme le font Riegel *et al.* [1994, p. 18] : « Une telle collection ne comprenant que des données attestées (des énoncés effectivement produits) constitue un corpus ».

Afin de résumer les différents aspects que nous venons d'évoquer, et de montrer à la fin de ce chapitre (*cf.* § 4.3.3 p. 75) que les textes que nous avons recueillis constituent bien un corpus, nous proposons, en nous appuyant sur les définitions précédentes, de définir les corpus de la manière suivante :

> **Définition 4.1** ***Un corpus*** *est une collection électronique de textes authentiques, regroupés pour servir d'échantillon langagier pour une étude linguistique donnée, et sélectionnés dans ce but selon des critères définis spécifiquement.*

Une telle ressource présente des intérêts multiples pour nos travaux. Elle nous permettra de travailler sur des erreurs réelles pour les analyser, en faire leur typologie, élaborer un modèle pour leur détection et vérifier la validité de ce modèle. Par ailleurs, seul un corpus rend également possible la réalisation d'analyses statistiques, que ce soit sur les erreurs elles-mêmes ou sur les erreurs selon les contextes de scription et les individus. Notre modèle de détection d'incohérences grammaticales pourra s'appuyer en partie sur ces analyses. Nous pourrons privilégier

4. Un corpus est une collection de morceaux de textes sous forme électronique, sélectionnés selon des critères externes afin de représenter, dans la mesure du possible, une langue ou une variété de langue comme source de données pour la recherche linguistique.

5. Les corpus devraient être conçus et construits en fonction de critères externes exclusivement.

par exemple le traitement des erreurs les plus fréquentes, ou bien mettre un oeuvre un support probabiliste pour orienter la prise de certaines décisions par l'outil de vérification grammaticale en cas d'ambiguïté dans la détection des erreurs.

En revanche, malgré un intérêt certain, les corpus ont l'inconvénient d'être très coûteux à constituer. Nous avons donc essayé, en vain cependant, de trouver dans la littérature et les projets existants dans ce domaine un corpus que nous aurions pu réutiliser pour notre recherche

4.1.2 Les corpus disponibles

Les études précédemment menées sur les erreurs à l'écrit s'appuient très souvent sur des corpus, mais il n'existait pourtant pas, à l'heure où nous avions besoin, de corpus d'erreurs disponible adapté à nos besoins, c'est-à-dire constitué de divers documents tapuscrits électroniques rédigés par des scripteurs variés.

Le corpus FRIDA [Granger, 2007], par exemple, constitué dans le cadre d'un projet de développement d'un logiciel d'ELAO (Enseignement des Langues Assisté par Ordinateur), est relativement important avec plus de 450 000 mots. Cependant, pour plusieurs raisons, ce corpus ne correspond pas à nos besoins. Tout d'abord, il est constitué uniquement de productions d'apprenants du FLE. Même si cette catégorie de locuteurs/scripteurs du français n'est pas à ignorer dans notre corpus, ce dernier ne serait en aucun cas représentatif de la variété des dactylographes s'il ne contenait que des écrits de francophones non natifs. Ensuite, les textes constituant le corpus FRIDA sont initialement rédigés sur papier, puis numérisés, ce qui pose problème par rapport à notre objectif. Nous souhaitons en effet étudier non pas les erreurs commises dans des textes manuscrits, qui ont déjà fait l'objet d'études (*cf.* § *Les études existantes* p. 54), mais celles des écrits dactylographiés. Par ailleurs, les textes originaux manuscrits ont été retranscrits sur clavier. Nous pouvons donc émettre l'hypothèse que des erreurs supplémentaires ont été ajoutées par inadvertance, avec par exemple des fautes de frappe, ou bien que des erreurs authentiques ont été corrigées inconsciemment par les transcripteurs. Ceci introduit un biais dans l'authenticité des erreurs d'écriture commises. Nous avons donc préféré écarter ce corpus.

Pour des raisons similaires, nous avons également écarté le corpus de « variations orthographiques » COVAREC [1994]. Contrairement au corpus FRIDA, il est constitué de textes de francophones natifs, ce qui correspond mieux à nos besoins. Mais il s'agit exclusivement des documents manuscrits (copies d'étudiants, notes, lettres à un service administratif ou de demande d'emploi).

Quant aux corpus constitués par les industriels pour le développement de leurs outils, tel celui rassemblé pour le logiciel Cordial [Synapse Développement, 2006], à supposer qu'ils correspondent à nos besoins, ils ne sont bien sûr pas accessibles et nous n'avons donc pas pu les utiliser.

Le récent corpus d'erreurs orthographiques et grammaticales constitué par Wisniewski *et al.* [2010], que nous avons déjà évoqué dans le paragraphe *Études des erreurs d'adultes francophones* p. 57, semble mieux correspondre à notre problématique. Il est extrait d'un corpus plus large, WiCoPaCo [Max & Wisniewski, 2010], contenant les modifications dites « mineures » apportées aux articles de l'encyclopédie collaborative Wikipédia, et disponibles dans l'historique des éditions. Nous n'avons pas pu utiliser ce corpus, pour la simple raison que la publication de ces travaux est récente et que nous n'en avions donc pas connaissance lorsque nous avons entrepris la constitution de notre corpus. Cependant, même si nous en avions disposé plus tôt, nous l'aurions tout de même écarté, et ce pour diverses raisons. Tout d'abord, il n'est constitué que d'un seul type de

textes : les modifications des articles de Wikipédia. Ensuite, le corpus global WiCoPaCo contient énormément de modifications qui ne sont pas des erreurs d'orthographe ou de grammaire, et nécessiterait donc un important travail de « tri » afin que nous puissions l'exploiter. C'est ce qu'ont tenté de faire Wisniewski *et al.* [2010] avec leur sous-corpus d'erreurs orthographiques et grammaticales, mais celui-ci est pour le moment trop lacunaire pour répondre à nos besoins. En effet, seules ont été conservées les modifications de graphie (d'orthographe lexicale et grammaticale) et lorsqu'elles ne concernent pas plus d'un mot. Certains types d'erreurs grammaticales ne sont ainsi pas représentés dans le corpus, comme par exemple les erreurs de ponctuation, de fusion de mots, ou bien de manière plus problématique les erreurs se rapportant à la syntaxe (oubli d'une négation, répétition d'un mot, ordre des mots, etc.). Ces lacunes sont sans doute en grande partie dues à l'application de filtres successifs dans le but de limiter le bruit dans la première version du corpus, et seront probablement corrigées, au moins en partie, dans les versions suivantes.

Aucun corpus ne semble donc répondre à nos besoins, mais notre résolution de nous appuyer sur une telle ressource pour nos travaux, du fait de l'intérêt qu'elle présente, nous a conduite à en constituer un de toute pièce.

4.2 Méthodologie de constitution du corpus

Nous attribuons plusieurs objectifs à notre corpus, en fonction desquels nous avons défini les critères qu'il doit idéalement satisfaire :

- il doit, dans un premier temps, nous permettre de dresser une typologie des erreurs morphosyntaxiques présentes dans les écrits dactylographiés, et par la suite de modéliser ces erreurs ;
- il doit également nous permettre de réaliser des analyses statistiques sur les différents types d'erreurs ;
- enfin, il doit nous être utile pour tester et valider notre modèle de détection d'erreurs.

4.2.1 Caractéristiques communes des données

Tout d'abord, nous travaillons sur les erreurs dans la langue française, même si nous aspirons à développer un outil qui puisse être facilement adapté à diverses langues. Notre corpus doit donc être constitué de textes rédigés en français contemporain. En effet, pour être utiles aux utilisateurs présents et futurs, les vérificateurs linguistiques doivent se fonder sur la grammaire et l'orthographe contemporaines. L'outil que nous développerons s'appuiera sur le corpus, qui doit donc refléter la langue actuelle.

Les textes doivent également bien évidemment contenir des erreurs. Nous avons ainsi choisi d'inclure à notre corpus les textes contenant au minimum une phrase avec une erreur de grammaire ou d'orthographe. De plus, afin que le contexte de chaque erreur soit présent et complet, nous avons conservé les textes entiers et non pas seulement les phrases erronées. En effet, bien que la vérification grammaticale automatique soit pour le moment limitée à la phrase, et bien que la détection d'erreurs à un niveau supérieur (par ex. les références anaphoriques) ne soit pas encore prise en charge par les outils TAL, des travaux sont en cours sur le traitement des anaphores notamment [Boudreau & Kittredge, 2005]. Par ailleurs, pour une analyse linguistique des erreurs, nous avons besoin du contexte complet, même si celui-ci est pour le moment inutile d'un point de vue informatique.

> « *Any instance of language depends on its surrounding context. The details of choice shown in any segment of a text depend –some of them– on choices made elsewhere in the text, and so no example is ever complete unless it is a whole text.*[6] »
>
> [Sinclair, 1991, p. 5]

Ensuite, la vérification automatique ne s'appliquant que dans le cadre de la rédaction de textes électroniques, le corpus ne doit être constitué que de textes dactylographiés. Ces derniers peuvent en effet contenir des erreurs particulières dues à l'utilisation d'un clavier pour la saisie, et qui n'apparaissent pas dans des textes manuscrits. Par ailleurs, pour simplifier les analyses ultérieures du corpus, nous avons décidé de collecter uniquement des documents en version numérique, aucun sous format papier. Ce dernier format impliquerait soit une transcription des textes sur ordinateur, avec le risque de les modifier, même de façon minime, soit une numérisation des documents avec un logiciel de reconnaissance de caractères (OCR : *Optical Character Recognition*), la fiabilité de ces logiciels étant parfois aléatoire, en particulier pour les textes manuscrits.

Enfin, nous souhaitons pouvoir identifier l'auteur de chacun des textes collectés pour d'une part, obtenir son accord pour l'utilisation de son texte et d'autre part, utiliser les informations associées comme filtre de perception du corpus. Nous souhaitons notamment connaître son âge, son sexe, son niveau d'études et sa langue maternelle, qui sont autant de paramètres pouvant influer sur la quantité et le type d'erreurs commises.

Nous résumons ci-après les différentes caractéristiques que doivent posséder les documents que nous recherchons :

- texte intégral,
- en français contemporain,
- tapuscrit,
- sous forme électronique,
- avec des erreurs de grammaire,
- avec l'auteur identifiable.

4.2.2 Variété des scripteurs

Certaines caractéristiques des individus peuvent avoir un impact sur les erreurs qu'ils commettent, il est donc indispensable que notre corpus représente leur variété.

Tout d'abord le niveau d'études des scripteurs est un paramètre important car il est généralement corrélé à la fréquence des erreurs. Plus le niveau est élevé, plus les connaissances en grammaire sont sollicitées, et moins les erreurs sont nombreuses. En effet, Lucci & Millet [1994] ont montré que « la fréquence et le type des variantes orthographiques produites par un scripteur dépendent crucialement de son niveau d'étude et de sa profession ». Pour refléter l'usage général de la langue, notre corpus doit donc comporter des écrits de tous niveaux, aussi bien de personnes non diplômées que de personnes ayant suivi des études supérieures. Nous avons cependant fait le choix de privilégier dans cette diversité les individus supposés avoir dépassé le stade d'apprentissage de l'écrit, en limitant les textes émanant d'une part, des apprenants du français comme langue étrangère ou seconde et d'autre part, des jeunes francophones natifs qui sont encore à

6. Tout exemple de langage dépend du contexte qui l'entoure. Les détails de choix visibles dans un segment donné d'un texte dépendent, pour certains du moins, de choix effectués ailleurs dans le texte. Ainsi, aucun exemple n'est complet à moins d'être un texte entier.

un âge d'apprentissage scolaire de l'écrit. Nous avons pour ces derniers fixé une limite à l'âge symbolique de 16 ans, qui correspond à la fois à l'âge approximatif de passage de l'épreuve de français du baccalauréat (16-17 ans) et à l'âge jusqu'auquel la scolarité est obligatoire en France. Nous pouvons donc penser qu'au-delà de 16 ans la langue écrite est censée être suffisamment maîtrisée.

Si nous avons fait le choix de marginaliser les écrits d'apprenants dans notre corpus, c'est parce qu'ils produisent généralement davantage d'erreurs, dont certaines sont typiques, telles les erreurs morphologiques. Elles peuvent être caractérisées par des erreurs de surgénéralisation [Totereau *et al.*, 1998] (règles de grammaire généralisées et appliquées dans des cas où elles ne devraient pas l'être ⇒ *`les internautes publies`) ou, pour les non-natifs, par des phénomènes d'interférences avec la langue maternelle (application au vocabulaire français de règles de grammaire de la langue maternelle) [Corder, 1980 ; Debyser, 1970]. Ce type d'erreurs mène souvent à des mots n'existant pas dans le lexique et sont donc assimilables à l'orthographe pour la vérification automatique. Elles nécessitent en outre des programmes spécifiques [Granger, 2007], à but pédagogique notamment, avec des rétroactions adaptées. Or, notre objectif est de concevoir un outil d'utilisation quotidienne grand public. Les apprenants constituent cependant une large population potentiellement utilisatrice de vérificateurs grammaticaux. Il nous paraît donc tout de même important d'inclure dans notre corpus des écrits émanant de ce type de scripteurs, même s'ils s'y trouvent en proportion moindre.

À l'inverse, les textes produits par des professionnels de l'écrit, c'est-à-dire les textes qui ont fait l'objet de relectures avant une publication par exemple, nous semblent peu intéressants. Ce type de documents ne contient normalement pas, ou alors très peu d'erreurs, à moins de disposer des premières versions, ce qui est rarement le cas. Nous préférons donc nous concentrer sur des écrits issus de situations quotidiennes d'écriture, n'ayant pas fait l'objet de correction professionnelle.

4.2.3 Variété des situations de scription

Outre la variété des auteurs, notre corpus doit également représenter la diversité des situations de rédaction. Dabène [1987, p. 26] les définit comme « l'ensemble des paramètres qui, à un moment donné, sont à l'oeuvre dans le recours à une pratique relevant de l'ordre scriptural [...] ». Dans notre cas, nous nous concentrons sur les contraintes de formalité et de temps.

Notre corpus doit ainsi refléter des degrés de formalité divers qui influent sur l'attention portée par le scripteur à ses écrits. Le plus haut niveau de formalité concerne les situations où l'écrit a pour destinataire une tierce personne, dans un rapport hiérarchique, avec pour objectif d'obtenir quelque chose, comme un emploi avec une lettre de candidature, ou une bonne note à un devoir. Dans ce cas, l'attention portée à la correction de l'écrit est supposée être maximale. Des travaux d'étudiants (devoir, exposé, mémoire, etc.) ou des documents rédigés dans un cadre professionnel (rapport, lettre, etc.) font donc partie des situations de rédaction avec un haut degré de formalité qui nous intéressent.

Au contraire, si l'écrit est destiné à une personne proche, s'il n'y a pas de relation de hiérarchie, s'il n'y a pas d'enjeu particulier, le niveau de formalité est faible et l'attention portée à la qualité de la rédaction peut-être plus relâchée. Les erreurs se trouvent alors généralement en plus grand nombre. Cela peut être le cas dans des mails, des discussions instantanées, des messages postés sur certains *blogs*, etc. L'auto-surveillance est moindre lorsque le scripteur et le lecteur ne sont

qu'une seule et même personne, c'est-à-dire lorsque l'écrit n'est destiné qu'à soi, comme dans le cas de prise de notes par exemple. Le niveau de formalité étant nul, le risque de trouver de nombreuses erreurs est beaucoup plus élevé.

Nous souhaitons également que notre corpus représente différents niveaux de contrainte de temps lors de la production d'un écrit. Plus le temps dont dispose le scripteur pour rédiger ou relire son texte est réduit, plus le risque de commettre des erreurs est important. En situation d'examen par exemple, le temps limité ne permet pas à l'étudiant d'effectuer toutes les relectures nécessaires pour vérifier son texte, qu'il doit par ailleurs également rédiger rapidement. Bien que le degré de formalité soit très élevé puisque l'enjeu est généralement important (avoir une bonne note, obtenir un diplôme, etc.), l'étudiant laisse des erreurs de grammaire qu'il aurait sans doute corrigées en l'absence de contrainte temporelle. Dans d'autres situations, c'est le temps disponible pendant l'écriture uniquement qui est très limité. Lors de la prise de notes, en plus de l'absence de formalité que nous mentionnions dans le paragraphe précédent, le scripteur dispose de très peu de temps pour réfléchir à l'orthographe et la grammaire. Il est alors fort probable qu'il soit amené à commettre beaucoup plus d'erreurs.

4.2.4 Variété des types de documents

À la variété des scripteurs et des situations de rédaction, s'ajoute la diversité des textes pouvant être produits. Nous avons tenté de répertorier, de manière non exhaustive, ces écrits qui nous intéressent, des écrits électroniques et dactylographiés.

a) Documents de correspondance

Les documents de correspondance peuvent être rédigés avec des niveaux de formalité variés qui dépendent de l'objectif du courrier et de la relation entre l'expéditeur et le destinataire. Ils ne sont en revanche généralement par soumis à des contraintes temporelles. Une lettre de motivation, par exemple, reflète un niveau de formalité d'autant plus élevé que l'enjeu est important (obtenir un emploi). Une lettre à une administration peut-être moins formelle, de même qu'une lettre de réclamation. Les lettres à une personne proche sont, quant à elles, généralement peu formelles, mais cela peut varier en fonction du degré d'intimité entre le scripteur et le destinataire.

Ce type de documents est relativement difficile à obtenir pour notre corpus. D'une part, ils sont souvent manuscrits et d'autre part, ils sont relativement personnels et privés, et donc rarement librement disponibles.

b) Documents rédigés dans un cadre professionnel

Dans les documents rédigés dans un cadre professionnel, qu'il s'agisse de rapports, de comptes-rendus, de notes, etc., le niveau de formalité est généralement assez élevé et la contrainte temporelle plutôt faible. Les erreurs présentes dans de tels documents dépendent beaucoup du niveau d'étude et du poste occupé. Mais ces documents sont souvent régis par le secret professionnel, ce qui les rend également difficilement accessibles.

c) Documents rédigés dans un cadre scolaire

Les écrits produits par des élèves dans un cadre scolaire présentent presque toujours un niveau de formalité important puisqu'ils sont généralement associés à une évaluation. Ce qui influera le plus sur la quantité et les types d'erreurs sera bien sûr l'âge et le niveau d'étude, mais aussi le niveau de contrainte temporelle. Les devoirs « maison » par exemple ne sont normalement pas concernés par la contrainte d'un temps limité. Il peut s'agir par exemple de dissertations, de rapports de stage, de mémoires, d'exercices divers, etc. En revanche, les examens ou devoirs sur table sont l'exemple type d'écrits rédigés avec une très forte contrainte de temps. Toutefois, ils ne sont que très rarement tapuscrits. Quant aux textes rédigés sans visée d'évaluation, ils peuvent avoir des contraintes de formalité et de temps très variables selon la situation et le but dans lesquels ils sont rédigés.

Ce genre de documents s'est avéré être l'un des plus accessibles pour notre corpus car nous nous trouvons dans un milieu universitaire avec un accès facilité aux étudiants et à leurs productions.

d) Textes issus d'Internet

Au sein des sites Internet, les erreurs dans les textes dépendent surtout des auteurs et du sujet traité. Par exemple, les sites d'information, marchands, officiels, etc. sont généralement formels, rédigés avec soin, et contiennent par conséquent peu d'erreurs linguistiques, contrairement aux sites personnels ou *blogs* qui sont susceptibles d'en contenir davantage.

Les forums de discussion ou les commentaires présents sur de nombreux sites sont un autre exemple de ressources en erreurs linguistiques. Ils contiennent en effet une multitude de textes courts écrits par des personnes très diverses et dans lesquels la langue ne fait pas toujours l'objet d'une grande attention. Les niveaux de contrainte et de formalité y sont en effet souvent très faibles.

Évoquons enfin les courriers électroniques qui peuvent contenir plus ou moins d'erreurs. Cela dépend beaucoup, comme pour les documents de correspondance évoqués plus haut, de la relation entre l'expéditeur et le destinataire, et donc du degré de formalité. Mais les listes de diffusion par exemple sont une bonne source de textes. Il est en effet assez facile de contacter les auteurs de mails nous intéressant et de recueillir ainsi des écrits rédigés par des personnes très diverses.

4.3 Caractérisation du corpus de l'étude

Nous venons de présenter un certain nombre de types de données intéressants pour notre corpus. Dans la pratique, nous avons vite pris conscience des difficultés à pouvoir utiliser une grande partie d'entre eux, du fait de leur indisponibilité ou de l'identification impossible des auteurs. Nous avons ainsi dû renoncer à une grande variété de textes, que nous avons tenté de compenser par les textes accessibles que nous avons collectés, mais qui nous a amenée à reconsidérer la question de la représentativité de nos données. Cependant, malgré cet écueil, les textes rassemblés n'en constituent pas moins un corpus, en regard des définitions que nous avons apportées au début de ce chapitre (*cf.* § *Définition de la notion de corpus* p. 64)

4.3.1 Écueils de la collecte des textes

Les textes dactylographiés sont produits en masse avec l'essor de l'informatique, mais derrière cette abondance se cache finalement une grande proportion de documents inutilisables pour constituer notre corpus. Sur Internet notamment, qui est une manne en textes tapuscrits variés, « *a vast and ever-growing repository of texts of every conceivable type*[7] » [Atkins & Rundell, 2008, p. 78], nous avons constaté que les scripteurs sont en fait souvent difficiles, voire impossibles à identifier. Cependant, pour des raisons juridiques évidentes, nous voulions avoir l'autorisation des auteurs pour utiliser leurs textes, lorsque ceux-ci ne faisaient pas mention d'une publication sous licence libre. Nous souhaitions également rassembler des informations associées aux auteurs pour les utiliser ultérieurement comme filtres d'analyse du corpus.

Ainsi, par exemple, la tâche d'identification des auteurs des sites officiels ou professionnels est compliquée par le fait que les textes sont généralement rédigés par plusieurs personnes différentes. Les auteurs des messages postés dans les forums de discussion ou en commentaires de certains sites sont malheureusement aussi presque impossibles à identifier du fait de l'usage de pseudonymes, et aussi parce que beaucoup sont seulement de passage sur le site. Les auteurs de *blogs* sont plus accessibles, cependant nous avons rapidement décidé de laisser ce type de sources de côté. En effet, la lecture de différents sites pour y rechercher des erreurs a monopolisé énormément de notre temps pour ne trouver au final que très peu de textes éligibles : beaucoup de *blogs* ne contiennent quasiment pas d'erreurs, beaucoup d'autres utilisent le langage SMS en masse. Nous avons écarté ces derniers de notre corpus car l'orthographe et la grammaire y sont parfois très éloignées du français standard. Faute de suffisamment de temps disponible, nous avons donc privilégié les textes plus « rentables » (plus courts et donc plus vite lus). Nous avons d'ailleurs également été confrontée au refus de nombreux auteurs identifiés de donner leur autorisation pour l'utilisation de leurs textes, et ce généralement par désintérêt ou par méfiance. Cette contrainte a elle aussi restreint considérablement les possibilités de collecte de textes sur Internet, malgré leur profusion.

À l'inverse, pour beaucoup de documents nous intéressant (issus d'Internet ou pas) l'auteur est très facilement identifiable et disposé à donner son accord. Il s'agit des documents ou correspondance de proches (amis et famille). Cependant, par souci éthique, nous avons décidé de ne pas les inclure à notre corpus.

Nous avions également à disposition de nombreux textes dactylographiés, mais sous format papier. Les auteurs pouvaient être identifiés, mais il aurait fallu numériser ces textes, soit par OCR, soit par saisie au clavier, pour pouvoir les exploiter. Comme nous l'avons indiqué précédemment (*cf.* § 4.1.2 p. 67), outre le coût supplémentaire d'un tel travail, la retranscription de textes peut conduire à l'ajout ou la correction involontaire d'erreurs et en altérer alors l'authenticité. Nous n'avons donc pas non plus utilisé ces documents.

4.3.2 Contenu du corpus et représentativité

> « La représentativité d'un corpus est sa capacité de fonctionner comme une base fiable pour des généralisations sur une langue particulière (générale ou spécialisée). Ceci implique à la fois une taille et une variété de textes adéquates. »
>
> [Marshman, 2003, p. 4]

7. un dépôt énorme et sans cesse grandissant de textes de toutes les sortes imaginables

Un « bon » corpus doit respecter des contraintes de représentativité, d'équilibre et de taille. Pour nos travaux, nous souhaitions obtenir un échantillon représentatif des erreurs d'orthographe, et surtout de grammaire, communément commises à l'écrit, sur un ordinateur, par des scripteurs variés en termes d'âge, de niveau d'étude ou de langue maternelle, dans des situations de scription diverses en termes de contrainte de temps ou de formalité, et dans différents types de documents. Cette représentativité est en fait difficilement atteignable pour notre corpus, même au prix d'un travail extrêmement coûteux, étant donné l'immense panel des textes qui peuvent être produits en fonction des contextes de rédaction très divers évoqués, et également du fait de la difficulté d'obtenir certains types de documents. Nous aurions pu faire l'impasse sur l'identification des auteurs, ce qui nous aurait donné accès à une variété bien plus grande de textes, mais cela n'aurait pas pour autant favorisé la représentativité. En effet, d'après Lucci & Millet [1994], les types et les fréquences des erreurs sont corrélés au niveau d'étude et à la profession. En ignorant les caractéristiques des scripteurs des textes de notre corpus, il devient impossible de savoir si les erreurs qu'il contient sont représentatives, et, le cas échéant, de qui. Notons par ailleurs que, sans une typologie des utilisateurs d'ordinateur et des contextes de rédaction de textes, avec la proportion de chaque type, il est difficile de construire un corpus qui représente fidèlement les erreurs commises pour chacun.

> « Le corpus équilibré est sans doute celui qui a "de tout un peu", mais encore faudrait-il savoir ce qu'est "tout", c'est-à-dire quelles sont les classes à représenter, – ce qui nécessite un modèle complet de la variation –, et avoir accès à des textes les représentant. »
>
> [Péry-Woodley, 1995, p. 218]

Par ailleurs, il ne faut pas confondre la représentativité quantitative des erreurs avec celle des types d'erreurs commises. Nous n'avons sélectionné que des textes qui contenaient des erreurs. Ainsi, notre corpus ne peut et n'aspire pas à représenter la densité de chaque type d'erreur de manière générale dans tel ou tel type d'écrit de tel ou tel type de scripteur. Il ne peut que représenter la proportion des erreurs les unes par rapport aux autres pour chacun.

Devant les difficultés à trouver des textes pour alimenter notre corpus, auxquelles s'ajoutait la contrainte de temps induite par le cadre d'un doctorat pour effectuer nos travaux de recherche, nous n'avons pu faire autrement que de revoir à la baisse nos objectifs de représentativité et d'équilibre. Pour le recueil de nos données, que nous présentons de manière détaillée dans le chapitre suivant (*cf.* chapitre *Constitution du corpus* p. 77), nous nous sommes principalement focalisée sur les scripteurs et les documents du milieu dans lequel nous évoluons, à savoir le milieu universitaire. Ainsi, près de 80% de notre corpus de 252 textes est constitué d'écrits d'étudiants, limitant drastiquement la variété des scripteurs, et par conséquent également les tranches d'âge représentées, puisque la moitié des scripteurs a entre 18 et 20 ans, et que les moins de 30 ans en général constituent à eux seuls 83% des auteurs. Nous ne pouvons cependant pas affirmer que la représentativité du corpus en est d'autant affectée, puisque la part d'étudiants parmi la population des producteurs de textes sur ordinateur n'est pas connue. En revanche, les types de textes récoltés, et surtout les proportions dans lesquelles ils ont été collectés, influent sur cette représentativité et sur l'équilibre du corpus.

Les textes d'étudiants sont constitués de trois types de documents : des commentaires postés sur un *blog*, en FLE, par des étudiants étrangers (14 textes), des résumés de textes scientifiques (34 textes) produits par des francophones natifs et des dictées (151 textes) majoritairement produits par des francophones natifs. Ces dernières représentent une proportion très importante de notre corpus (plus de la moitié), et influent donc nettement sur l'équilibre général du corpus, mais nous avons obtenu un nombre important de documents à moindre coût grâce à elles. Nous

avons mis en place un protocole de saisie des textes, qui nous a permis de pallier les difficultés de collecte d'écrits en récoltant des textes contenant de nombreuses erreurs. Mais nous avons introduit des biais dans l'acquisition de ces données, ne serait-ce que par la situation de scription qui est peu naturelle. Il nous semble en effet que la dictée n'est pas l'exercice le plus couramment réalisé sur ordinateur. Ensuite, nous avons sélectionné les textes à dicter en fonction de leur contenu grammatical et des erreurs qui pouvaient être commises. Nous avons donc ainsi sans doute induit des erreurs. Par ailleurs, les dictées ont été réalisées dans un navigateur Internet, certains ayant la fonction de vérification orthographique activée. Même si ce sont prioritairement les erreurs morphosyntaxiques qui nous intéressent, cela introduit une disparité entre les textes dont les scripteurs ont pu corriger l'orthographe et les autres. Enfin, ce sont des personnes tierces qui ont dicté les textes aux étudiants, et nous n'avons donc pas eu de contrôle sur la manière dont ces dictées ont été réalisées. Nous avons ainsi relevé les mêmes erreurs syntaxiques (mauvaises segmentations des phrases) dans plusieurs dictées d'un même groupe d'étudiants, erreurs qui sont sans conteste dues à la manière dont le texte source a été dicté.

Pour ce qui est des documents qui ne proviennent pas d'étudiants, ils sont constitués de 53 mails échangés au sein de trois listes de diffusion : une liste d'utilisateurs d'OpenOffice.org, une liste traitant de la langue française, et une liste sur la nature. Ces mails nous ont permis d'avoir des écrits de personnes qui ne sont pas étudiantes et qui ont plus de 30 ans.

La répartition des types de scripteurs dans les différentes tâches n'est donc pas homogène. Les dictées ne représentent que des étudiants de premier cycle (L1 à L3), de même que les résumés ne représentent que les deuxième et troisième cycles (M1, M2 et Doctorat). Les mails n'ont été rédigés que par des personnes n'étant plus étudiantes mais ayant pour moitié un niveau M2, et enfin les commentaires de blogs ne sont issus que d'étudiants en FLE.

Les obstacles que nous avons rencontrés dans notre collecte de textes ne nous ont donc pas permis de constituer un corpus véritablement équilibré en ce qui concerne les types de scripteurs ou de documents. Il est difficile en revanche d'évaluer si les différentes situations de scription sont bien représentées. La contrainte de temps est présente sans aucun doute dans les dictées, mais nous ne pouvons pas être sûre qu'il n'en est pas de même pour une partie des autres documents. Certains résumés, par exemple, ont été rédigés comme s'ils avaient été contraints par une durée limitée, alors qu'il n'existait pas de limite de temps pour réaliser l'activité. De même, le degré de formalité attaché aux textes est difficile à mesurer. Les commentaires du *blog* de FLE étant destinés à être lus par un enseignant, il est probable que des étudiants se soient particulièrement appliqués, dans l'idée d'une possible évaluation. Mais pour les dictées, alors que l'exercice était réalisé de manière anonyme, l'analyse des textes montre que beaucoup se sont pris au jeu et ont fait particulièrement attention à leur correction, quand d'autres au contraire ont réalisé l'exercice « à la va-vite ». Nous ne pouvons donc pas attribuer de manière systématique le même critère de temps ou de formalité à des documents du même type. Nous pouvons seulement constater, sans pouvoir le quantifier, que notre corpus contient des documents variés en termes de contraintes de temps et de formalité.

4.3.3 Positionnement de notre corpus

Malgré les réserves que nous venons d'avancer sur l'équilibre et la représentativité des textes que nous avons recueillis et que nous présentons dans leur ensemble dans le chapitre suivant (*cf.* § *Recueil des textes* p. 77), au regard des définitions citées au début de ce chapitre (*cf.* § *Définition de la notion de corpus* p. 64), nous pouvons dire que nos textes constituent bien un corpus et non

une simple collection de documents. En accord avec la définition de Sinclair [2005], ils ont avant tout été « sélectionnés selon des critères externes » [8], le principal étant leur mode de production (tapuscrits). D'autres critères sont également entrés en jeu, tel le type de documents, afin de privilégier ceux qui sont le plus susceptibles de contenir des erreurs (les textes de professionnels de l'écrit sont ainsi écartés), ou encore l'identification possible de l'auteur. Bien sûr, la présence d'erreurs morphosyntaxiques fait partie des critères importants qui ont guidé la sélection des textes. Il s'agit-là d'un critère plutôt linguistique, et donc interne, mais pour Clear [1992, p. 29], « *A corpus selected entirely on external criteria would be liable to miss significant variation among texts since its categories are not motivated by textual (but by contextual) factors* [9] ». Il n'est donc pas problématique que des critères internes interviennent dans la constitution d'un corpus. Dans notre cas, ne pas sélectionner les textes en fonction de la présence ou non d'erreurs aurait pu nous conduire à un corpus dans lequel le nombre d'occurrences d'erreurs aurait été en faible quantité. Nous n'aurions alors pas pu l'exploiter de manière satisfaisante.

Nos textes forment ainsi un échantillon du langage, « *a sample of the language* » [Sinclair, 1996, p. 4] : le langage erroné. Ils sont par ailleurs stockés (et générés) électroniquement ce qui, selon Williams [2005] et Sinclair [2005], est un critère de définition de la notion de corpus, et constitue même pour Williams [2005] un point commun aux diverses approches.

Enfin, la taille modeste de notre corpus, avec 33 232 mots à ce jour (*cf.* § 6.2 p. 109), peut lui conférer le statut de petit corpus d'étude, se rapprochant du corpus spécialisé par son contenu uniquement formé de textes dactylographiés contenant des erreurs grammaticales et orthographiques. Nous faisons référence ici à une distinction fondamentale qui est faite, parmi les différents types de corpus (écrits, oraux, comparables, parallèles, monolingues, multilingues, etc.), entre les corpus de référence et les corpus de spécialité. Les premiers sont conçus dans le but de représenter toutes les variétés de la langue, toutes les situations de communication, afin de servir de fondement pour leur étude et la constitution d'usuels langagiers.

> « *A reference* corpus *is one that is designed to provide comprehensive information about a language. It aims to be large enough to represent all the relevant varieties of the language, and the characteristic vocabulary, so that it can be used as a basis for reliable grammars, dictionaries, thesauri and other language reference materials.* [10] »
>
> [Sinclair, 1996, p. 10]

Contrairement aux corpus de référence, les corpus spécialisés regroupent des documents sur une thématique particulière, une situation de communication précise, un domaine, etc. Ces corpus sont toujours relativement restreints de par leur spécialité, et il est plus difficile de recueillir des données pour les alimenter. L'équilibre et la représentativité y sont également plus difficiles à atteindre que dans les corpus de référence [Habert, 1998], mais n'en demeurent pas moins une préoccupation importante de leurs concepteurs. Ils sont par ailleurs généralement constitués pour l'observation d'un phénomène linguistique particulier, comme par exemple les erreurs morphosyntaxiques en contexte tapuscrit dans notre cas, et perdent leur utilité en dehors de la recherche pour laquelle ils ont été créés. Ils sont alors rarement réutilisables dans le cadre

8. « *selected according to external criteria* » [Sinclair, 2005]. (en anglais dans le texte original).

9. Dans un corpus sélectionné uniquement selon des critères externes, il risquerait de manquer des variations significatives parmi les textes, puisque ses catégories ne sont pas motivées par des facteurs textuels, mais contextuels.

10. « Un corpus de référence est conçu pour fournir une information en profondeur sur une langue. Il vise à être suffisamment grand pour représenter toutes les variétés pertinentes de cette langue et son vocabulaire caractéristique, de manière à pouvoir servir de base à des grammaires, des dictionnaires et d'autres usuels fiables. » (traduit par [Habert, 1998], p. 36).

d'autres recherches.

Notre corpus s'apparente ainsi à un petit corpus spécialisé. Il nous a permis d'établir une première typologie des erreurs tapuscrites, que nous pourrons faire évoluer par la suite en complétant le corpus avec d'autres types de documents.

Chapitre 5

Constitution du corpus

Sommaire

5.1 Recueil des textes 77
5.1.1 Dictées 77
5.1.2 Résumés 79
5.1.3 Courriers électroniques 80
5.1.4 Commentaires de blog 80
5.2 Normalisation des données 81
5.2.1 Stockage homogène des données 81
5.2.2 Standards d'annotation : XML, TEI, CES 83
5.2.3 Normalisation des données 86

Ayant montré l'intérêt d'un corpus pour nos travaux et défini le type de données textuelles que nous souhaitions rassembler pour le constituer, nous présentons à présent le recueil de celles que nous avons effectivement collectées et que nous avons brièvement évoquées dans le chapitre précédent. Nous avons recueilli des textes de quatre types : des dictées, des résumés et des commentaires de blog, tous trois réalisés par des étudiants, ainsi que des mails issus de listes de diffusion. Nous expliquons dans ce chapitre comment nous avons procédé pour réunir ces documents, puis nous décrivons les traitements que nous leur avons appliqués afin de les normaliser et de les préparer à l'annotation des erreurs. Nous abordons pour cela également le sujet des standards d'encodage de corpus en XML.

5.1 Recueil des textes

5.1.1 Dictées

Notre corpus est né de la collecte de dictées réalisées par des étudiants à partir de textes issus d'Internet. Le principe de la dictée, avec une contrainte temporelle forte, nous a semblé intéressant pour recueillir un grand nombre d'erreurs de grammaire. Il a en effet été montré qu'un temps limité pour la rédaction est un facteur augmentant les erreurs [Lucci & Millet, 1994]. L'article de Largy *et al.* [2005], synthétisant les résultats d'une série de recherches sur la production et la révision de textes (en particulier l'accord verbal en nombre), apporte une explication à ce phénomène.

> « [...] les experts comme les apprenants sont à même d'appliquer une règle d'accord (en production comme en révision) en mettant en oeuvre de manière consciente l'algorithme correspondant, mais [...] seuls les experts peuvent être performants dans des situations où le scripteur ne dispose pas du temps et/ou des ressources nécessaires à la mise en oeuvre de l'algorithme. L'expert semble ainsi s'appuyer sur un autre type de procédure pour gérer l'accord. »
>
> [Largy *et al.*, 2005, p. 347]

L'article met aussi en évidence que l'expert, toujours dans le cas de l'accord du verbe en nombre, bénéficie d'une « capacité à juger très efficacement du caractère exact ou erroné d'un accord en se fondant sur le repérage de cooccurrences proximales entre morphèmes flexionnels (en particulier l'association -s-nt) » [Largy *et al.*, 2005, p. 348]. Paradoxalement, en situation de temps limité, pendant une dictée notamment, l'expert est exposé à des erreurs dues à ce principe d'accord de proximité hautement automatisé [Largy, 2003] (par ex. *`*la correction des exercices se limitent à...*`).

Les études réalisées dans ce domaine s'attachent en grande majorité aux accords nominaux et verbaux en nombre, mais nous pouvons émettre l'hypothèse que d'autres phénomènes grammaticaux sont concernés par le recours privilégié aux processus cognitifs automatisés de l'expert, au détriment de l'utilisation de règles de grammaire, lors de situations de scription contraintes par le temps.

Le choix des textes à donner en dictée s'est fait selon plusieurs critères. Nous voulions des textes de longueur réduite, qui soient attrayants par leur thématique pour ne pas ennuyer les étudiants, et qui comportent si possible quelque difficulté grammaticale. Nous avons ainsi porté notre choix sur des articles de nouvelles insolites et avons conservé dix dépêches, composées de 121 à 230 mots, avec une moyenne de 180 mots par article.

Nous avons en second lieu demandé la collaboration de collègues enseignants en informatique pour qu'ils fassent réaliser les dictées sur ordinateur par leurs étudiants. Ils ont accepté de prendre 5 à 10 minutes sur leur temps de cours pour faire faire l'exercice. Nous leur avons donné pour consignes de choisir un texte parmi les dix proposés et de le dicter à leur(s) groupe(s) d'étudiants, comme un professeur des écoles le ferait avec ses jeunes élèves, avec pour seule différence de ne pas leur accorder de temps à la fin pour la relecture. Nous leur avons par ailleurs demandé de ne pas révéler le but de l'exercice avant la fin de l'expérience, afin que les étudiants ne soient pas influencés. Ces derniers étaient seulement informés de leur participation anonyme à une expérience de recherche, sans évaluation de leur écrit et sans lien avec leur cursus.

Du côté des étudiants participant à l'expérience, la dictée était à réaliser sur une page Web dédiée que nous avons développée spécifiquement. L'interface [1], constituée d'un champ de saisie de la dictée et de quatre champs de saisie des informations sur l'étudiant, a permis l'enregistrement dans une base de données de la dictée et de toutes les informations utiles sur le scripteur. Les informations à saisir par l'étudiant concernaient :

- son genre : M ou F à cocher pour masculin ou féminin ;
- son âge : l'âge saisi n'a pas été pas conservé tel quel mais enregistré sous la tranche d'âge correspondante, à savoir 15-20, 21-30, 31-40, 41-50, 51-60 et plus de 60 ;
- son niveau d'études : un menu déroulant permettait d'indiquer son niveau d'études actuel, à savoir L1, L2, L3, M1 ou M2 ;
- sa langue maternelle : la plupart ont saisi « français ».

1. http ://www.asouque.fr/THESE/corpus/dictees.php

D'autres informations, non visibles sur le formulaire, ont également été enregistrées à la fin de chaque dictée :

- la présence d'une contrainte de temps ;
- un degré de formalité faible [2] ;
- la date de saisie du texte ;
- le type du document, à savoir « Dictée ».

Outre l'avantage de permettre l'enregistrement automatique des données tout en conservant l'anonymat des étudiants, cette interface empêche que l'authenticité de certaines erreurs soit altérée par l'utilisation d'un correcteur grammatical intégré à l'outil de saisie. Cette fonctionnalité n'était en effet pas activée dans les navigateurs Web des ordinateurs utilisés. En revanche, en faisant dicter les textes par des personnes tierces, un certain nombre de points ont échappé à notre contrôle. Il nous a par exemple été rapporté après l'expérience que, sur quelques ordinateurs, la fonction de correction orthographique du navigateur était activée. Par ailleurs, dans certains groupes, le nombre d'ordinateurs était inférieur au nombre d'étudiants qui étaient alors deux par poste. Il en résulte que certaines dictées ont probablement été réalisées à deux. Cependant, nous n'avons pas de moyen de savoir quelles sont les productions concernées, ni les éventuelles implications sur les erreurs commises ou non. De même, nous ne connaissons pas la durée qui s'est écoulée entre la fin de la dictée et l'enregistrement des données par l'étudiant, durée qui correspond à une ou plusieurs relectures du texte, et donc vraisemblablement à la correction d'erreurs. Enfin, nous n'avons pas pu contrôler la manière dont les textes ont été dictés : la vitesse d'énonciation, la durée laissée pour la saisie ou encore le nombre de répétitions des propositions sont autant de facteurs pouvant influer sur la quantité et la nature des erreurs. Des textes d'un même groupe présentent ainsi des similitudes au niveau d'erreurs de segmentation des phrases, erreurs qui ne se retrouvent pas dans les mêmes textes dictés à d'autres groupes par d'autres personnes.

Nous avons finalement recueilli 151 dictées, réalisées par des étudiants de L1 à L3 de filières de lettres et langues. Parmi eux, six sont non francophones natifs.

5.1.2 Résumés

Nous avons également intégré à notre corpus 34 résumés réalisés par 9 étudiants francophones, du M1 au doctorat, au cours d'une expérience de recherche menée dans le cadre d'un projet européen (projet LTfLL [3]). L'objet de l'expérience était le test d'APEX 2.1 [4], outil d'aide à la révision de cours pour la préparation d'examen. Plus précisément, il s'agissait de rédiger sur ordinateur des résumés à partir de textes scientifiques longs, et de faire évaluer ces résumés par le logiciel. Ce dernier devait alors indiquer, d'après le contenu de l'écrit produit, si le texte long avait été bien compris ou pas. Toutes les actions des sujets étaient par ailleurs enregistrées à des fins d'analyse.

Chaque sujet de l'expérience avait donc pour tâche de lire des textes scientifiques dans le domaine du TAL et/ou des Environnements Informatiques pour l'Apprentissage des Langues (EIAL), et de rédiger au fur et à mesure sur l'ordinateur les résumés d'au moins 5 d'entre eux,

2. Ceci peut être discutable. En effet, les dictées sont des exercices généralement associés à une évaluation, et donc à un degré de formalité élevée. Cependant, les étudiants étaient informés de la non évaluation et de l'anonymat de leur texte, ce qui réduit fortement ce degré de formalité.

3. Projet *Language Technologies for Lifelong Learning* - http ://partners.ltfll-project.org/

4. APEX : Aide à la Préparation d'EXamen [Dessus *et al.*, 2009 ; Zampa & Dessus, 2010].

au choix, afin de montrer leur bonne compréhension. L'expérience avait une durée d'environ 50 minutes, et interdisait d'effectuer des copier-coller du texte original. À la fin de la rédaction du résumé d'un texte, le sujet devait le faire évaluer automatiquement par le logiciel APEX.

Les résumés n'étant pas destinés à être lus par un humain, mais évalués par un logiciel, le degré de formalité était par conséquent plutôt faible. En revanche, les données temporelles enregistrées montrent que le temps passé à la rédaction est très variable selon les sujets et les textes, allant de 1'49 s à 27'40 s, avec une moyenne de 7'05 s et un écart-type de 4'16s. Nous pouvons alors émettre l'hypothèse que ces variations pourraient avoir un impact sur la quantité d'erreurs d'écriture contenues dans les résumés.

5.1.3 Courriers électroniques

Pour diversifier les types de scripteurs et sortir du milieu étudiant, nous avons aussi recueilli des mails envoyés sur des listes de diffusion. Étant nous-même inscrite sur celle des utilisateurs francophones d'OpenOffice.org (« users-fr »), nous avons sollicité les autres utilisateurs afin qu'ils nous autorisent à utiliser leurs textes et acceptent de nous indiquer leur âge, leur niveau d'études, et leur langue maternelle. 16 personnes non étudiantes nous ont donné leur accord. Nous avons alors passé en revue leurs courriers électroniques disponibles, dans les archives de la liste de diffusion, et nous avons conservé 44 documents qui contenaient au moins une erreur de grammaire ou d'orthographe.

Afin de compléter ces données, et puisque nous avions fait le choix de ne pas collecter les mails de notre correspondance personnelle pour des raisons éthiques, nous nous sommes inscrite à deux autres listes : une ayant pour thème assez général la nature (« natur-naute »), et une sur la langue française (« mots_passion »). Cela n'a malheureusement pas été très fructueux. Peu de personnes ont accepté que nous utilisions leurs textes et, parmi celles ayant donné leur accord, la plupart n'avaient pas fait d'erreur dans leurs mails. Nous n'avons finalement recueilli que neuf documents, issus de trois scripteurs différents.

S'agissant de courriers électroniques, il n'y avait en principe aucune contrainte de temps pour la rédaction. En revanche, le degré de formalité était très variable. Nous avons notamment remarqué que les mails issus de la liste de diffusion sur la langue française, probablement du fait même du sujet traité, avaient fait l'objet de plus de soin dans la correction que ceux issus des deux autres listes.

5.1.4 Commentaires de blog

Enfin, pour que la population de scripteurs francophones non natifs soit représentée dans notre corpus, nous avons également collecté, sur le blog [5] d'une enseignante de FLE, 14 commentaires rédigés en français par 7 étudiants étrangers. Sur ce blog, spécialement destiné aux étudiants en FLE de l'enseignante en question, chaque billet est une consigne de rédaction d'un court texte. Chaque étudiant réalise ensuite l'exercice de rédaction par l'ajout d'un nouveau commentaire au billet. Il est aisé, à l'aide des signatures des étudiants, d'identifier par qui a été écrit chaque texte, et d'associer ainsi les informations sur l'âge, le sexe, le niveau d'études et la langue maternelle.

Le tableau 5.1 ci-après résume les caractéristiques des différents types de documents compo-

5. http ://mp-campus.blogspot.com

sant notre corpus.

Type	Nombre	Scripteurs	Français	Temps	Situation	Limites
Dictées	151	étudiants de L1 à L3	natif en majorité	limité	variable	variété des textes limitée, formalité difficilement appréciable
Résumés	34	étudiants de M1 à D	natif	plutôt libre	informelle	contrainte de temps ajoutée par certains scripteurs
Mails	53	non étudiants	natif	libre	informelle	pas d'informations sur le contexte de rédaction
Blogs	14	étudiants	FLE	libre	formelle	nombre de textes et variété des scripteurs très limités

Tableau 5.1 : Synthèse des textes du corpus

5.2 Normalisation des données

Notre corpus ainsi constitué n'échappe pas au constat que fait Habert [1998, p. 12] : « L'hétérogénéité est une constante, même au sein des documents qui ont été directement dactylographiés ». Les textes que nous avons collectés ont en effet été rédigés dans des logiciels divers (traitement de texte, client mail, navigateur Internet, etc.), chacun ayant ses propres spécificités d'encodage de caractères et de formatage de document. Il devient alors nécessaire, pour exploiter les données textuelles, de les homogénéiser, et donc, comme le précise Habert [1998, p. 51], de « les transformer pour les mettre dans un format commun ».

Dans cette section, nous décrivons la première étape de normalisation de nos textes, qui consiste en leur enregistrement dans une base de données. Nous présentons ensuite le langage de représentation *eXtensive Mark-up Language* (XML), dont les propriétés en font un langage incontournable dans la description de documents et l'annotation de corpus. C'est sur ce langage que reposent notamment deux propositions de standard pour le balisage de corpus, à savoir la *Text Encoding Initiative* (TEI), prééminente dans le domaine, ou encore le *Corpus Encoding Standard* (CES). Plusieurs raisons nous ont cependant conduite à ne pas nous conformer à ces propositions de standard pour l'annotation normalisée de nos textes, au profit d'un modèle *ad hoc* plus simple, qui repose tout de même toujours sur le langage XML, et que nous détaillons ci-après.

5.2.1 Stockage homogène des données

Nous avons précisé, dans la section précédente (*cf.* § 5.1.1 p. 77), que les dictées ont été enregistrées directement dans une base de données à la fin de leur saisie. Afin de garantir une homogénéité dans le mode de stockage de nos textes et de faciliter le traitement ultérieur de toutes les données, nous avons également procédé à l'enregistrement des résumés, des mails et des commentaires de blog dans la même base de données.

Pour simplifier ces enregistrements, nous avons conçu une interface Web qui permet de saisir

dans un formulaire, pour chaque texte, toutes les informations importantes concernant le scripteur, le contexte de rédaction, le type de document et le texte lui-même. Nous renseignons ainsi le genre, l'âge, le niveau d'études et la langue maternelle du scripteur. Nous indiquons, dans la mesure du possible, si le texte a été rédigé avec ou sans contrainte de temps ou dans un contexte formel ou informel. Nous précisons également le type de texte (mail, résumé, etc.), le format original du document électronique (.doc, .html, etc.), la date de rédaction, et enfin nous insérons le texte lui-même dans son intégralité. La validation du formulaire entraîne l'enregistrement de toutes ces informations ainsi que le nombre de mots du texte, comptabilisé automatiquement par le script PHP d'enregistrement.

Nous avons également procédé, avant leur sauvegarde dans la base de données, à l'anonymisation des mails et des commentaires de blog. Certains d'entre eux contenaient en effet des prénoms qui remettaient en cause l'anonymat. L'article de Rock [2001, p. 1] met en garde contre le danger de penser que « l'anonymisation peut être obtenue en modifiant ou en enlevant uniquement les noms[6] », mais après observation de nos textes, il nous est apparu que seuls les prénoms pouvaient effectivement constituer un problème. Les méthodes d'anonymisation répertoriées par Rock [2001] pour les informations de ce type montrent qu'il est généralement d'usage d'utiliser des codes ou des prénoms de remplacement qui permettent de conserver le genre du nom anonymisé. Cependant, dans les textes d'étudiants en FLE, il nous était souvent impossible de déterminer si les prénoms mentionnés, tous d'origine étrangère, désignaient des personnes de sexe féminin ou masculin. Nous avons donc pris le parti de substituer aux prénoms le code `NOM` dans les textes, précédé de la lettre `f` ou `m` (`fNOM`, `mNOM`) pour les prénoms féminins ou masculins, et de la lettre `u` (`uNOM`) dans les cas ambigus. Les dictées et résumés, quant à eux, n'étaient pas concernés par l'anonymisation. S'appuyant respectivement sur des articles de presse publics et des articles scientifiques, ils ne contenaient en effet pas de données personnelles relatives aux scripteurs.

L'enregistrement en base de données, en plus de permettre le stockage d'informations de mêmes types pour chaque document, a été l'occasion de normaliser l'encodage des caractères de tous les textes. En effet, dès que l'on travaille sur des données textuelles électroniques de provenances variées (en termes de langues, de logiciels, de systèmes d'exploitation, etc.), les différents encodages dans lesquels elles ont été créées sont à l'origine de nombreuses difficultés pour leur exploitation ou leur échange. Il est alors indispensable d'effectuer une normalisation en convertissant les textes dans un seul et même code. Dans notre cas, il s'agit du standard de codage de caractères UNICODE (pour *UNIfication CODE*), et plus précisément du code UTF-8 qui en est issu. McEnery & Xiao [2005] recommandent « *UTF-8 as a universal format for data exchange in Unicode, and for corpus construction so as to avoid the textual Tower of Babel*[7] ». Gillam [2003, p. 204] pense également que « *UTF-8 is likely to remain the most popular way of exchanging Unicode data between entities in a heterogeneous environment*[8] ». Par ailleurs, UTF-8 constitue la norme pour l'encodage des caractères des fichiers XML, format de fichier qui, comme nous l'avons évoqué, est celui que nous avons utilisé pour l'annotation de notre corpus.

Notre base de données contient donc finalement chaque texte collecté pour notre corpus dans un format brut, un codage de caractères unique et, pour chacun des textes, un ensemble de métadonnées constituées des informations sur le scripteur (âge, sexe, niveau d'études et langue

6. « *[...] anonymisation can be achieved by altering or removing just names [...]* »

7. UTF-8 comme format universel pour l'échange de données en Unicode et pour la constitution de corpus, pour éviter la Tour de Babel textuelle.

8. Il est probable qu'UTF-8 reste la manière la plus répandue d'échanger des données Unicode entre des entités dans un environnement hétérogène.

maternelle), sur les conditions de scription (temps et formalité) et sur le texte (type, format et nombre de mots). Nous avons également conservé les fichiers originaux contenant les textes (non anonymisés), à l'exception des dictées qui ont été directement enregistrées dans la base de données.

5.2.2 Standards d'annotation : XML, TEI, CES

Nos données ainsi sauvegardées et uniformisées en base de données, nous avons pu entreprendre de les exploiter, en commençant par leur annotation. À cette fin, nous nous sommes dans un premier temps naturellement tournée vers le langage XML et les propositions de standards d'annotation qui l'utilisent, que nous présentons dans cette section, et dont l'observation nous a finalement conduite à créer un modèle d'étiquetage sur mesure pour notre corpus.

a) XML

XML, « langage extensible de balisage », est un langage de représentation né en 1998 de la simplification de son prédécesseur SGML (*Standard Generalized Markup Language*)[9]. Il permet comme lui de décrire des contenus en se fondant sur un système de balises de délimitation (avec une balise ouvrante et une fermante, par ex. `<DOCUMENT>...</DOCUMENT>`), appelées éléments, et présente diverses caractéristiques intéressantes pour l'annotation des données de corpus notamment.

Par exemple, le principe des balises, qui peuvent s'enchâsser mais pas se chevaucher, est particulièrement adapté à la description de la structure des documents et des éléments qui le constituent, en les délimitant de manière logique. Il est idéal notamment pour les corpus de textes, dans lesquels des délimitations peuvent être effectuées sur différents niveaux hiérarchiques, tels que les mots dans les phrases, les phrases dans les textes, les textes dans le corpus, etc., ou encore sur des segments de textes spécifiques, comme des citations, des passages dans une autre langue ou, dans notre cas, des erreurs morphosyntaxiques. Le langage XML permet également, grâce à des attributs associés aux balises, de spécifier des propriétés à chaque élément délimité. Une balise délimitant un mot, par exemple, pourra posséder des attributs spécifiant les traits morphosyntaxiques du mot (`<MOT cat="nom" genre="fem" nombre="pluriel">solutions</MOT>`). Concernant notre corpus, cette propriété est intéressante notamment pour indiquer dans les balises délimitant les erreurs le type des erreurs concernées. Elle permet également d'affecter un numéro unique à chaque élément, afin de l'identifier facilement par la suite, et pouvoir par exemple faire référence à une erreur en indiquant son identifiant, ou celui de la phrase ou du texte dans lesquels elle se trouve.

Les documents XML sont par ailleurs généralement associés à une grammaire *Document Type Definition* (DTD)[10], qui définit et contraint les balises et les attributs qu'ils doivent utiliser pour être valides. Ceci permet notamment de normaliser les descriptions de différents documents effectuées dans plusieurs fichiers XML distincts, lorsqu'ils sont liés à une même DTD, et de s'assurer ainsi de l'homogénéité de la description de tous les textes d'un corpus. Un très grand nombre de documents à annoter peut en effet facilement conduire à des divergences d'encodage, qui gênent ensuite de manière plus ou moins importante l'exploitation des données annotées. Pour

9. Langage normalisé de balisage généralisé.
10. Définition de type de document.

notre corpus, sans utiliser une DTD qui fixe par exemple la liste des types d'erreurs possibles, nous aurions probablement des variations dans l'annotation des erreurs, comme par exemple `type="accord-nom"`, ou `type="accord-nb-nom"` ou encore `type="Acc-Det-Nom"` pour l'erreur **cet outils*. Ceci aurait pour conséquence de générer des difficultés à faire le lien entre des erreurs appartenant au même type mais annotées de manière variable, et donc à les exploiter par la suite, notamment via des traitements automatiques. L'accès à la grammaire associée à un document XML facilite en effet également la création d'un script adapté pour le parcourir et en extraire des informations ou en indexer le contenu dans une base de données notamment. Tous les éléments utilisés pour structurer et annoter le document étant préalablement définis et contraints par les règles de la DTD, l'automatisation d'opérations comme l'extraction de certaines données pour des analyses statistiques ou linguistiques, ou l'indexation d'informations telles que les erreurs annotées pour notre corpus, s'en retrouvent par la suite facilitée. De plus, l'attribution de noms explicites aux éléments et aux attributs dans la DTD permet par la suite une lecture plus aisée et une meilleure compréhension par l'humain des fichiers XML se conformant à cette DTD. Un corpus annoté peut alors être parcouru sans avoir besoin d'outil particulier.

Il est également possible de lier une feuille de style *Cascading Style Sheet* (CSS)[11] à un document XML, ce qui permet la mise en forme et la visualisation du contenu du document, sans les balises, dans un navigateur Web par exemple. Cette fonctionnalité peut nous permettre, par exemple, d'afficher des extraits du corpus en mettant en évidence spécifiquement les erreurs et leur correction. Il est également possible d'envisager plusieurs feuilles de style associées à un unique document XML, qui permettront autant de mises en forme différentes pour visualiser le contenu, en fonction des besoins : par exemple un affichage des textes en bloc ou bien phrase par phrase, un affichage des erreurs en rouge et des corrections en vert, un affichage sans les corrections, etc.

Enfin, une dernière caractéristique intéressante dans le langage XML consiste en la possibilité, via des feuilles de style *eXtensible Stylesheet Language* (XSL)[12], de transformer un document XML en un document d'un autre type, ou de même type mais structuré différemment. Si nécessaire, nous pourrons ainsi exporter nos données dans d'autres formats pour les partager.

Avec ces différentes propriétés, le langage XML est ainsi très prisé pour la description de ressources linguistiques, dont bien sûr les corpus. Il sert d'ailleurs de support à deux propositions de standard d'annotation de corpus très répandues, la TEI et le CES qui en dérive, auxquelles nous nous sommes donc intéressée afin d'évaluer leur adéquation avec nos besoins pour l'annotation de nos données.

b) TEI

La TEI est un projet communautaire né en 1987, à l'initiative de trois associations professionnelles[13] du domaine du TAL, afin de répondre au besoin d'une normalisation de l'encodage des corpus, pour en faciliter l'échange notamment, dans la communauté des chercheurs en lettres, langues et sciences humaines.

« *The goal of the TEI is to develop and disseminate a set of Guidelines for the in-*

11. Feuille de style en cascade.
12. Langage extensible de feuille de style.
13. *Association for Computational Linguistics (ACL), Association for Literary and Linguistic Computing (ALLC) et Association for Computing and the Humanities (ACH).*

> *terchange of machine-readable texts among researchers, so as to allow easier and more efficient sharing of resources for textual computing and natural language processing.* [14] »
>
> Burnard [1991]

Initialement fondée sur le métalangage SGML, elle est aujourd'hui entièrement conforme au langage XML, son successeur, dont nous venons de présenter l'intérêt pour l'annotation de corpus. Le consortium TEI a ainsi mis à disposition des chercheurs des directives pour la création de DTD pour le balisage des textes, qui simplifient à la fois le partage des corpus numérisés et leur traitement par ordinateur.

Ces recommandations de description sont reconnues internationalement et utilisées par de nombreux projets. Il semblerait donc naturel de les adopter pour l'encodage de notre corpus, plutôt que de concevoir un nouveau système. Cependant, sans pour autant nier leur intérêt et leur utilité, nous avons, pour des raisons surtout pratiques, préféré élaborer un modèle *ad hoc*.

Tout d'abord, parce qu'elle se veut utilisable par toutes les disciplines des lettres, langues et sciences humaines, la TEI est très riche et relativement lourde à assimiler. Elle définit un nombre très important de descripteurs (plus de 400), dont les noms ne sont pas toujours explicites, dont les attributs sont parfois complexes, et parmi lesquelles il faut sélectionner, voire modifier, celles qui semblent correspondre à nos besoins. Cette assimilation du balisage TEI requiert un investissement en temps relativement élevé dont nous ne disposions pas et qui ne nous pas paru pertinent à notre échelle. L'annotation de notre corpus ne nécessitait en effet qu'un nombre très restreint d'éléments.

Par ailleurs, pour analyser nos données, nous n'avions pas besoin des outils puissants conçus autour de la TEI. Faute d'outil adéquat, nous avons développé le nôtre. Nous avons alors préféré avoir à traiter un modèle simplifié pour décrire nos données, contenant uniquement les éléments strictement nécessaires.

Enfin, la TEI a principalement pour but de faciliter l'échange des données, ce qui n'est pas notre objectif premier. Nous souhaitons en effet avant tout annoter les erreurs de notre corpus de manière aussi simple et efficace que possible, afin de pouvoir les exploiter plus facilement pour nos travaux. Mais cela ne nous empêche pas de nous préoccuper également du partage et de la réutilisabilité de notre corpus. Notre système d'annotation, simple et épuré, peut ainsi être aisément modifié afin de s'adapter, si besoin, aux recommandations de la TEI (*cf.* § 5.2.3 p. 86).

c) CES

Le *Corpus Encoding Standard* (CES) est issu du projet EAGLES [Sinclair, 1996 ; EAGLES, 2000]. Il dérive de la TEI dont il reprend et simplifie les propositions, tout en les complétant. Comme elle, il vise à la réutilisation des corpus et donc à la facilitation de leur échange. En revanche, alors que la TEI ne propose que des directives pour construire une DTD pour l'annotation d'un corpus, le CES fournit directement trois types de DTD auxquelles doivent se conformer les fichiers annotés. Une première DTD (cesDoc) définit le balisage général de la structure et des divers éléments du corpus, une seconde (cesAna) est consacrée à l'annotation morphosyntaxique des données, et enfin une troisième (cesAlign) est spécifique à l'alignement de corpus.

14. Le but de la TEI est de développer et diffuser un ensemble de directives pour l'échange de textes électroniques entre chercheurs, afin de permettre un partage plus facile et plus efficace de ressources pour l'exploitation informatique de textes et le traitement automatique des langues.

Seul le premier type de DTD nous serait utile pour notre corpus, puisque nous n'envisageons pas de faire pas de description morphosyntaxique ni d'alignement. Il définit trois niveaux distincts d'annotation, de la plus large (niveau 1, jusqu'au paragraphe) à la plus fine (niveau 3, au sein des paragraphes). Le niveau 1 constitue le minimum requis pour être conforme au CES et est nécessaire pour pouvoir ensuite effectuer les annotations de niveau 2, ces dernières étant à leur tour un préalable pour les annotations de niveau 3, les plus précises.

Bien que simplifiées par rapport à la TEI, les propositions d'encodage du CES ne conviennent pas au balisage que nous souhaitions faire de notre corpus. En effet, l'annotation minimale obligatoire est la délimitation des paragraphes (niveau 1), qui ne nous est pas utile pour nos travaux. En revanche, nous voulions délimiter les phrases et des segments de phrases, ce qui correspond au niveau 3 d'annotation CES et nécessite donc que les annotations des niveaux précédents soient préalablement effectuées. Réaliser un balisage de notre corpus en nous conformant au CES nous conduirait ainsi à surcharger nos données d'annotations non nécessaires à nos analyses ultérieures.

Notre modèle d'annotation ne suit donc pas les propositions du CES, ni celles de la TEI. Cependant, comme nous allons le voir dans le paragraphe suivant, il en reste relativement proche en se fondant comme eux sur le langage XML, parfaitement adapté à cette tâche. Une transposition vers l'un ou l'autre modèle est ainsi tout à fait envisageable pour des besoins d'échange ou de partage du corpus, notamment grâce aux transformations par les feuilles de styles XSL. Nous avons en outre également conservé, tout en le simplifiant, l'en-tête (*header*) qu'utilisent la TEI et le CES et qui permet de spécifier, en en-tête d'un texte, toutes les métadonnées le concernant (format, origine, dates, etc.).

5.2.3 Normalisation des données

Le recueil de nos différents textes nous a mise en présence de documents sous des formats variés (des pages Internet, des mails et des documents Word), et leur enregistrement en base de données en a constitué une première normalisation (*cf.* § 5.2.1 p. 81).

Pour pouvoir exploiter aisément ces enregistrements, et notamment les annoter, nous les avons extraits de la base de données afin d'obtenir un document XML sur lequel nous avons travaillé. Le fichier est constitué d'un élément racine `<CORPUS>`, qui lui-même contient les textes. Chaque texte est délimité par les balises `<DOCUMENT> </DOCUMENT>`, avec un identifiant unique comme attribut. Entre ces balises, deux éléments sont insérés :

- les balises `<ENTETE> </ENTETE>` sont destinées à recevoir toutes les informations recueillies lors de la collecte du texte, concernant le scripteur (âge, niveau d'études, langue maternelle et sexe), la situation de rédaction (contrainte temporelle et niveau de formalité) et le fichier (date d'acquisition, date d'annotation, format du support et type de document). Ces informations sont écrites automatiquement dans le fichier XML à partir de la base de données, à l'exception de la date d'annotation qui n'est bien sûr pas encore connue lors de la création du fichier.
  L'en-tête contient également une balise d'informations statistiques sur le texte, telles que le nombre de mots, d'erreurs, de phrases au total et de phrases erronées. Seule la valeur pour le nombre de mots est renseignée au départ, par un calcul automatique effectué directement par le programme. Les autres valeurs seront saisies manuellement lors de l'analyse et l'annotation.
- les balises `<CONTENU> </CONTENU>`, comme leur nom l'indique, délimitent le texte lui-même. Lors de l'ajout d'un document dans le fichier XML, le texte n'étant pas encore annoté, il

est copié en l'état entre les balises. Il fera par la suite l'objet d'un balisage manuel.

```
<CORPUS>
    <DOCUMENT iddoc="2">
        <ENTETE>
            <Scripteur age="21-30" etudes="M1" langue="Oui" sexe="F"/>
            <Texte contrainte="Non" formalite="Non"/>
            <Fichier contexte="résumé" date_acquisition="2009-04-17" date_annotation="" format="doc"/>
            <Stat nb_erreurs="" nb_mots="78" nb_phrases="" nb_phrases_err=""/>
        </ENTETE>
        <CONTENU>
            ce texte traite la problématique de la valence des mots. Le calcul de la valence d'un mot
              permet de calculer la proximité sémantique entre les mots qui ont "aproximativement" un
              même sens. Ce calcule se base sur un calcul mathématique à partir du du nombre d'oocurence
              du mot dans un tableau nous allons calculer le vectuer de ce mot et son cosinus. Si
              l'indice se rapproche de 1 c'est que la valence est forte.
        </CONTENU>
    </DOCUMENT>
    <DOCUMENT iddoc="3">
    ...
    </DOCUMENT>
    ...
</CORPUS>
```

FIGURE 5.1 : Structure du fichier XML

Nous obtenons ainsi un premier fichier, dont la figure 5.1 montre un extrait de la structure, contenant chaque texte de notre corpus qui a été anonymisé, normalisé et sauvegardé dans la base de données. C'est sur ce fichier que nous avons ensuite réalisé nos annotations d'erreurs à l'aide d'un logiciel[15] doté d'une interface WYSIWYG (*What You See Is What You Get*[16]), facilitant l'édition du code XML.

Dans le prochain chapitre, nous présentons les annotations que nous avons effectuées sur les différents écrits collectés pour notre corpus, à savoir les dictées d'étudiants de premier cycle, les résumés d'étudiants de deuxième et troisième cycles, les commentaires de blog d'étudiants en FLE, et les mails de personnes non étudiantes.

15. Morphon XML-Editor, Morphon Technologies, 1998-2003
16. Ce que vous voyez est ce que vous obtenez.

Chapitre 6

Annotation et analyse des erreurs

Sommaire

6.1 Typologies des erreurs et annotation descriptive 89
6.1.1 Adaptation de typologies existantes 89
6.1.2 Balisage du corpus 100
6.1.3 Réajustements de la typologie 103
6.2 Analyse quantitative des erreurs 109
6.2.1 Traitements statistiques des données 109
6.2.2 Description quantitative du corpus 112
6.3 Résumé des principaux résultats 136

6.1 Typologies des erreurs et annotation descriptive

Pour annoter les textes de notre corpus, préalablement normalisés et insérés dans des fichiers XML, nous avons dû déterminer précisément les éléments à considérer dans l'annotation, ainsi bien sûr que la manière de les encoder. Pour les erreurs morphosyntaxiques notamment, point central de notre travail, disposer d'une typologie était indispensable. Nous avons ainsi étudié quelques travaux proposant des typologies d'erreurs, à partir desquelles nous avons défini une première version de notre propre grille d'annotation des erreurs, que nous avons ensuite réajustée au fil du balisage du corpus et des difficultés rencontrées.

6.1.1 Adaptation de typologies existantes

Nous avons déjà évoqué différentes études sur les erreurs d'orthographe et/ou de grammaire (*cf.* § *Les études existantes* p. 54), pour lesquelles nous avons indiqué qu'elles ne tenaient pas compte du type de support (manuscrit ou dactylographié) des écrits. Si notre objectif d'étude des erreurs des tapuscrits nous a conduite à laisser ces travaux de côté dans un premier temps, ils constituent néanmoins des ressources précieuses pour l'élaboration d'une typologie d'erreurs morphosyntaxiques tapuscrites. Certains d'entre eux nous ont ainsi fourni des éléments intéressants à cette fin.

a) Quelques typologies d'erreurs

Parmi les typologies étudiées, nous avons repris à celle de Debyser *et al.* [1967] la distinction entre « faute relative » et « faute absolue », qui correspondent parfaitement, du moins au niveau lexical, à la distinction entre erreur de grammaire et erreur d'orthographe d'un point de vue informatique (*cf.* § 3.2.1 p. 54). Il nous semblait en effet primordial de distinguer dans notre corpus ces deux types d'erreurs, étant donnée l'application informatique que nous visons à plus long terme.

En revanche, l'autre principale distinction réalisée par Debyser *et al.* [1967], à savoir entre « faute graphique » et « faute orale », nous a paru moins fondamentale pour notre annotation d'erreurs, même si nous souhaitions tout de même identifier des erreurs à caractère phonétique, telles que les confusions entre homophones.

La figure 6.1 p. 92 illustre le classement des erreurs proposé par les auteurs, avec des exemples extraits de leur ouvrage, et montre également les correspondances en couleur, avec la typologie à laquelle nous avons abouti. Nous voyons ainsi que nos erreurs orthographiques, en rouge, constituent les fautes absolues au niveau lexical et morphologique. Nous voyons également, dans les erreurs relatives de morphologie, que notre typologie ne fait pas de différence de classement entre le caractère « oral » ou « graphique », et que les deux pourraient être superposés.

La typologie de Catach *et al.* [1980] (voir figure 6.2 p. 93) est fondée sur des critères didactiques. Elle différencie notamment les erreurs directement liées à l'apprentissage de l'écriture et de la lecture, comme les transcriptions erronées de certains sons (dues soit à une mauvaise perception ou prononciation, soit à la méconnaissance de la graphie correspondante), les erreurs issues d'une mauvaise application des règles de grammaire, qui interviennent après l'apprentissage de la lecture, les erreurs induites par la non compréhension d'un texte, par la méconnaissance des diverses graphies d'homophones, etc. L'annotation de notre corpus ne se situant pas dans une approche didactique mais descriptive des erreurs, indépendamment du milieu scolaire, il ne nous semble pas pertinent d'utiliser une typologie qui repose sur des critères didactiques, comme celle-ci. Nous voyons d'ailleurs sur la figure 6.2 qu'environ la moitié des types d'erreurs répertoriés par les auteurs est classée dans notre typologie finale sous deux catégories différentes : `Orthographe` et `Lexique`, ou `Orthographe` et `Ponctuation` (représentées par la double coloration rouge-jaune et rouge-marron). Nous pouvons remarquer également que la typologie est exclusivement orthographique et n'inclut aucune erreur de syntaxe (que nous aurions représentée en vert le cas échéant).

Contrairement aux deux travaux que nous venons de mentionner, Jacquet-Pfau [2001] propose une typologie des erreurs tapuscrites, mais seulement celles détectées par les outils de correction de l'époque. Elle se fonde sur les notions de compétence et de performance de Chomsky [1965] et distingue ainsi les erreurs de compétence du scripteur, ses erreurs de performance, et les erreurs du système (*cf.* § b) p. 56). Si cette classification des erreurs est intéressante pour l'interprétation informatique des erreurs d'écriture (*cf.* § 1.2.2 p. 21), elle est difficile à mettre en œuvre pour l'annotation de notre corpus. Elle nécessiterait en effet soit de connaître la cause de chaque erreur (inattention ou lacune des connaissances), ce qui est difficile dans la plupart des cas à moins d'interroger le scripteur directement, soit d'induire le type de chaque erreur en fonction de différents paramètres (contexte syntaxique, langue maternelle du scripteur, etc.), ce qui relève de l'interprétation, et non plus simplement de la description des erreurs. Or, nous souhaitions réaliser dans un premier temps une annotation principalement descriptive et non interprétative.

Par ailleurs, la figure 6.3 p. 93, dans laquelle nous représentons le classement des erreurs de Jacquet-Pfau [2001], montre très distinctement que seules deux catégories de notre typologie sont représentées : les erreurs d'orthographe principalement, et de lexique (substitution de lemmes plus précisément) dans une moindre mesure. Nous pouvons noter également que les erreurs répertoriées sous l'étiquette « erreurs du système » dans la figure ne sont pas incluses à notre typologie. Il s'agit en effet des formes détectées par le système comme des incohérences car il ne les reconnaît pas (mot inconnu de son lexique), mais ce ne sont en général pas de réelles erreurs.

La typologie des variations orthographiques élaborée par Lucci & Millet [1994], dont nous donnons une représentation avec la figure 6.4 p. 94, repose quant à elle sur une différenciation entre les erreurs « phonographiques » et les erreurs « visuographiques ». Les premières concernent la correspondance entre les sons et les lettres, que les auteurs considèrent comme le principe fondamental du système d'écriture français. Les secondes englobent les graphies distinctives d'homophones (lexicaux et grammaticaux) et celles d'origine étymologique. Encore une fois, cette différenciation ne nous paraît pas adaptée à l'annotation de notre corpus. En effet, c'est la graphie finale des mots erronés qui nous intéresse en premier lieu, indépendamment de l'origine phonographique ou visuographique des erreurs constatées. La figure 6.4 montre d'ailleurs que nous avons effectué un classement différent sur ce point, privilégiant une distinction fondée sur l'existence ou l'inexistence dans la langue de la graphie incriminée (respectivement `Lexique` et `Orthographe`) .

Il ne nous a pas paru non plus pertinent de distinguer les erreurs d'omission, d'ajout, de substitution ou de sélection de caractère comme le font Lucci & Millet [1994]. Une telle différenciation présenterait davantage un intérêt au niveau de la vérification orthographique que de la vérification grammaticale, pour les propositions de correction de mots inconnus par exemple.

Enfin, les erreurs ayant trait à la syntaxe ou à la ponctuation n'apparaissent pas dans cette typologie qui se concentre sur les erreurs de graphie.

La dernière typologie à laquelle nous faisons référence est celle établie par Granger [2007], et de laquelle nous nous sommes le plus largement inspirée. Elle distingue plusieurs domaines d'erreurs dont la forme graphique, la morphologie, la grammaire, le lexique, la syntaxe, le registre, le style, la ponctuation et les coquilles. Certains de ces domaines sont très orientés vers le FLE, puisqu'ils s'appuient sur un corpus d'apprenants dans le cadre d'un projet d'ALAO en FLE [Granger *et al.*, 2001 ; Granger, 2007]. Les erreurs de morphologie typiques des apprenants, par exemple, font ainsi l'objet d'une classification propre qu'il ne nous sera pas nécessaire de conserver en tant que telle, nos textes étant en très grande majorité rédigés par des francophones natifs. Cette différence est visible sur la figure 6.5 p. 94 où les erreurs de morphologie apparaissent en rouge et jaune, signifiant qu'elles sont classées dans notre typologie soit dans la catégorie `Orthographe`, soit dans la catégorie `Lexique`. La classement proposé par Granger [2007] reste toutefois le plus proche de ce que nous recherchions pour décrire les erreurs.

De ces diverses typologies d'erreurs mentionnées, nous ne retenons donc finalement que la distinction entre « faute relative » et « faute absolue » de Debyser *et al.* [1967], ainsi qu'une partie des domaines et catégories d'erreurs établis par Granger [2007].

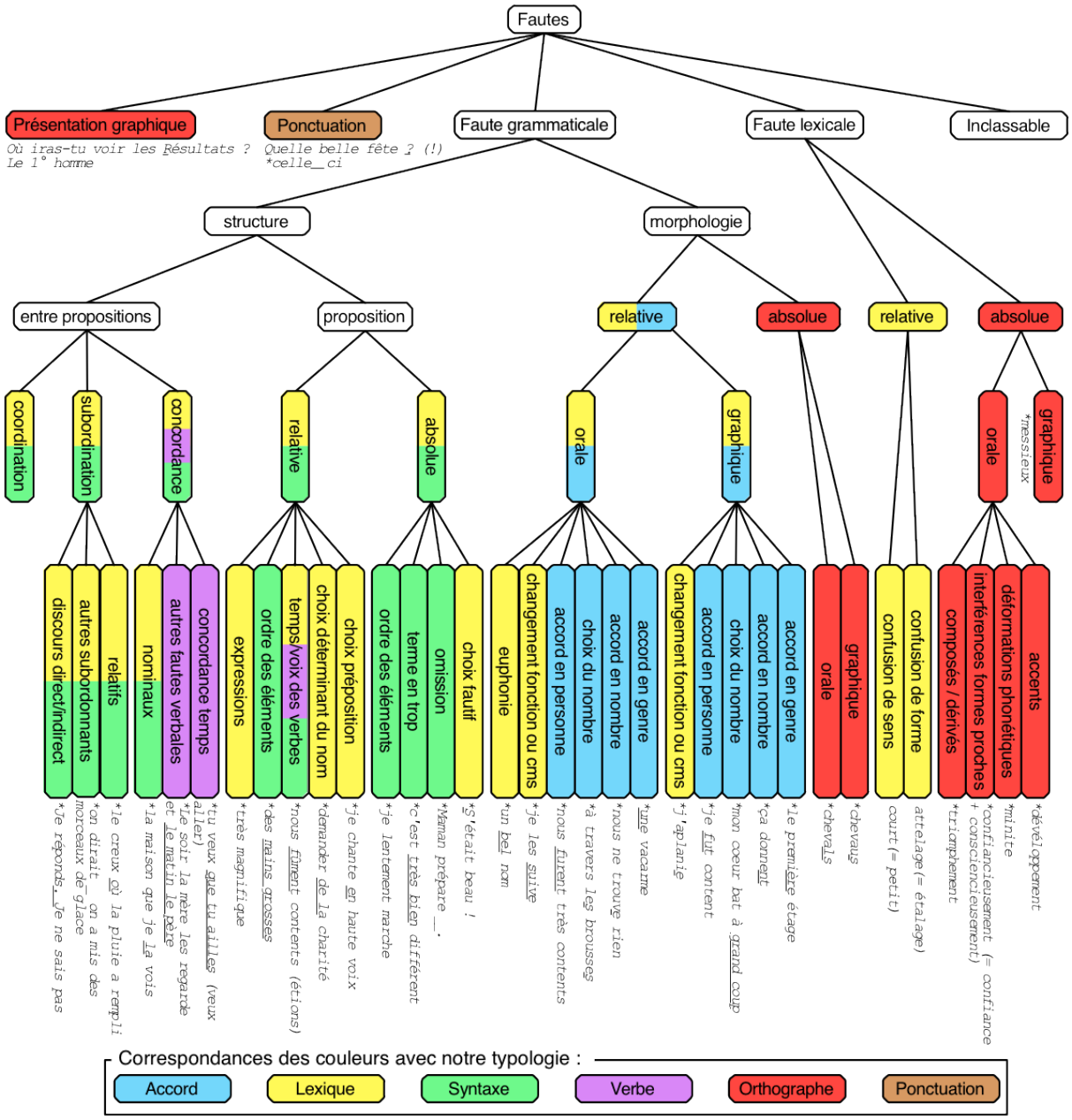


FIGURE 6.1 : Typologie d'erreurs d'après Debyser *et al.* [1967]

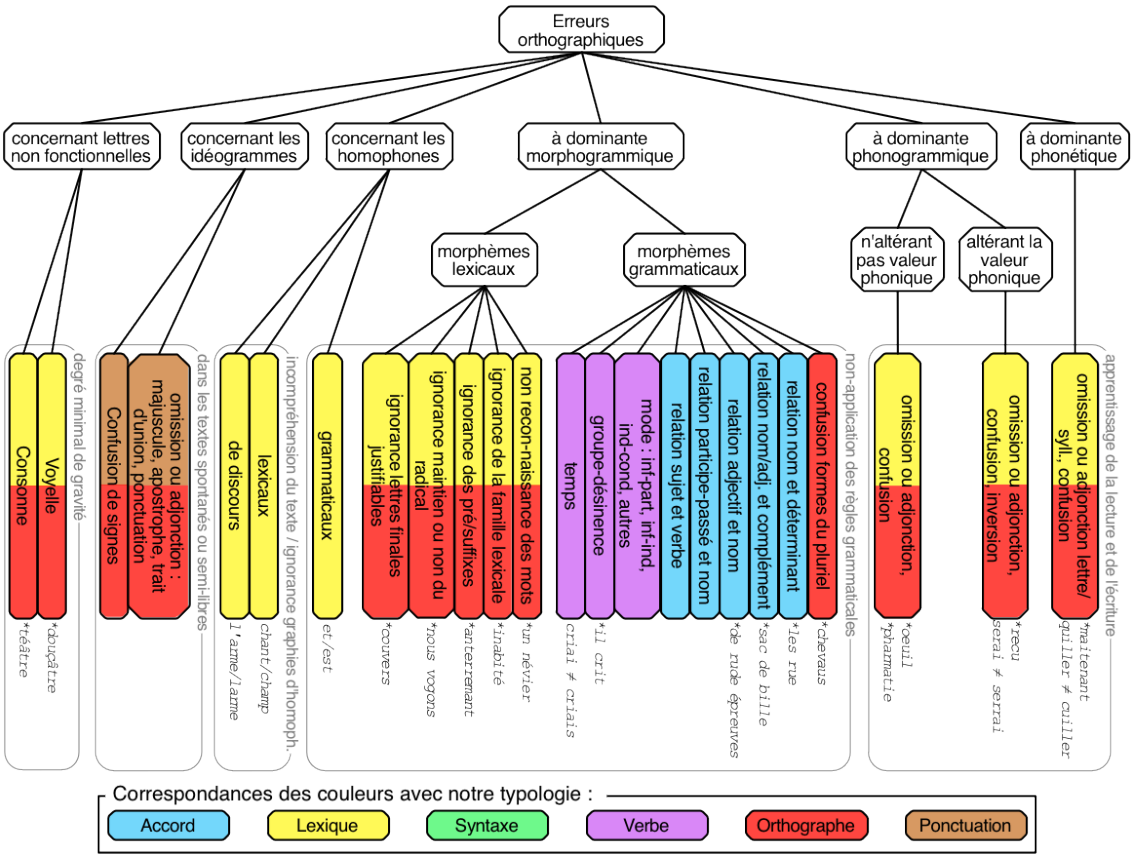


FIGURE 6.2 : Typologie d'erreurs d'après Catach *et al.* [1980]

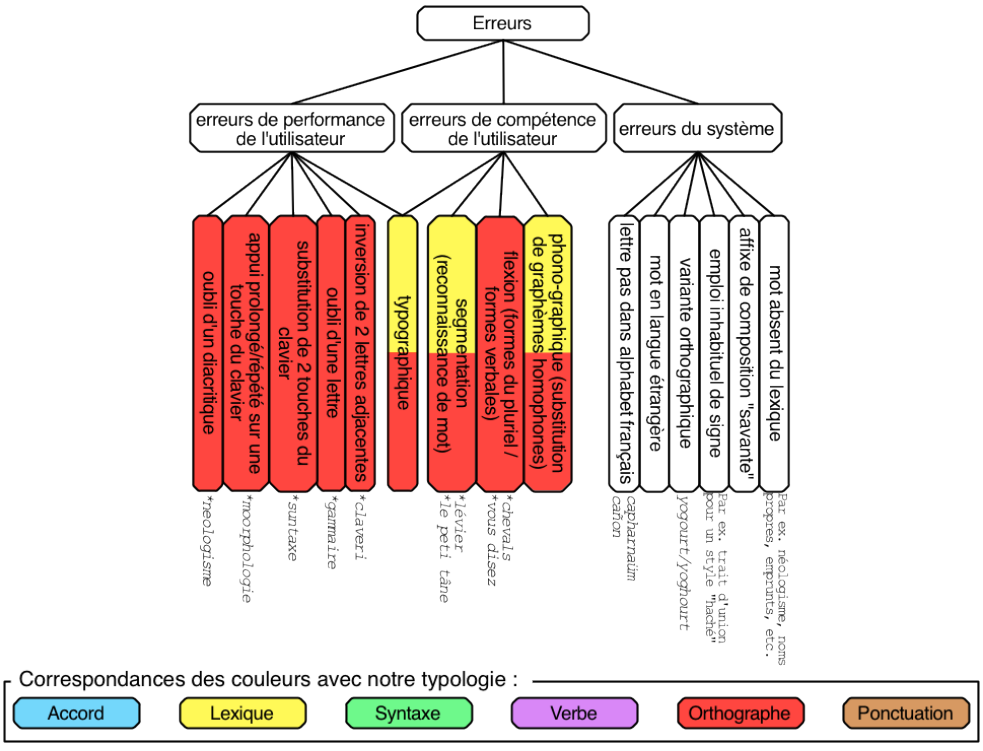


FIGURE 6.3 : Typologie d'erreurs d'après Jacquet-Pfau [2001]

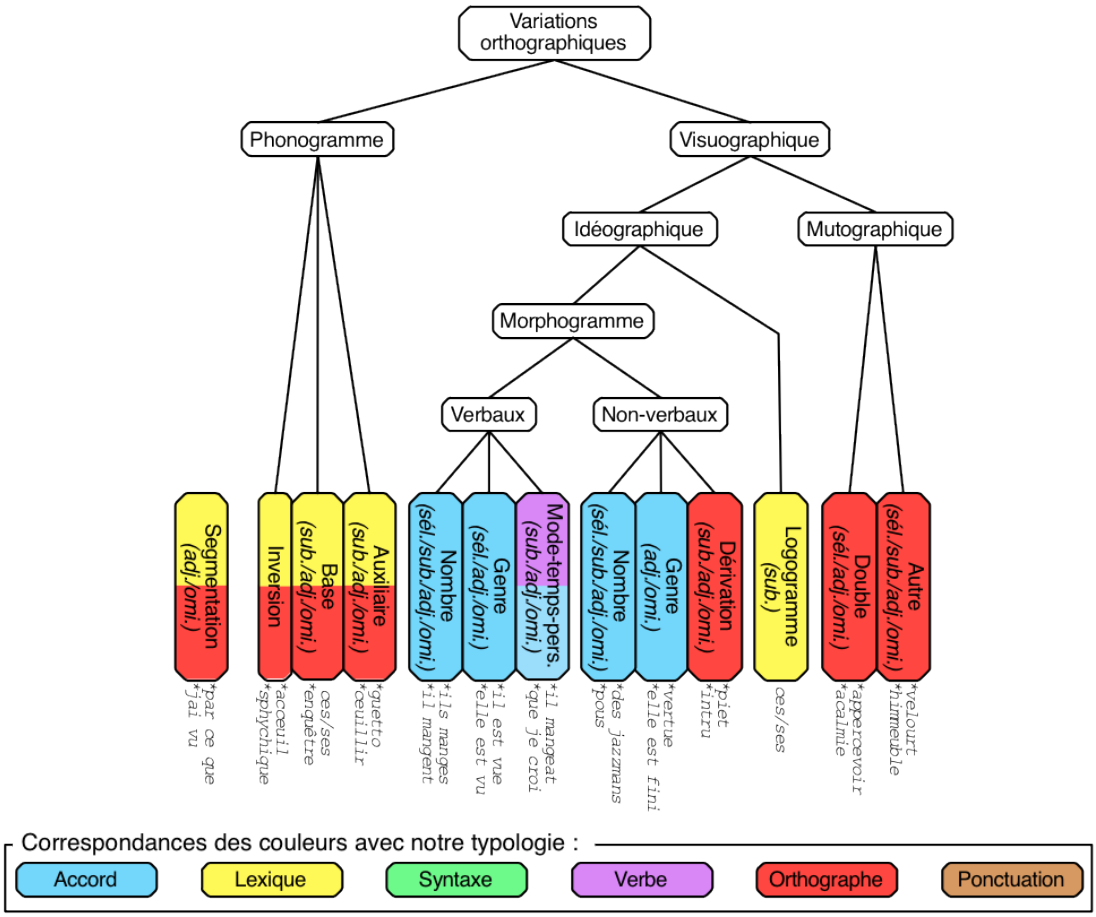


FIGURE 6.4 : Typologie d'erreurs d'après Lucci & Millet [1994]

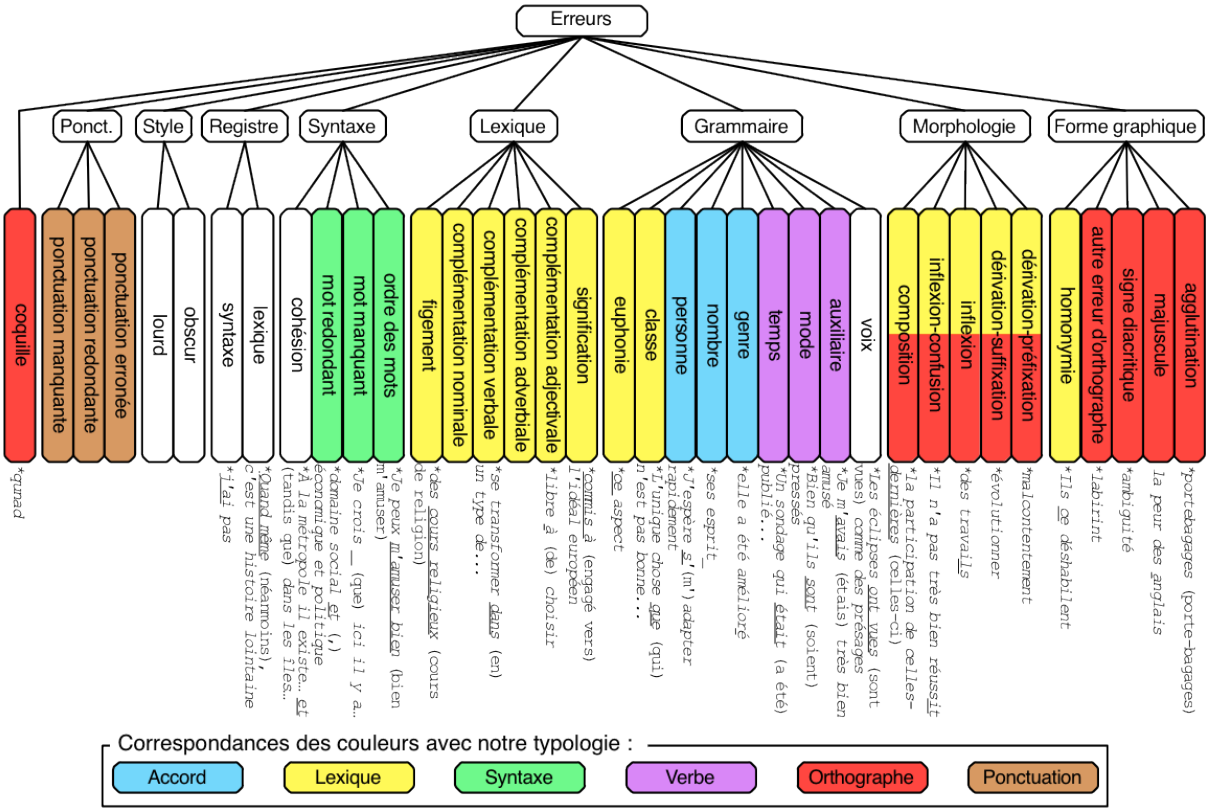


FIGURE 6.5 : Typologie d'erreurs d'après Granger [2007]

b) Définition d'une première typologie *ad hoc*

La première typologie que nous avons définie, à l'aide des travaux existants et en visant à long terme une retombée informatique de notre travail, est divisée en six catégories et 21 sous-catégories que nous synthétisons dans le tableau 6.1 ci-après. Elles sont centrées essentiellement sur les erreurs morphosyntaxiques car ce sont elles que nous souhaitons détecter, dans un premier temps, avec notre modèle de vérification grammaticale. Nous avons également choisi d'effectuer des distinctions autour des types de mots portant les erreurs : les verbes, les lemmes, les différentes catégories de mots concernées par les accords. Ces distinctions nous semblent davantage pertinentes, en vue d'une application informatique, que des distinctions reposant sur des critères didactiques ou phono-/visuo-graphiques.

C'est alors la typologie de Granger [2007] que nous avons reprise et dont nous avons adapté les domaines d'erreurs, afin de regrouper dans une seule et même catégorie les erreurs proprement orthographiques d'un point de vue informatique (graphie inconnue), ce qui correspond aux « fautes absolues » de Debyser *et al.* [1967] (voir figure 6.1 p. 92). Nous avons ainsi défini une catégorie `ORTHO` (pour orthographe) incluant, comme le domaine « Forme graphique » de la typologie de Granger, les erreurs :

- d'agglutination (`agglu`) : pour l'absence d'une espace entre deux mots qui se retrouvent ainsi collés (par ex. **àl'oral*, **undosage*) ;
- de majuscule (`maj`) : lorsqu'elle est manquante en début de phrase ou dans un nom propre, ou au contraire utilisée abusivement (par ex. **au brésil*, **le quotidien D'upsala*) ;
- de graphie (`graph`) : lorsque la graphie d'un mot est erronée, erreurs d'accents incluses, et conduit à un mot inconnu (par ex. **héléphant*, **companies*), à l'exception des erreurs de frappe et de morphologie que nous différencions.

Nous leur avons ajouté les erreurs :

- de frappe (`typo`) : qui se manifestent surtout par des inversions, substitutions, ajouts ou omissions de caractères (par ex. **pendantr*, **dasn*), classées dans « coquilles » chez Granger ;
- de morphologie (`morpho`) : concernant les erreurs de construction de mots d'où résultent des mots inconnus (par ex. **On fesait*, **effets spécials*). Il s'agit d'une partie des erreurs regroupées sous la catégorie « Morphologie » chez Granger. L'autre partie, qui conduit à des formes existantes, fait l'objet d'un classement dans la catégorie `LEXIQUE` de notre typologie.

Dans notre catégorie `LEXIQUE`, nous avons placé les erreurs portant principalement sur la substitution d'un mot par un autre, ne résultant pas d'accord ou de conjugaison erronés. Nous avons ainsi une sous-catégorie pour :

- les homophones (grammaticaux et lexicaux) (`homo`), qui englobe également les paronymes et autres confusions entre formes phonétiquement proches (par ex. *a/à*, *emprunte/empreinte*, *à leurs/alors*) ;
- les erreurs spécifiques d'euphonie (`eupho`), c'est-à-dire la transformation de certains mots en fonction de celui qui suit (par ex. **parce que on*, **cet système*) ;
- la substitution (`substitution`), plus généralement, entre deux mots qui ne se ressemblent pas phonétiquement (par ex. *à/dans*, *de/qui*).

Chez Granger (voir figure 6.5 p. 94), ces trois sous-catégories sont réparties dans trois domaines d'erreurs distincts (« Forme graphique », « Grammaire » et « Lexique »). Il nous a paru pertinent au contraire de les regrouper, dans une perspective de détection automatique des er-

reurs, car elles ont en commun le fait qu'un mot incohérent, grammaticalement par sa catégorie ou sémantiquement, a été utilisé à la place d'un autre.

Les erreurs d'accord ne sont pas prises en compte dans cette catégorie `LEXIQUE`, bien que dans les faits elles conduisent généralement à des homophones. Elles sont d'ailleurs classées par Lucci & Millet [1994] (voir figure 6.4 p. 94) parmi les erreurs visuographiques (qui ne concernent pas la transcription de sons) dans les erreurs idéographiques, qui impliquent des graphies distinctives d'homophones. Nous avons au contraire choisi d'en faire une catégorie propre car, toujours dans une perspective de détection automatique des erreurs, les accords peuvent être pris en charge de manière unique à l'aide de la segmentation en syntagmes, associée à l'unification de traits (*cf.* § **??** p. **??**). Dans cette catégorie `ACCORD`, nous distinguons d'une part, les accords dans les syntagmes nominaux et d'autre part, ceux des syntagmes verbaux. Concernant les nominaux, nous avons ainsi trois sous-catégories pour les erreurs d'accord en genre et en nombre :

- du déterminant avec le reste du syntagme nominal (`SN-D`), par exemple **un position*, **leur subalternes* ;
- du nom avec le reste du syntagme nominal (`SN-N`), par exemple **trois mouvement*, **pendant un ans* ;
- de l'adjectif avec le reste du syntagme (`SN-A`), par exemple **des injections quotidienne*. Nous incluons à cette sous-catégorie les participes passés à valeur adjectivale, non associés à l'utilisation d'un auxiliaire (par ex. **après trois ans passé en cure*). Nous adoptons ici la même position que l'équipe du corpus COVAREC [1994], qui distingue les participes passés employés comme des adjectifs, sans auxiliaire, de ceux qui suivent l'auxiliaire être et que nous classons dans la sous-catégorie `Suj-Ppas` des erreurs d'accords verbaux.

Nous avons quatre autres sous-catégories pour les erreurs d'accords verbaux

- entre le sujet et son verbe (`Suj-V`), accord de nombre et de personne, que le sujet soit un pronom ou un syntagme nominal (par ex. **tu pourra*, **les internautes publies*) ;
- entre le sujet et l'attribut (`Suj-Attr`), en genre et en nombre, s'appliquant aux adjectifs attributs après un verbe d'état (par ex. **beaucoup d'accidents qui sont souvent consécutif*) ;
- entre le sujet et le participe passé (`Suj-Ppas`), en genre et en nombre, quel que soit l'auxiliaire employé (par ex. **une variante est associé*, **les américains ont déjà choisis*) ;
- entre l'objet et le participe passé (`Obj-Ppas`), en genre et en nombre pour tenir compte notamment des accords entre le participe passé et l'objet du verbe lorsque celui-ci se trouve avant l'auxiliaire avoir (par ex. **la solution que j'ai utilisé*).

Cette classification correspond à une partie du domaine d'erreurs « Grammaire » de Granger (en bleu sur la figure 6.5 p. 94), mais dans une approche différente. En effet, alors que Granger classe les erreurs en question selon le trait morphosyntaxique altéré (genre, nombre ou personne), nous les classons en fonction du rôle syntaxique ou de la catégorie morphosyntaxique (nom, verbe, sujet, etc.) du mot mal accordé. Ce choix découle de la volonté de pouvoir distinguer d'une part les erreurs d'accord à caractère nominal et verbal, et d'autre part, au sein de ces erreurs, le type de mot concerné. Un tel étiquetage nous apparaît en effet davantage pertinent pour nos analyses ultérieures des erreurs, par rapport à un étiquetage des traits morphosyntaxiques erronés, puisque la méthode de détection par l'unification de traits que nous envisageons pour ces erreurs repose sur la catégorie ou la fonction des mots ou syntagmes.

Le domaine « Grammaire » de Granger inclut, en plus des erreurs de personne et de nombre que nous avons déjà intégrées sous la forme sujet-verbe dans notre catégorie `ACCORD`, d'autres erreurs spécifiques aux verbes (en violet sur la figure 6.5 p. 94). Nous les avons réunies dans une

catégorie `VERBE`, qui est donc composée des sous-catégories suivantes :

- temps (`temps`), lorsque le temps employé n'est pas celui attendu, entraînant une mauvaise concordance des temps par exemple (par ex. *`On fesait la bataille de boules de neige dehors pendant environ 20 minutes.`) ;
- mode (`mode`), qui concerne tous les cas où un mode inadapté est utilisé. Le plus souvent, il s'agit de confusions entre infinitifs, participes passés et formes conjuguées (par ex. *`pour le dompté`, *`La municipalité qui présentée donc cet arrété`) ;
- auxiliaire (`aux`), pour l'utilisation de « avoir » à la place de « être », et *vice versa* (par ex. *`pour explique qu une personne est des pb personnel`) .

Nous n'avons pas conservé la catégorie « voix » également proposée par Granger (ex. « Les éclipses ont vues (sont vues) comme des présages » [Granger, 2007]) car nous l'assimilons davantage à une erreur d'auxiliaire.

Nous avons aussi conçu une catégorie `SYNTAXE`, fondée sur le domaine du même nom chez Granger (voir figure 6.5 p. 94), duquel nous avons seulement exclu un type d'erreurs. Notre catégorie contient ainsi trois sous-catégories :

- les oublis (`oubli`), pour tous les mots omis, comme par exemple la particule « ne » de la négation (par ex. *`c'est pas grave`) ;
- les répétitions (`repet`), quand un mot est doublé (par ex. *`à partir du du nombre`) ;
- l'ordre des mots (`ordre`)(par ex. *`On tous était super heureus`).

Nous avons exclu la catégorie « cohésion [1] » de Granger, car nous l'apparentons plutôt au mauvais choix d'un mot (erreur de lexique), voire à une question de style. Or, nous préférons laisser les erreurs de style de côté dans notre annotation, si tant est que nous puissions parler d'erreurs dans ce cas. En effet, l'appréciation du style est variable d'un individu à l'autre, et un vérificateur linguistique peut difficilement intervenir sur ce point. C'est la raison pour laquelle nous n'avons pas non plus repris le domaine « Style » de Granger, ni celui de « Registre ».

Enfin, nous avons ajouté à notre typologie la catégorie `PONCT`, présente aussi chez Granger (« Ponctuation »), mais nous ne l'avons pas détaillée, les questions de ponctuation ne nous paraissant pas essentielles pour notre étude.

1. Exemple tiré de [Granger, 2007] : « À la métropole, il existe plus d'allocations et d'aide pour les chômeurs, les handicapés et les personnes âgées, *et* (tandis que) dans les îles, il n'y a pas beaucoup de soutien. »

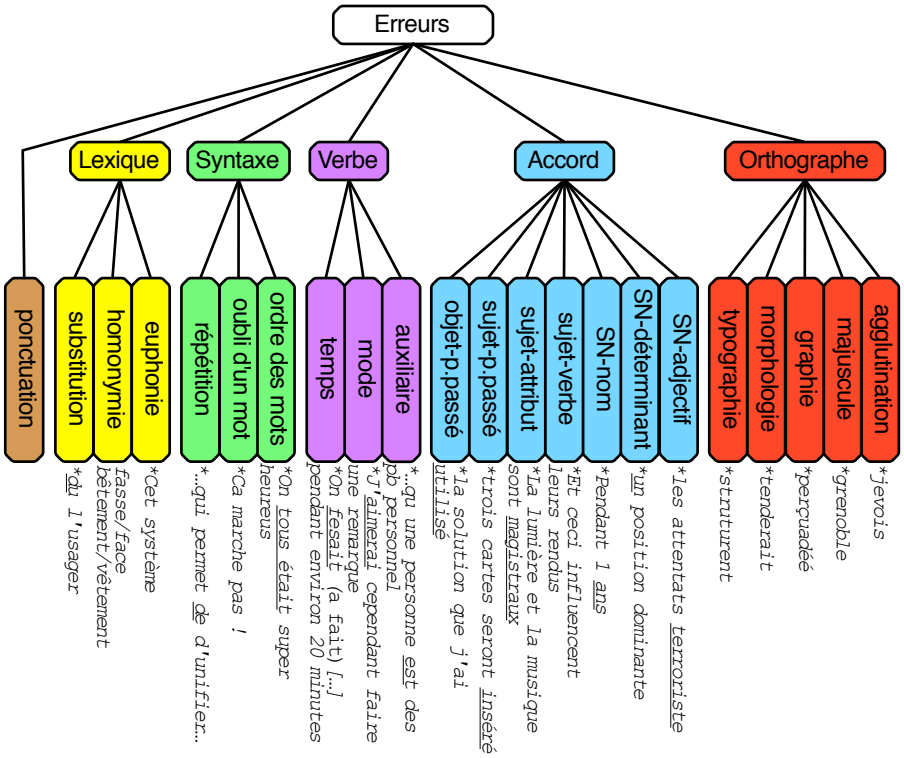


FIGURE 6.6 : Synthèse de la première typologie d'erreurs

Catégorie	**Sous-cat.**	**Description**	**Exemple**
ORTHO	agglu	Oubli d'une espace qui cause l'agglutination de 2 mots	**jevois*
	graph	Erreur de graphie, sauf typographique ou morphologique	**perçuadéé*
	maj	Erreur de majuscule	**grenoble*
	morpho	Erreur de construction morphologique	**tenderait*
	typo	Erreur de frappe	**struturent*
ACCORD	SN-A	Erreur de genre/nombre de l'adjectif dans un SN ou du participe passé employé comme adjectif	**les attentas terroriste*
	SN-D	Erreur de genre/nombre du déterminant	**un position dominante*
	SN-N	Erreur de genre/nombre du nom	**Pendant 1 ans*
	Suj-V	Erreur d'accord du verbe avec son sujet	**Et ceci influencent leurs rendus*
	Suj-Attr	Erreur d'accord de l'adjectif attribut avec son sujet	**La lumière et la musique sont magistraux*
	Suj-Ppas	Erreur d'accord du participe passé avec le sujet	**aucun sondage similaire n'a été réalisée*
	Obj-Ppas	Erreur d'accord du participe passé avec l'objet antéposé	**la solution que j'ai utilisé*
VERBE	aux	Erreur de choix de l'auxiliaire	**...qu une personne est des pb personnel*
	mode	Erreur de mode	**J'aimerai cependant faire une remarque*
	temps	Erreur de temps	**On fesait* (a fait)
SYNTAXE	ordre	Mots mal ordonnés	**On tous était super heureus...*
	oubli	Oubli d'un mot	**Ca marche pas !*
	repet	Répétition d'un même mot	**...qui permet de d'unifier ...*
LEXIQUE	eupho	Euphonie non respectée	**Cet système...*
	homo	Confusion entre 2 homophones ou mots très proches phonétiquement	*fasse/face*, *bêtement/vêtement*
	subs	Mot inadapté au contexte	**du l'usager*
PONCT		Erreur de ponctuation (signe, espace)	**Pourquoi pas?*

Tableau 6.1 : Première typologie d'erreurs

6.1.2 Balisage du corpus

La première typologie que nous avons élaborée nous a permis de commencer à annoter nos textes, dans les fichiers que nous avons créés précédemment en utilisant le langage de description XML (*cf.* § 5.2.3 p. 86), dont nous avons présenté les avantages dans le chapitre précédent (*cf.* § 5.2.2 p. 83). La figure 5.1 p. 87 par exemple montre la structure générale de chaque document, et notamment de l'entête. Nous nous consacrons à présent au balisage du texte lui-même, qui se trouve à l'intérieur des balises `<CONTENU></CONTENU>`.

a) Éléments et attributs utilisés

Nous avons commencé par segmenter les textes en phrases. Chacune d'elle a ainsi été délimitée par des balises `<Phrase></Phrase>` et a reçu un numéro unique pour l'identifier, sous la forme d'un attribut (`idp`). Par exemple :

```
<Phrase idp="ID694">Il te faut simplement sélectionner toute ta cellule.</Phrase>
```

Nous avons ensuite recherché les erreurs au sein de chaque phrase. Pour les baliser, nous nous sommes inspirée du modèle d'annotation du corpus FRIDA-bis [Antoniadis *et al.*, 2010], dérivé du corpus d'apprenants FRIDA que nous avons brièvement présenté dans un chapitre précédent (*cf.* § 4.1.2 p. 67). La figure 6.7 ci-après, tirée d'Antoniadis *et al.* [2010], présente l'annotation d'une erreur contenue dans le corpus FRIDA-bis. Nous y voyons deux balises `<tok></tok>` entourant l'une la forme erronée et l'autre la version corrigée, et contenant des attributs permettant de caractériser ces formes (lemme, catégorie et traits morphosyntaxiques). Chacune de ses deux balises `<tok></tok>` se trouve respectivement dans la balise `<ini></ini>` pour la forme initiale erronée, et `<cor></cor>` pour la forme corrigée. Le tout est entouré de la balise `<err></err>`, qui délimite donc l'erreur, et qui possède plusieurs attributs pour la spécifier : dans l'ordre sont indiqués le numéro de l'erreur (`ide`), son domaine (`dom1`) et sa catégorie (`cer1`), et enfin sa catégorie grammaticale (`cgr1`). Hormis l'identifiant numérique, les autres attributs prennent pour valeurs les données définies dans la typologie d'erreurs de Granger [2007].

```
<err ide="2" dom1="G" cer1="NBR" cgr1="ADJ">
    <ini>
        <tok lemme="public" cat="adj" tags="fem sing">publique</tok>
    </ini>
    <cor>
        <tok lemme="public" cat="adj" tags="fem plu">publiques</tok>
    </cor>
</err>
```

FIGURE 6.7 : Exemple d'annotation de FRIDA-bis [Antoniadis *et al.*, 2010]

De ce modèle d'annotation, nous avons repris les balises `<err></err>`, `<ini></ini>` et `<cor></cor>` pour délimiter l'erreur, la forme initiale erronée et la forme corrigée, et nous les avons simplement renommées respectivement `<ERREUR></ERREUR>`, `<Initial></Initial>` et `<Correction></Correction>`. Afin de décrire l'erreur ainsi délimitée, chaque balise `<ERREUR>` est complétée, comme pour le corpus FRIDA-bis, par plusieurs attributs. Les deux premiers que nous présentons sont proches de ceux utilisés dans FRIDA-bis, en revanche les suivants nous sont propres.

1. `ide` définit un identifiant unique pour chaque erreur (comme pour les phrases) ;

2. `type` spécifie le type de l'erreur, conformément à la typologie que nous avons élaborée ;

3. `att` permet d'indiquer qu'elle forme était attendue à la place de la forme erronée (par exemple un singulier au lieu d'un pluriel)

4. `idref` est un attribut particulier qui renvoie au « référent » de la forme erronée, référent en ce qu'il constitue l'entité à laquelle nous nous référons pour détecter l'incohérence grammaticale et la justifier, et non pas référent au sens linguistique du terme qui désigne les éléments de la réalité représentés par les signes linguistiques.

   Dans le cas d'une erreur d'accord sur un adjectif par exemple, le nom dont il dépend sera le référent, de même pour le groupe nominal antécédent d'un pronom anaphorique de genre erroné.

   L'attribut `idref` reprend donc le numéro d'identifiant associé au référent dans l'étiquetage du texte. Si le référent constitue lui-même une erreur, il s'agira d'un identifiant d'erreur (`ide`). Sinon, il sera lui-même délimité par les balises `<REF></REF>` et possèdera son propre identifiant, qui sera repris par l'attribut `idref` de l'erreur.

Un même mot peut contenir plusieurs types d'erreurs simultanément. Dans ce cas, elles sont toutes décrites dans un seul et même élément `<ERREUR>`. Afin de différencier les attributs propres à chaque type d'erreurs, nous rajoutons un chiffre, de 1 à 4 (dans de rares cas nous avons pu associer trois ou quatre types d'erreurs différents à une même forme), à chacun des attributs que nous venons d'énumérer. Les attributs « 1 » (par ex. `ide1`, `type1`, etc.) font ainsi référence à un premier type, « 2 » à un deuxième (par ex. `ide2`, `type2`, etc.), etc.

Dans la figure 6.8 nous montrons un exemple de document annoté, composé de cinq phrases, dont deux contiennent chacune deux erreurs, avec annotation des référents. Dans cet exemple, chaque erreur n'appartenant qu'à une seule catégorie, tous ses attributs sont suivis du chiffre 1.

```
<DOCUMENT iddoc="225">
    <ENTETE>
        <Scripteur age="15-20" etudes="Ukn" langue="Non" sexe="M"/>
        <Texte contrainte="Non" formalite="Non"/>
        <Fichier contexte="liste de diffusion users OOo" date_acquisition="2009-04-15" date_annotation="2009-04-27" format="mail"/>
        <Stat nb_erreurs="4" nb_mots="58" nb_phrases="5" nb_phrases_err="2"/>
    </ENTETE>
    <CONTENU>
        <Phrase idp="ID693">mNOM, </Phrase>
        <Phrase idp="ID694">Il te faut simplement sélectionner toute ta cellule. </Phrase>
        <Phrase idp="ID2466">
            <REF idref="ID870">Tu</REF>
            <ERREUR att1="V2s" ide1="ID869" idref1="ID870" type1="ACC-Suj-V">
                <Initial>clique</Initial>
                <Correction>cliques</Correction>
            </ERREUR> dans ta cellule en laissant
            <REF idref="ID868">le bouton</REF> de la souris
            <ERREUR att1="Ams" ide1="ID867" idref1="ID868" type1="ACC-SN-A">
                <Initial>enfoncée</Initial>
                <Correction>enfoncé</Correction>
            </ERREUR>.
        </Phrase>
        <Phrase idp="ID2467">Puis
            <REF idref="ID866">tu</REF> te
            <ERREUR att1="V2s" ide1="ID865" idref1="ID866" type1="ACC-Suj-V">
                <Initial>déplace</Initial>
                <Correction>déplaces</Correction>
            </ERREUR> en dehors de la cellule et tu y reviens et
            <REF idref="ID864">tu</REF>
            <ERREUR att1="V2s" ide1="ID863" idref1="ID864" type1="ACC-Suj-V">
                <Initial>relâche</Initial>
                <Correction>relâches</Correction>
            </ERREUR> le bouton de la souris.
        </Phrase>
        <Phrase idp="ID695">De cette manière le remplissage qui suit sera sur tout le fond de ta cellule. </Phrase>
    </CONTENU>
</DOCUMENT>
```

FIGURE 6.8 : Exemple de balisage d'un texte

b) Difficultés rencontrées

Le balisage du corpus nous a conduite à quelques questionnements pour lesquels nous avons dû prendre des décisions. Par exemple, dans certains cas, il était impossible de déterminer la correction à indiquer pour une erreur, en particulier pour les mots manquants. Nous avons alors défini la notation « `?UKN?` » à insérer entre les balises `<Correction></Correction>` dans ces cas précisément où la correction était inconnue (ex. 7). En complément, pour certaines situations où le mot manquant était sans équivoque un nom propre (concerne les dictées, pour lesquelles nous avions le texte original pour comparaison), nous avons défini la notation « `?NP?` » sur le modèle de « `?UKN?` » (ex. 8).

(7) *`la frequence des probleme de couple chez les homme sporteur de l.... a ete 2 fois plus eleve que chez les utre hommes.`*

(8) *`Aujourd"hui, désintoxiqué, Big Birhter rejoindra samedi un parc national à __, captiale du Yonan`*

Le traitement des noms propres a fait l'objet d'autres hésitations. Dans un premier temps, nous avons balisé comme erreur ceux que nous savions mal orthographiés, c'est-à-dire en général des noms propres très répandus (par ex. `Washington`). Nous avons parallèlement ignoré ceux qui sont peu ou pas connus (par ex. `capitale du Yunnan`), au motif qu'il peut être difficile de savoir s'ils sont écrits correctement. Cette différence d'annotation n'était cependant pas pertinente. Elle correspondait en quelque sorte à ce qui existe au niveau des vérificateurs linguistiques, qui intègrent désormais souvent un dictionnaire des noms propres les plus répandus. Mais leur liste varie d'un outil à l'autre, et la nôtre, pour déterminer si un nom propre devait donc être considéré comme connu ou pas, se fondait sur notre propre connaissance du monde. Nous avons donc décidé, pour plus de cohérence, d'ignorer les erreurs d'orthographe de tous les noms propres dans notre balisage, à l'exception des erreurs de majuscules manquantes.

Nous avons également parfois hésité, lors de la détermination du référent de certaines erreurs, entre la solution strictement grammaticale et celle qui nous semblait plus pertinente. L'énoncé (9) ci-après en est un exemple :

(9) *`elle a attéri dans le entre de tri bagage où le personnel de l'aéroport d'Arlanda la réceptionné`*

Il contient plusieurs erreurs, dont deux seulement nous intéressent ici. La première concerne la fusion du pronom objet `l'` avec l'auxiliaire auquel il est antéposé, pour donner la forme `la` au lieu de la forme homophone `l'a`. La seconde porte sur le participe passé `réceptionné` dont l'accord avec son antécédent est erroné. De manière stricte, son antécédent est le pronom objet que nous venons de mentionner et c'est avec lui qu'il s'accorde. C'est donc lui que nous devrions désigner comme le référent. Cependant, il ne permet pas directement de connaître le genre pour déterminer l'accord à effectuer. Il faut se reporter au début de la phrase pour retrouver l'élément dont il est l'anaphore et qui est marqué en genre, à savoir le pronom personnel sujet `elle`. C'est l'antécédent le plus proche du pronom objet `l'` qui permet de savoir que c'est un féminin, et donc que le participe passé doit également être au féminin. Pour cette raison, il nous a semblé plus pertinent que le rôle de référent lui revienne pour l'erreur sur `réceptionné`, plutôt qu'au pronom objet `l'`.

Enfin, c'est au niveau de la typologie elle-même que nous avons eu le plus d'hésitations, qui nous ont conduite à divers remaniements.

6.1.3 Réajustements de la typologie

Si la version initiale de notre typologie nous a permis d'annoter un certain nombre de textes, de nombreuses difficultés et questions se sont présentées au fur et à mesure de notre progression dans l'étiquetage et ont nécessité des adaptations dans les différentes catégories.

Catégorie `ORTHO`

Concernant la catégorie `ORTHO` (*cf.* figure 6.9), nous avons constaté la présence dans le corpus d'erreurs de segmentation autres que celles d'agglutination de deux mots, déjà prises en compte dans notre typologie par la sous-catégorie `agglu`. Il s'agit d'erreurs de segmentation au mauvais endroit (par ex. **pou rêtre*). Nous les avons donc incluses à la sous-catégorie `agglu` que nous avons alors renommée plus justement `segm`, pour segmentation.

Nous avons également ajouté à la catégorie `ORTHO` une sous-catégorie pour les erreurs d'accents. Elle existait chez Granger [2007] dont nous nous sommes inspirée, mais nous ne l'avions pas intégrée à notre typologie initiale, pensant étiqueter ces erreurs comme des erreurs de graphie. Cependant, l'annotation de certains textes dans lesquels tout accent avait été délibérément omis (de même que les apostrophes), nous a incitée à finalement distinguer les erreurs d'accents des autres erreurs de graphie. Nous avons donc créé une sous-catégorie `dia` prenant en compte de manière plus générale tous les diacritiques, et au sein de laquelle nous avons également annoté les erreurs sur les apostrophes (des omissions le plus souvent).

Enfin, nous avons complété cette catégorie `ORTHO` avec une sous-catégorie `abrev` pour les mots qui ont été abrégés lors de la rédaction (par ex. **tjrs*), et nous avons finalement supprimé la sous-catégorie `typo` pour les fautes de frappe, que nous avons englobée dans les erreurs de graphie `graph`. En effet, pour classer une erreur d'orthographe dans l'une ou l'autre des deux sous-catégories, il aurait fallu sortir de la simple description pour essayer de deviner l'origine probable de l'erreur, à savoir une lacune du scripteur ou une faute de frappe.

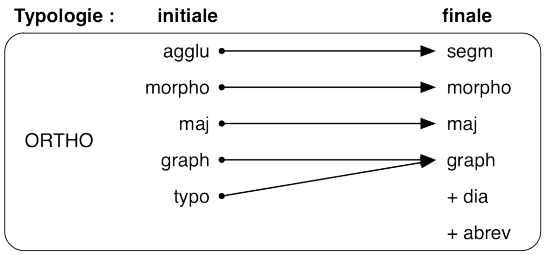


FIGURE 6.9 : Évolutions de la catégorie `ORTHO`

Catégorie `SYNTAXE`

La seule modification que nous avons apportée à la catégorie `SYNTAXE` (*cf.* figure 6.10) concerne la sous-catégorie `repet` que nous avons renommée en `ajout`, de manière à y inclure plus naturellement les cas où un mot est en trop dans la phrase, sans toutefois être un doublon ou une répétition, ce qui n'était pas prévu dans la typologie initiale. La sous-catégorie ainsi rebaptisée regroupe désormais tous les ajouts d'un mot superflu dans la phrase.

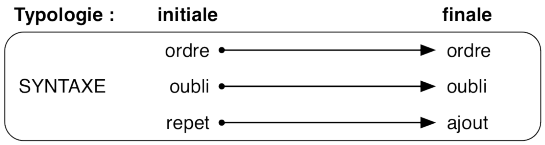


FIGURE 6.10 : Évolutions de la catégorie `SYNTAXE`

Catégorie `ACCORD`

Dans la catégorie `ACCORD` (*cf.* figure 6.11), alors que nous avions initialement créé une seule sous-catégorie pour les erreurs d'accord des participes passés après un auxiliaire (`Suj-Ppas`), nous avons rapidement choisi de la scinder en deux sous-catégories distinctes pour d'une part, les participes passés après l'auxiliaire être, et d'autre part, ceux après l'auxiliaire avoir, les comportements étant différents pour l'un et l'autre au niveau des accords. Nous avons également simplifié les différents noms.

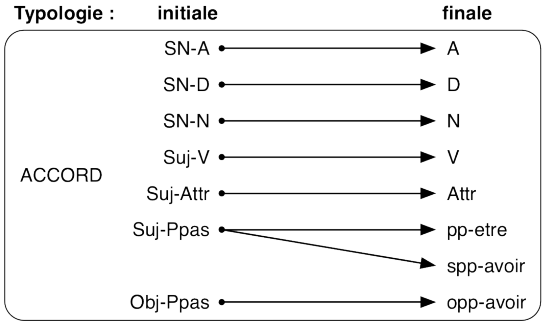


FIGURE 6.11 : Évolutions de la catégorie `ACCORD`

Catégorie `VERBE`

Nous avons retiré dans la catégorie `VERBE` (*cf.* figure 6.12 p. 105) la sous-catégorie `aux` relative aux erreurs d'auxiliaires. En effet, nous les avons finalement étiquetés comme des substitutions de mots, donc dans la catégorie `LEXIQUE`.

Catégorie `LEXIQUE`

La catégorie `LEXIQUE` (*cf.* figure 6.12 p. 105) est celle qui s'est vue la plus modifiée. Dans un premier temps, afin de distinguer les erreurs entre vrais homophones de celles entre mots plus ou moins ressemblants, nous avons ajouté une sous-catégorie `confu`. Dédiée à la confusion entre mots proches phonétiquement, elle est venue en complément de `homo` que nous avons alors restreint aux homophones stricts.

Nous avons également créé une sous-catégorie `typo` pour les erreurs typographiques. La catégorie `ORTHO` disposait déjà d'une sous-catégorie du même nom, mais nous avons constaté à plusieurs reprises que des erreurs s'apparentant manifestement à des fautes de frappe conduisaient à des mots existants, et ne pouvaient donc pas rentrer dans la catégorie `ORTHO`. La sous-catégorie `typo` dans `LEXIQUE` avait pour objectif l'annotation d'erreurs de frappe créant un autre mot que celui attendu. Nous n'avons cependant pas conservé cette sous-catégorie pendant longtemps. En

effet, en même temps que nous avons supprimé celle de `ORTHO`, et pour la même raison, nous avons éliminé cette sous-catégorie `typo` de la catégorie `LEXIQUE`. Plus exactement, nous avons effectué une fusion entre les sous-catégories `subs` (substitution d'un mot par un autre non ressemblant phonétiquement), `confu` (substitution d'un mot par un autre proche phonétiquement) et `typo` (substitution d'un mot par un autre à cause d'une faute de frappe). Toutes les trois concernent en effet une erreur de substitution d'un mot par un autre, c'est-à-dire d'utilisation d'un mauvais lemme, la distinction entre les trois reposant davantage sur de l'interprétation que sur de la description. Nous avons donc créé une sous-catégorie `lemme` pour les regrouper.

Nous n'avons pas inclus la sous-catégorie `homo` à ce regroupement, bien qu'elle concerne également souvent la substitution d'un lemme par un autre. Tout d'abord, l'observation des erreurs au fur et à mesure de l'annotation nous a fait prendre conscience de la très grande proportion des erreurs finalement concernées par la question de l'homophonie, comme les erreurs d'accord par exemple. L'ajout ou l'oubli du « -s » du pluriel, du « -e » du féminin dans une moindre mesure, la confusion dans les désinences verbales, etc. conduisent très fréquemment à un mot homophone, comme le laisse voir la typologie de Lucci & Millet [1994] (voir figure 6.4 p. 94), dont la branche la plus complète concerne les erreurs idéographiques, c'est-à-dire les graphies distinctives d'homophones. Pour rester cohérente, iI aurait fallu que nous étiquetions chaque erreur concernée avec à la fois son type principal et le type `homo` pour l'homophonie, ce qui aurait surchargé les annotations. Ensuite, dans les situations que nous venons de citer, à savoir les accords ou les affixes verbaux, il n'y a généralement pas de substitution de lemmes. Il n'était donc pas pertinent d'ajouter la sous-catégorie `homo` à celle que nous venions de former, `lemme`, puisque les erreurs à caractère homophone n'impliquent pas systématiquement un changement de lemme. Toutefois, comme cette caractéristique particulière éventuelle des erreurs nous semblait intéressante à observer, et donc à annoter, nous avons finalement choisi de définir un nouvel attribut pour les balises `<ERREUR>`, auquel nous avons donné la valeur `homo="oui"` pour les cas d'homophonie. Chaque erreur dont la forme erronée est homophone avec la forme correcte dispose ainsi d'un trait spécifique pour cette propriété, renseigné dans la balise. Nous avons bien sûr conjointement retiré de la catégorie `LEXIQUE` la sous-catégorie `homo` qui n'avait plus lieu d'être.

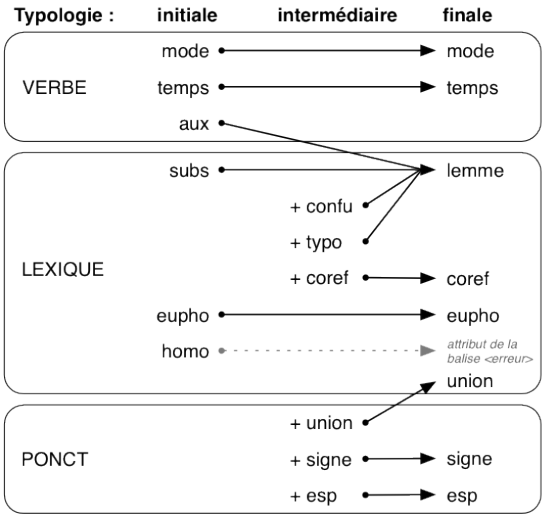


FIGURE 6.12 : Évolutions des catégories `VERBE`, `LEXIQUE` et `PONCTUATION`

Cette modification a également apporté une solution au problème d'annotation qui se posait à nous, concernant par exemple la confusion entre les formes infinitives en « -er » et les participes passés en « -é », ou d'autres du même type, que nous avons évoquées dans le précédent paragraphe. Il nous semble que ce genre d'erreur est d'une part assez fréquent, et d'autre part induit par l'homophonie. Bien qu'il s'agisse d'erreurs de mode verbal, nous tenions à ce que le paramètre de l'homophonie apparaisse également dans leur étiquetage, ce que la création de l'attribut `homo` a rendu plus facilement réalisable.

Pour finir, nous avons ajouté à la catégorie `LEXIQUE` une sous-catégorie `coref` pour l'étiquetage des erreurs portant sur un problème de coréférence. Nous avons trouvé peu d'erreurs de ce type, et elles consistent à chaque fois en un choix d'un pronom ne gardant pas le genre et/ou le nombre de l'antécédent qu'il reprend (par ex. **[...]toute question est la bienvenue, quelle que soit son fond ou sa forme*).

Catégorie `PONCTUATION`

(voir figure 6.12 p. 105) Au vu des erreurs de ponctuation que nous avons rencontrées dans les textes, contrairement à notre choix initial, nous avons finalement décidé de détailler la catégorie `PONCTUATION`. Nous avons d'abord distingué trois sous-catégories :

- `esp` rassemble les erreurs d'espace redondante ou manquante avant ou après un autre signe de ponctuation. Nous avons remarqué en effet qu'il est fréquent que les règles typographiques (Imprimerie Nationale [2002]) les concernant ne soient pas suivies. La plupart du temps, ce type d'erreurs est sans conséquence sur la structure et la grammaticalité de la phrase.
  Nous les distinguons en revanche des erreurs d'espaces (appelés aussi « blancs ») entre mots, que nous avons classées dans la catégorie `ORTHO`, assimilées à une segmentation erronée, et qui conduisent généralement à un ou plusieurs mots inconnus.
  Nous distinguons également ces erreurs de celles affectant les signes de ponctuation eux-mêmes, qui peuvent avoir un impact sur le sens ou la grammaire, et qui sont plus problématiques en matière de vérification linguistique.
- `signe` concerne les signes de ponctuation manquants, superflus ou erronés, en dehors du trait d'union qui fait l'objet d'un classement à part, et de l'espace dont nous venons de parler. Il pourra s'agir par exemple d'un point se substituant à un point d'interrogation, d'une virgule mal placée, d'une parenthèse non refermée, etc.
- `union` enfin rassemble les cas d'ajout ou d'omission d'un trait d'union, dans les mots composés ou les composés grammaticaux (par ex. **existe t il*). Nous avons initialement classé ces erreurs avec les erreurs de ponctuation, comme le font Debyser *et al.* [1967], mais nous avons rapidement pris conscience du fait qu'il serait plus approprié de considérer le trait d'union davantage comme un signe non alphanumérique liant deux mots, voire un signe diacritique, plutôt que comme une ponctuation au même titre que les points ou les virgules. D'ailleurs, ce signe est classé parmi les ponctuations de mot (avec le blanc et l'apostrophe) par certains auteurs (Tournier [1980] ; Arrivé *et al.* [1986] ; Riegel *et al.* [1994]), le distinguant ainsi des ponctuations de phrase. Catach *et al.* [1980] différencie également la majuscule, l'apostrophe et le trait d'union des autres signes de ponctuation, même s'ils les regroupent tous au sein des erreurs d'idéogrammes (voir figure 6.2 p. 93).
  En outre, la majorité des erreurs de trait d'union dans notre corpus ont pour conséquence la présence dans la phrase de deux ou plusieurs mots non pertinents sémantiquement ou grammaticalement, au lieu de l'unité lexicale attendue, et *vice versa* (par ex. **[...] ce*

qui permet peut être des adaptations). Par analogie avec les erreurs de substitution de lemmes, nous avons finalement considéré qu'il s'agissait ici également d'une substitution, mais impliquant des composés et des suites de mots. Nous avons donc déplacé la sous-catégorie `union` de `PONCTUATION` vers `LEXIQUE`, et y avons rassemblé toutes les erreurs de trait d'union conduisant à une ou des unité(s) lexicale(s) existant dans la langue.
Dans quelques rares autres cas de notre corpus, l'ajout ou l'omission d'un trait d'union engendre une erreur d'orthographe (par ex. **souentendait*, **il-y-a*). Nous avons donc reclassé dans cette catégorie les erreurs de trait d'union concernées.

Ces différents réajustements ont conduit à la version finale de notre typologie (voir tableau 6.2 p. 108) à partir de laquelle nous avons fini d'annoter notre corpus et repris les annotations déjà effectuées pour les conformer aux modifications. Elle comporte toujours 6 catégories, avec désormais 25 sous-catégories, comme l'illustre la figure 6.13 suivante :

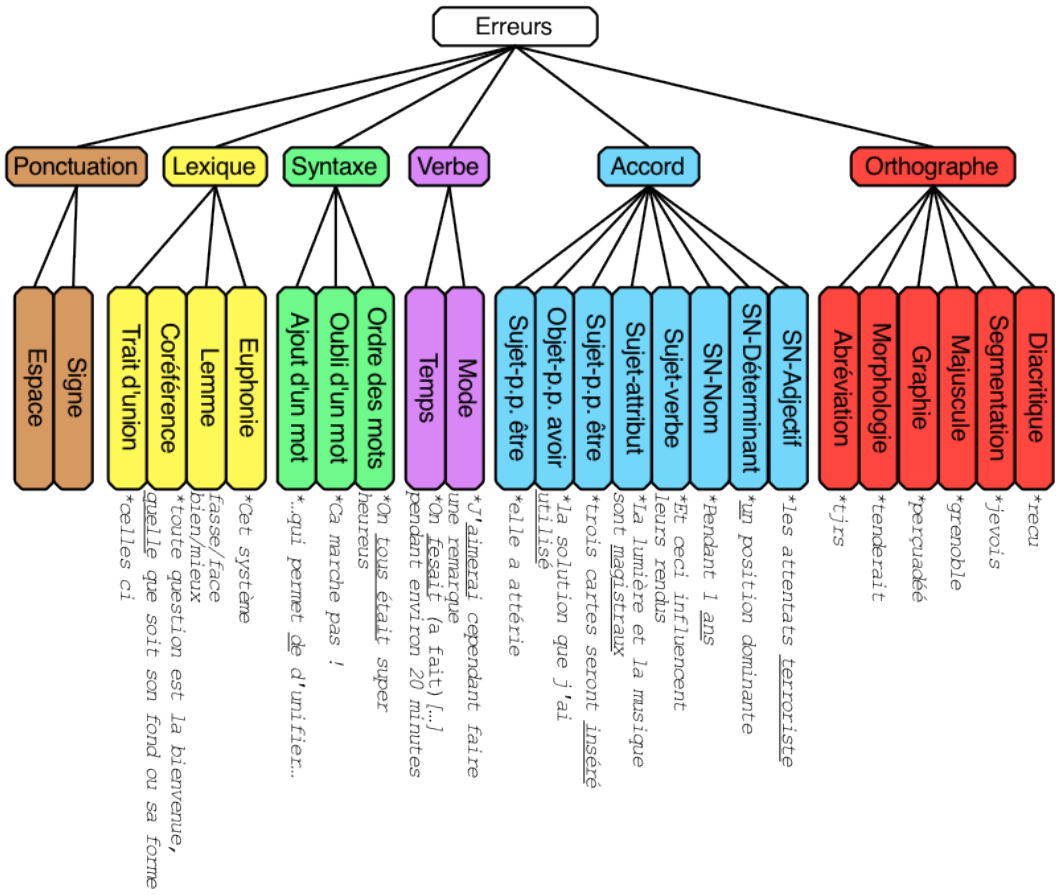


FIGURE 6.13 : Synthèse de la typologie d'erreurs finale

Catégorie	**Sous-cat.**	**Description**	**Exemple**
ORTHO	segm	Erreur de segmentation	*jevois, *pou rêtre
	graph	Erreur de graphie non morphologique	*perçuadéé
	maj	Erreur de majuscule	*grenoble
	morpho	Erreur de construction morphologique	*tenderait
	abrev	Mot abrégé	*tjrs, *pb
	dia	Erreur sur un diacritique (ou apostrophe)	*recu, *l auteur
ACCORD	A	Erreur de genre/nombre de l'adjectif dans un SN ou du participe passé employé comme adjectif	*les attentas terroriste
	D	Erreur de genre/nombre du déterminant	*un position dominante
	N	Erreur de genre/nombre du nom	*Pendant 1 ans
	V	Erreur d'accord du verbe avec son sujet	*Et ceci influencent leurs rendus
	Attr	Erreur d'accord de l'adjectif attribut avec son sujet	*La lumière et la musique sont magistraux
	pp-etre	Accord du participe passé avec le sujet après l'auxiliaire être	*trois cartes seront inséré
	spp-avoir	Accord du participe passé avec le sujet après l'auxiliaire avoir	*elle a attérie
	opp-avoir	Erreur d'accord du participe passé avec l'objet antéposé	*la solution que j'ai utilisé
VERBE	mode	Erreur de mode	*J'aimerai cependant faire une remarque
	temps	Erreur de temps	*On fesait (On a fait)
SYNTAXE	ordre	Mots mal ordonnés	*On tous était super heureus...
	oubli	Oubli d'un mot	*Ca marche pas !
	ajout	Présence d'un mot superflu	*...qui permet de d'unifier...
LEXIQUE	eupho	Euphonie non respectée	*Cet système...
	lemme	Substitution d'un lemme par un autre	fasse/face, bien/mieux
	coref	Pronom non accordé avec son antécédent	*toute question est la bienvenue, quelle que soit son fond ou sa forme
	union	Trait d'union manquant ou superflu	*dix huit, *celles ci
PONCT	esp	Espace manquante ou superflue autour d'une ponctuation	*Pourquoi pas?
	signe	Erreur sur un signe de ponctuation	

Tableau 6.2 : Typologie d'erreurs finale

6.2 Analyse quantitative des erreurs

6.2.1 Traitements statistiques des données

Afin de nous confronter à la réalité des erreurs tapuscrites et de tenter d'observer des régularités dans ces erreurs, nous avons soumis nos données à des traitements statistiques. Nous présentons ces traitements ici, en suivant l'ordre de leur application dans la suite de cette section.

a) Calcul des pourcentages moyens d'erreurs

En premier lieu, nous avons ramené les effectifs d'erreurs à des pourcentages afin de pouvoir observer leur répartition dans les différentes catégories et sous-catégories d'erreurs. Pour cela, deux possibilités de calculs se sont présentées. La première était de considérer le corpus comme une unité et de réaliser les calculs sur la base des effectifs totaux d'erreurs dans une (sous-)catégorie et/ou un sous-corpus (type d'écrits), avec par exemple des formules du type :

$$\frac{Nombre\ total\ d'erreurs\ de\ la\ (sous-)catégorie\ dans\ le\ \textbf{\textit{sous-corpus}}}{Nombre\ total\ d'erreurs\ du\ \textbf{\textit{sous-corpus}}} \times 100$$

Une autre possibilité était de travailler à l'échelle plus fine des textes, en calculant les proportions des différents types d'erreurs pour chacun d'entre eux et en effectuant ensuite des moyennes de ces proportions afin de disposer de pourcentages d'erreurs à un niveau de granularité supérieur (sous-corpus ou corpus). Dans ce cas, les formules utilisées suivent le modèle :

$$Moyenne\left(\frac{Y}{n} \times 100\right)$$

avec Y = *nombre total d'erreurs de la catégorie dans le* ***texte***
et n = *nombre total d'erreurs dans le* ***texte****.*

Le premier calcul permettrait de dégager des indices de dispersion par sous-corpus, mais il fait l'hypothèse *a priori* d'une homogénéité des scripteurs. Nous avons donc opté pour la seconde solution car elle permet de tenir compte de l'hétérogénéité des textes et des caractéristiques individuelles de chaque scripteur. Le tableau 6.4, que nous commentons p. 113, présente en effet des écarts-types élevés pour la plupart des moyennes indiquées (nombre de mots par texte, nombre d'erreurs par texte et nombre de mots pour une erreur). Ceci illustre une dispersion importante, et confirme la une grande hétérogénéité des données.

b) Comparaison de moyennes

Pour nos statistiques inférentielles, nous nous sommes tournée généralement vers l'analyse de variance (ou ANOVA) pour tester l'effet éventuel d'une ou plusieurs variables indépendantes (par ex. la langue maternelle du scripteur ou le type de texte) sur une variable dépendante (par ex. le pourcentages d'erreurs d'un certain type).

L'ANOVA est un test statistique paramétrique utilisé pour comparer les moyennes de plusieurs groupes. Son application est soumise à certaines contraintes, dont notamment [Howell, 1998 ; Judd *et al.*, 1995] :

- la normalité : les valeurs de la variable dépendante doivent suivre une distribution normale.
- l'homogénéité des variances : les différents groupes doivent avoir des variances homogènes.
- l'homogénéité des effectifs par cellule : les différents groupes doivent être composés d'échantillons de taille équivalente.

Nos données ne remplissaient pas les conditions nécessaires pour faire l'objet d'une ANOVA. Nous avons alors utilisé une méthode statistique permettant d'homogénéiser les variances. Il s'agit de la « transformation arc sinus », appelée aussi « transformation angulaire », qui est préconisée en particulier lorsque les valeurs sont des pourcentages ou des proportions. La formule initiale pour transformer un pourcentage p est $arcsin\sqrt{p}$, mais nous trouvons de nombreuses adaptations dans la littérature [Chanter, 1975 ; Freeman & Tukey, 1950 ; Johnson & Kotz, 1969]. Nous avons choisi la plus appropriée aux données qui comme les nôtres contiennent des valeurs extrêmes (proches de 0% ou 100%) [Pollard, 1979]. La formule est la suivante [2] :

$$\theta = arcsin\sqrt{\frac{Y + \frac{3}{8}}{n + \frac{3}{4}}}$$

avec Y = *nombre d'observations considérées*
et n = *effectif total des observations.*

Nos données exprimées en pourcentage ont toutes été transformées en valeurs θ suite à l'application de cette formule, avec comme effet des variances homogénéisées, comme nous pouvons l'observer dans le tableau 6.3. Nous présentons à gauche les pourcentages moyens d'erreurs par catégorie et les écarts-types (ET) correspondants avant transformation, et à droite la moyenne des valeurs θ des pourcentages transformés et leurs écarts-types respectifs. Les écarts-types initiaux se situent dans un intervalle allant de 11,9 à 28,4 ; après la mise en œuvre de la transformation arc sinus, cet intervalle ne s'étend plus que de 0,15 à 0,26

Catégories d'erreurs	**Sans transformation**		**Après transformation**	
	Moyenne des pourcentages d'erreurs	(ET)	Moyenne des θ des pourcentages d'erreurs	(ET)
ACCORD	24,9%	(26,8)	0,53	(0,24)
LEXIQUE	11,5%	(16,3)	0,41	(0,17)
SYNTAXE	3,6%	(11,9)	0,31	(0,15)
VERBE	4,3%	(12)	0,32	(0,15)
ORTHOGRAPHE	36,1%	(28,4)	0,66	(0,26)
PONCTUATION	19,6%	(26,4)	0,49	(0,26)

Tableau 6.3 : Résultats de la transformation arc sinus sur les pourcentages moyens d'erreurs par catégorie dans le corpus.

2. Par exemple, la valeur θ indiquée dans la 1re ligne du tableau 6.3 représente la moyenne des valeurs calculées pour chaque texte d'après la formule, avec pour valeur de Y le nombre d'erreurs d'accord du texte, et pour valeur de n le nombre total d'erreurs du texte.

Dans ce chapitre, toutes les analyses de variance que nous effectuons se fondent sur les données transformées en arc sinus. Cependant, dans un souci de clarté, ce sont les données initiales exprimées en pourcentages (et leurs écarts-types) que nous présentons dans les tableaux et que nous commentons.

Si des différences significatives entre les moyennes comparées sont mises en évidence par l'ANOVA, nous complétons l'analyse par une comparaison deux à deux de l'ensemble des moyennes à l'aide du test *post-hoc* PLSD de Fisher [Howell, 1998]. Ce test permet d'identifier plus précisément les paires entre lesquelles se situent les différences révélées par l'ANOVA. Afin de ne pas surcharger notre texte, nous donnerons généralement la valeur de p (*cf.* § *Seuil de significativité* p. 112) sans préciser systématiquement qu'elle se rapporte au test PLSD de Fisher. Une valeur de p seule sera donc associée à ce test.

Nous avons également appliqué des tests de Student pour échantillons appariés (tests t) afin de comparer des moyennes deux à deux dans un même groupe [Howell, 1998]. Les tests ont porté sur les pourcentages moyens d'erreurs des différentes catégories, dans chacun des sous-corpus considérés indépendamment.

c) Calcul des densités

Les pourcentages d'erreurs que nous avons calculés permettent de comparer pour un scripteur l'importance d'un type d'erreurs donné relativement aux autres types d'erreurs, mais ils ne permettent pas d'évaluer les fréquences absolues de ces types d'erreurs. Les proportions ne prennent en effet pas en compte la longueur des textes et le nombre de mots bien orthographiés.

Afin de compléter les analyses sur les proportions des catégories et sous-catégories d'erreurs, nous avons calculé les densités, ou fréquences d'apparition, de chaque type d'erreurs dans chaque texte. Il s'agit d'un ratio calculé de la manière suivante :

$$\frac{Nombre\ d'erreurs\ de\ la\ (sous-)catégorie\ dans\ le\ texte}{Nombre\ de\ mots\ du\ texte}$$

Comme pour les proportions, nous avons travaillé pour chaque sous-corpus avec les moyennes des valeurs obtenues par ce calcul. Pour les raisons déjà évoquées pour les proportions, nous avons appliqué aux densités la transformation en arc sinus et utilisé les données transformées pour la mise en œuvre des tests statistiques. Ajoutons que les densités telles qu'elles sont calculées indiquent le nombre d'erreurs d'un certain type pour un mot et consistent en des valeurs très faibles (<1) qui sont peu parlantes. Par exemple, pour le sous-corpus Dictées, la densité moyenne d'erreurs d'accord est de 0,0132 erreur par mot, ce qui permet difficilement de se représenter la fréquence effective de ces erreurs dans les textes. Nous avons donc choisi de présenter les densités avec des valeurs plus représentatives. Nous avons pour cela pris l'inverse des valeurs, soit $\frac{1}{densité}$, qui donne un résultat équivalent mais exprimé de manière plus claire en nombre de mots pour une erreur. La densité moyenne d'erreurs d'accord dans les dictées, donnée en exemple, sera ainsi présentée sous sa valeur inverse ($\frac{1}{0{,}0132}$) soit comme « 76 mots par erreur », ou encore « une erreur tous les 76 mots ».

Précisons que les statistiques inférentielles n'ont pas été effectuées à partir de l'inverse des densités afin d'être en mesure de traiter les cas des scripteurs n'ayant pas commis d'erreur dans une ou plusieurs catégories. En effet, avec un nombre d'erreurs de la (sous-)catégorie égal à 0 au

dénominateur [3], l'inverse n'est pas calculable.

d) Traitement des homophones

Nous avons également soumis à des tests statistiques les données concernant l'homophonie. Dans nos annotations, nous avons indiqué pour chaque erreur si la forme erronée est homophone de la forme correcte (*cf.* p. 105). Nous avons une fois encore calculé dans un premier temps la proportion des erreurs homophones de chaque (sous-)catégorie par rapport à l'ensemble des erreurs de la même (sous-)catégorie au niveau des textes, puis nous avons ensuite travaillé sur les moyennes de ces proportions pour les sous-corpus :

$$Moyenne\left(\frac{Nombre\ d'erreurs\ homophones\ de\ la\ (sous-)catégorie\ dans\ le\ texte}{Nombre\ d'erreurs\ de\ la\ (sous-)catégorie\ dans\ le\ texte}\times 100\right)$$

e) Seuil de significativité

Tous les tests d'inférence statistiques que nous avons utilisés permettent d'estimer le risque pris à généraliser à l'ensemble de la population un résultat ne concernant qu'un échantillon de cette population. Rappelons que, selon l'usage en sciences humaines et sociales, la généralisation d'un résultat est admise si le risque, noté p, est inférieur ou égal à 0,05, auquel cas le résultat est dit significatif. Si p est supérieur à 0,05 mais inférieur ou égal à 0,10, le résultat est considéré comme tendanciel. Lorsque la valeur de p est supérieure à 0,10, le résultat est donné comme non significatif, donc non généralisable.

Notons également que nous n'avons pas effectué de test sur des effectifs inférieurs à 5%, considérant qu'ils ne donneraient pas des résultats suffisamment fiables. Dans les cas où les effectifs d'erreurs d'une catégorie (ou sous-catégorie) d'un sous-corpus représentent moins de 5% des erreurs du sous-corpus (ou de la catégorie), les effectifs en question sont présentés en gris dans le tableau récapitulatif A.1 p. 242. Nous indiquons également en gris les proportions ou les densités moyennes correspondantes dans les tableaux de ce chapitre.

6.2.2 Description quantitative du corpus

Dans cette section, nous nous intéressons aux données quantitatives que nous avons pu extraire de notre corpus suite à son annotation. Dans la présentation de ces résultats, nous distinguons quatre sous-corpus, que nous appellerons également « types de textes », et qui font chacun référence à la fois à un type de document, à sa situation de scription et au type de scripteur. Ces types de textes correspondent en majeure partie aux quatre types de documents recueillis : dictées, résumés d'articles scientifiques, mails de listes de diffusion et commentaires de blogs (voir tableau 5.1 p. 81). Les résumés et les mails constituent chacun un sous-corpus du même nom. En revanche, le sous-corpus Dictées ne contient que les dictées des scripteurs de langue maternelle française. Les six dictées effectuées par des scripteurs non natifs ont été adjointes à l'ensemble des commentaires de blogs, également rédigés par des non natifs, afin de constituer un sous-corpus de FLE exclusivement, que nous nommons Textes FLE. Cette décision de regrouper les écrits produits par des apprenants de FLE fait suite à l'observation d'une différence significative de

3. Densité inverse : $\frac{Nombre\ de\ mots\ du\ texte}{Nombre\ d'erreurs\ de\ la\ (sous-)catégorie\ dans\ le\ texte}$

densité d'erreurs entre ces textes et ceux de francophones natifs dans notre corpus [4] , ainsi qu'aux constats que fait Granger [2007] :

> « La langue de l'apprenant diffère de la langue maternelle tant quantitativement que qualitativement. Elle se manifeste par des fréquences de mots, d'expressions et de structures très différentes, certains éléments étant surutilisés et d'autres considérablement sous-utilisés. Elle est également caractérisée par un taux élevé d'usages impropres, à savoir des erreurs orthographiques, lexicales et grammaticales. »
>
> [Granger, 2007, p. 1]

a) Description générale du corpus

Nous présentons dans le tableau 6.4 quelques observations très générales sur le corpus et ses sous-corpus.

	Effectifs (%)			**Moyennes (ET)**		
	Textes	Mots	Erreurs	Mots/Texte	Erreurs/Texte	Mots/Erreur*
Dictées	145	21929 (66%)	1360 (64,9%)	151,2 (35,5)	9,4 (11)	15,8
Résumés	34	4978 (15%)	241 (11,5%)	146,4 (54,5)	7,1 (5)	20,6
Mails	53	4759 (14,3%)	231 (11%)	89,8 (75,2)	4,4 (5,1)	18,3
Textes FLE	20	1566 (4,7%)	264 (12,6%)	78,3 (47,8)	13,2 (7,4)	5,4
CORPUS	**252**	**33232 (100%)**	**2096 (100%)**	**131,9 (57,5)**	**8,3 (9,4)**	**14,5**

* *Rappelons que les valeurs que nous donnons ici sont les inverses des valeurs utilisées pour les calculs de densité. Nous omettons volontairement d'indiquer les écarts-types (ET) correspondant à ces valeurs inverses car ils ne pourraient être calculés qu'approximativement.*
Pour information, les densités calculées ($\frac{nb\ d'erreurs}{nb\ de\ mots}$) et leurs écart-types sont les suivants : Dictées : 0,063 (0,080) ; Résumés : 0,049 (0,028) ; Mails : 0,055 (0,048) ; Textes FLE : 0,185 (0,079).

Tableau 6.4 : Description du corpus

Dans la première colonne du tableau 6.4, nous retrouvons le nombre de textes constituant chacun des sous-corpus que nous avons définis. Les dictées représentent à elles seules plus de la moitié de notre corpus avec 145 textes, soit 57,5% d'un total de 252 textes. Cette proportion atteint même les deux tiers si nous nous plaçons à l'échelle du mot : le nombre de mots (seconde colonne) du sous-corpus Dictées représente 66% du corpus entier, avec près de 22 000 mots sur les 33 232 que compte l'ensemble du corpus. Le sous-corpus Mails, constitué de 53 textes, équivaut au sous-corpus Résumés en nombre de mots (respectivement 14,3% et 15% des mots du corpus), alors que ce dernier ne contient que 34 textes. Les textes FLE, quant à eux, correspondent à seulement 4,7% du nombre total de mots, soit environ trois fois moins que les résumés ou les mails, et pourraient de ce fait être considérés comme un sous-corpus plutôt marginal par rapport à l'ensemble. Nous constatons cependant qu'il n'est pas aussi marginal si nous nous intéressons aux effectifs d'erreurs de la troisième colonne du tableau. En effet, nous voyons que 12,6% des erreurs du corpus se situent dans ce sous-corpus Textes FLE, ce qui est légèrement supérieur au

4. ANOVA : $F_{(3,248)}$=20,802 ; p<0,0001. PLSD de Fisher : p<0,0001 pour les paires Textes FLE-Dictées, Textes FLE-Résumés et Textes FLE-Mails (*cf.* § a) p. 113)

nombre d'erreurs se trouvant dans les résumés ou les mails (respectivement 11,5% et 11%), alors que le nombre de mots de ces deux sous-corpus est environ trois fois plus grand.

Sans surprise, nous pouvons déduire des ces données que la concentration en erreurs est plus forte dans les textes FLE, et ceci est confirmé dans la dernière colonne du tableau 6.4 p. 113. Nous y indiquons la densité moyenne d'erreurs, sous la forme du nombre moyen de mots pour une erreur. Nous observons une nette différence entre le sous-corpus Textes FLE et les trois autres sous-corpus, confirmée d'une part par l'ANOVA qui montre un effet du type de sous-corpus sur la densité d'erreurs ($F_{(3,248)}$=20,802 ; p<0,0001), et d'autre part par le test PLSD de Fisher qui montre des différences significatives entre les textes FLE et les trois autres types de textes (p<0,0001 pour les trois paires comparées). Nous trouvons en moyenne une erreur tous les 5,4 mots dans les textes FLE, contre une erreur tous les 15,8 à 20,6 mots seulement dans les autres textes. Ces résultats appuient ceux obtenus par Granger [2007, p. 1] à partir d'un corpus de FLE : « Dans nos corpus de français langue étrangère, [le taux d'erreurs] varie d'une erreur tous les 5,2 mots à une erreur tous les 9,5 mots, la moyenne étant de 6,89 ». Notre moyenne de 5,4 mots par erreur n'est en outre pas significativement différente des 6,89 obtenus par Granger (Test t sur échantillon unique : p=0,41). Nous avons cependant une amplitude de densité d'erreurs plus importante allant de 2,70 à 11,1 mots par erreurs.

Ces résultats nous permettent de nous attendre à observer lors de nos différentes analyses des données, des différences fréquentes entre les textes FLE et les textes de francophones natifs.

Les colonnes Mots/Texte et Erreurs/Texte du tableau 6.4 donnent respectivement le nombre moyen de mots par texte et le nombre moyen d'erreurs par texte. En ce qui concerne le nombre moyen de mots par texte, nous relevons que la longueur des textes dépend des sous-corpus (ANOVA : $F_{(3,248)}$=28,517 ; p<0,0001), et que ceux-ci forment deux groupes au sein desquels les longueurs sont équivalentes : les dictées et résumés d'une part (p=0,612), avec en moyenne respectivement 151 et 146 mots par texte, et les mails et textes FLE d'autre part (p<0,0001), avec des moyennes plus faibles de 90 et 78 mots par texte.

Combinée au très grand nombre de textes du sous-corpus, la longueur des dictées explique que ces dernières occupent les deux tiers du corpus en nombre de mots, et presque autant en nombre d'erreurs. De même, la longueur moindre des mails par rapport aux résumés (p<0,0001) est compensée par une quantité de mails plus importante.

Notons cependant que, comme nous l'avons mentionné dans la section *Traitements statistiques des données* p. 109, les écarts-types correspondant à ces moyennes de mots par texte sont relativement importants et traduisent une grande hétérogénéité dans la longueur des textes d'un même sous-corpus [5].

Nous retrouvons un effet significatif du type de textes ($F_{(3,248)}$=6,054 ; p=0,001) sur le nombre d'erreurs par texte. Les dictées contiennent en moyenne 9,4 erreurs, avec une étendue de 82 (1 à 83 erreurs). Cette étendue [6] du nombre d'erreurs est plus faible dans les autres sous-corpus : 1 à 20 erreurs pour les résumés, 1 à 23 pour les mails et 3 à 29 pour les textes FLE. Ces derniers présentent le plus grand nombre moyen d'erreurs par texte et se distinguent significativement des mails (PLSD de Fisher : p<0,0001) et des résumés (p=0,019), et tendanciellement des dictées (p=0,081).

5. Dispersion, en nombre de mots : Dictées [102-237], Résumés [64-245], Mails [19-452], Textes FLE [20-184].
6. Différence entre la valeur la plus grande et la valeur la plus petite.

b) Distribution des catégories d'erreurs

Distribution dans le corpus

Dans le tableau 6.5, nous présentons la proportion moyenne de chaque catégorie d'erreurs par texte, tous types de textes confondus. La représentation graphique des données du tableau permet de mieux visualiser la distribution des erreurs dans les différentes catégories.

Nous voyons par exemple que la majorité des erreurs appartient à la catégorie `Orthographe` (36,1%). Cette valeur rappelle le taux d'erreurs que les outils de première génération (correcteurs orthographiques) étaient capables de détecter (voir tableau 3.1 p. 42) :

> « Les insuffisances de ces systèmes étaient évidentes, particulièrement en langue française où le taux de correction avoisinait les 40 %. Cette performance est comparable aux résultats d'une étude réalisée au Québec qui indiquait qu'en moyenne 39 % des fautes commises sont des fautes d'orthographe [Bureau, 1985]. »
>
> [Doll & Coulombe, 2004, p. 36]

Catégories d'erreur	CORPUS
ACCORD	24,9% (26,8)
LEXIQUE	11,5% (16,3)
SYNTAXE	3,6% (11,9)
VERBE	4,3% (12)
ORTHOGRAPHE	36,1% (28,4)
PONCTUATION	19,6% (26,4)

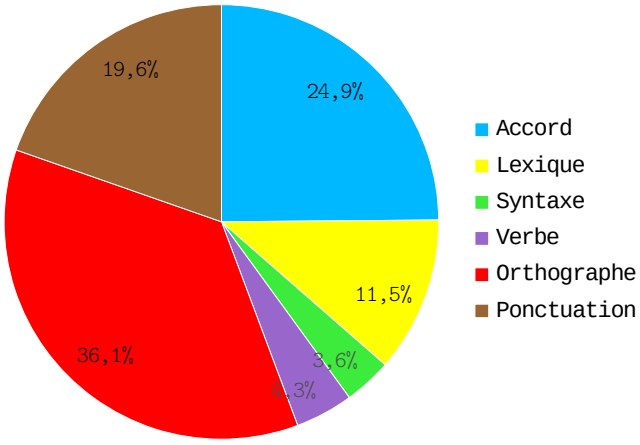


Tableau 6.5 : Pourcentages moyens d'erreurs dans chaque catégorie au sein du corpus (ET) [7]

Afin de comparer notre pourcentage d'erreurs d'orthographe aux données mentionnées ci-dessus, nous avons réalisé un test t sur échantillon unique. Il montre que notre moyenne de 36,1% n'est pas différente des 39% que mentionne Bureau [1985] (p=0,104) [8]. Par ailleurs, la proportion d'erreurs d'`Orthographe` se révèle être significativement supérieure aux cinq autres catégories (test t apparié : p<0,0001 pour toutes les paires comparées).

En ajoutant aux erreurs d'orthographe les erreurs d'accord, qui représentent la seconde part la plus importante avec 24,9%, nous couvrons plus de la moitié des erreurs relevées dans le corpus (61%). Un nouveau test t sur échantillon unique montre qu'il s'agit d'une valeur équivalente (p=0,624) aux 60% approximatifs de taux de correction des outils de seconde génération (voir tableau 3.1 p. 42), capables de détecter des erreurs d'accords locaux en plus des erreurs

7. L'abréviation « ET » renvoie à l'écart-type.
8. Test effectué à partir des données non transformées en arc sinus.

d'orthographe. Notons toutefois que dans les 24,9% d'erreurs d'accord que nous avons annotées, toutes ne concernent pas des accords locaux. Un certain nombre d'entre elles se situe au niveau d'accords distants, qui ne font donc pas partie du domaine de compétence de la deuxième génération d'outils. Il faudrait alors annoter la distance entre les éléments concernés par chaque erreur d'accord, afin de connaître la proportion d'erreurs dans des contextes locaux par rapport aux contextes distants et ainsi évaluer l'impact de la distance sur l'apparition d'erreurs. Il s'agit d'un complément d'annotation que nous avons effectué par la suite (*cf.* § g) p. 131).

Les catégories `Accord`, `Lexique`, `Orthographe` et `Ponctuation` se distinguent toutes significativement les unes des autres, avec toutefois une différence seulement tendancielle entre `Accord` et `Ponctuation` (test t apparié : $p<0{,}0001$ pour toutes les paires comparées sauf `Accord-Ponctuation` : $p=0{,}083$).

Quant aux catégories `Syntaxe` et `Verbe`, elles représentent chacune moins de 5% des erreurs du corpus et n'ont donc pas fait l'objet de tests statistiques.

Distribution dans les sous-corpus

Les différentes catégories d'erreurs dont nous venons de commenter la distribution au sein du corpus, n'apparaissent pas toujours dans des proportions équivalentes selon les sous-corpus, comme le montrent le tableau 6.6 et la figure 6.14 p. 117.

Catégorie Accord

Pour cette catégorie, les ANOVA effectuées révèlent que le type de sous-corpus a un impact sur les proportions d'erreurs ($F_{(3,248)}=4{,}926$; $p=0{,}002$). Le tableau et les graphiques montrent des proportions relativement proches entre les dictées (23,3%) et les résumés (20,6%), mais très inégales pour les mails (35,4%) et les textes FLE (15,4%). Il y a significativement plus d'erreurs d'accord dans les dictées que dans les textes FLE ($p=0{,}044$), mais également significativement plus dans les mails que dans chacun des autres types de textes (Mails-Dictées : $p=0{,}007$; Mails-Résumés : $p=0{,}015$; Mails-Textes FLE : $p=0{,}001$). De plus, les erreurs d'accord sont les plus représentées dans les mails, alors qu'elles n'arrivent qu'en seconde où troisième position pour les résumés, les dictées et les textes FLE, et toujours derrière l'orthographe.

Catégorie Lexique

Sur les graphiques, les mails semblent encore se démarquer pour la catégorie `Lexique`, avec une proportion d'erreurs environ de moitié plus petite que dans les autres sous-corpus (6,8% contre 12,5% à 14,2%). L'analyse de variance ne montre cependant aucun effet significatif du type de textes sur ces erreurs ($p=0{,}929$).

Catégories Syntaxe et Verbe

Ces deux catégories sont globalement les moins représentées dans tous les sous-corpus, avec des effectifs qui constituent moins de 5% des erreurs dans chaque type de textes, à l'exception des erreurs de syntaxe dans les textes FLE. Des différences entre les sous-corpus sont visibles à partir des graphiques, avec par exemple une proportion d'erreurs de syntaxe environ six fois plus grande dans les textes FLE (10,1%) par rapport aux dictées (1,7%), ou encore une proportion d'erreurs sur les verbes environ cinq fois plus grande dans les mails (7,4%) par rapport aux résumés (1,5%).

Catégories d'erreur	Dictées	Résumés	Mails	Textes FLE
ACCORD	23,3% (22,6)	20,6% (22,5)	35,4% (37,9)	15,4% (18,3)
LEXIQUE	12,7% (16,1)	12,5% (16,8)	6,8% (15,5)	14,2% (17,6)
SYNTAXE	1,7% (6,4)	2,6% (7,4)	6,8% (21,4)	10,1% (10,6)
VERBE	4,2% (10,1)	1,5% (5,4)	7,4% (19,3)	2% (4,5)
ORTHOGRAPHE	34,2% (27,8)	57% (20,5)	29% (31,3)	32,6% (21,6)
PONCTUATION	23,9% (28)	5,8% (10,5)	14,6% (25,7)	25,8% (26,3)

Tableau 6.6 : Pourcentages moyens des catégories d'erreurs dans les différents types de textes

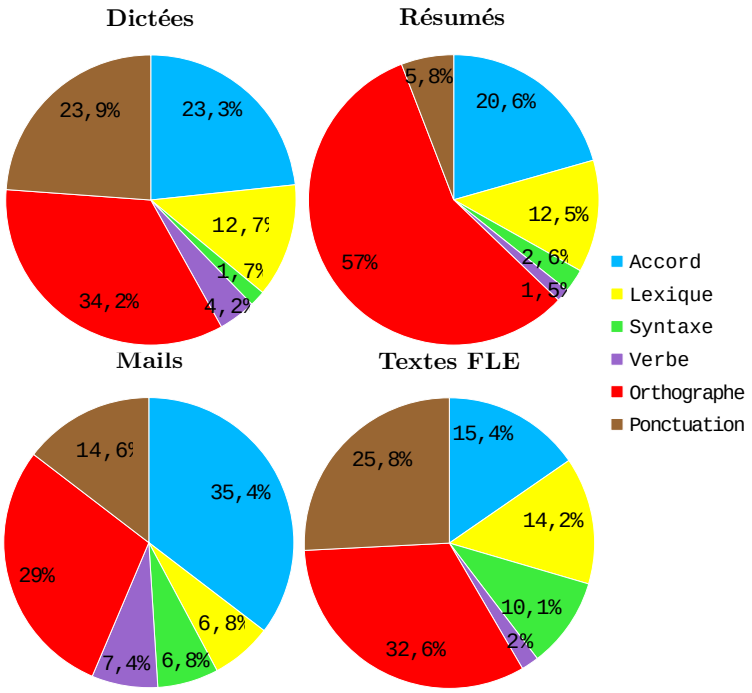


FIGURE 6.14 : Pourcentages moyens des catégories d'erreurs dans les différents types de textes

La proportion moyenne d'erreurs de syntaxe dans les mails est également plus élevée que dans les dictées (rapport de 4) ou les résumés (rapport de 2,6). Si nous examinons les erreurs de ce type relevées dans les mails, nous constatons que la plupart (7/9) résulte de l'utilisation à l'écrit d'un langage plutôt oral, avec notamment l'omission du « ne » de la négation (*`[...] j'ai pas d'idée immédiate`). Nous ne retrouvons aucune erreur de ce type dans les dictées et les résumés, dont les contenus ne consistent pas en des conversations écrites pouvant favoriser ces tournures typiques du discours oral. De même, bien que les commentaires de blogs soient souvent le support

d'écrits de ce type, ce n'est pas le cas avec nos commentaires de blogs en FLE, dans lesquels nous ne trouvons pas de telles formulations, probablement à cause du caractère formel apporté par le fait que ces commentaires soient amenés à être lus, voire évalués, par un enseignant.

Catégorie Orthographe

Concernant les erreurs d'orthographe, les graphiques montrent une proportion nettement plus grande dans les résumés, avec 57%, soit plus de la moitié des erreurs dans ce type de textes, alors que les autres sous-corpus sont plus proches du tiers (de 29% à 34,2%). Il existe bien un effet du type de textes (ANOVA : $F_{(3,248)}=7{,}331$; $p<0{,}0001$) ici, avec une différence entre les résumés et les autres types de textes, révélée par le test PLSD de Fisher (Résumés-Dictées et Résumés-Mails : $p<0{,}0001$; Résumés-Textes FLE : $p=0{,}001$). Nous verrons dans la suite de nos analyses (voir p. 130) que la fréquence des erreurs de frappe peuvent constituer une explication à ce résultat.

Catégorie Ponctuation

Les proportions d'erreurs dépendent également du type de textes pour la catégorie `Ponctuation` (ANOVA : $F_{(3,248)}=4{,}239$; $p=0{,}006$), avec de nouveau une différence entre les résumés (5,8%) et les autres sous-corpus (Dictées : 23,9%, Mails : 14,6%, Textes FLE : 25,8%), mais cette fois-ci à cause d'une proportion d'erreurs plus faible (Résumés-Dictées : $p=0{,}001$; Résumés-Mails : $p=0{,}004$; Résumés-Textes FLE : $p=0{,}012$).

Cette première analyse de la répartition des erreurs par catégorie permet de dégager des types d'erreurs majoritaires ou minoritaires :

- Les erreurs d'orthographe sont majoritaires dans tous les sous-corpus à l'exception des mails (dans les mails les erreurs majoritaires sont les erreurs d'accord) ;
- Les erreurs de syntaxe et de verbe sont minoritaires dans tous les sous-corpus ;

c) Distribution intra-catégories

Catégorie Accord

Le tableau 6.7 p. 119 expose les pourcentages moyens d'erreurs calculés pour chaque sous-catégorie d'erreurs d'accord dans les sous-corpus. Nous avons scindé ce tableau horizontalement afin de mettre en évidence deux ensembles qu'il nous a paru pertinent de distinguer : les erreurs d'accord au sein de syntagmes nominaux d'une part, et au sein de syntagmes verbaux d'autre part.

Dans le premier ensemble, nous avons regroupé les erreurs d'accord touchant les adjectifs, les déterminants et les noms. Dans le second ensemble se trouvent les erreurs d'accord sur les verbes, sur les adjectifs attributs et sur les participes passés.

L'ANOVA révèle un effet significatif du type de textes sur la proportion d'erreurs dans les syntagmes nominaux et verbaux ($F_{(3,248)}=6{,}411$; $p<0{,}0001$). Comme le montre le tableau, dans les dictées, les résumés et les textes FLE, les erreurs dans les syntagmes nominaux sont plus importantes que dans les syntagmes verbaux, tandis que dans les mails, les erreurs dans les syntagmes verbaux prédominent (Mails-Dictées et Mails-Résumés : $p<0{,}0001$; Mails-Textes FLE :

p=0,003).

ACCORD	Dictées	Résumés	Mails	Textes FLE	CORPUS
`adjectif`	21,2% (32,4)	24,3% (31,2)	7% (23,8)	24,3% (36,2)	19,0% (31,4)
`déterminant`	7,2% (23,4)	11% (26,6)	5,7% (23,6)	38% (43,3)	10% (27,2)
`nom`	32,7% (34,8)	39,2% (40,1)	17,7% (38,1)	11,3% (28)	28,6% (36,4)
Total dans les syntagmes nominaux	*61,1% (40)*	*74,5% (34,4)*	*30,4% (45,3)*	*73,7% (40,6)*	*57,6% (42,8)*
`verbe`	5,8% (17,2)	18,2% (33,3)	49,9% (49)	8,6% (15,9)	16,4% (33,1)
`attribut`	0,1% (1,1)	4% (12,8)	2,9% (16,9)	7,8% (20,8)	1,8% (10,8)
`part. passé avec v. d'état`	20,9% (33,7)	3,3% (11,5)	14% (33,1)	6,7% (17,6)	16,1% (31,1)
`sujet-pp avec avoir`	8,4% (21,1)	-	-	-	5% (16,7)
`objet-pp avec avoir`	3,8% (11)	-	2,9% (16,9)	3,3% (12,9)	3,1% (11,9)
Total dans les syntagmes verbaux	*38,9% (40)*	*25,5% (34)*	*69,6% (45)*	*26,3% (41)*	*42,4% (42,8)*

Tableau 6.7 : Proportions moyennes (ET) des erreurs de la catégorie `ACCORD` dans chaque type de textes

Syntagmes verbaux

La différence relevée au niveau des mails nous ayant semblé curieuse, nous avons cherché à comprendre ce qui pouvait l'expliquer. Nous voyons dans le tableau 6.7 que la moitié des erreurs d'accord dans les mails apparaissent au niveau des verbes (21/43). L'observation des erreurs en question a révélé qu'il s'agit pour les deux tiers (14/21) de l'omission de l'accord des verbes conjugués à la deuxième personne du singulier. Les mails de notre corpus sont en effet pour la majorité rédigés à la deuxième personne du singulier, ce qui n'est pas le cas de nos autres textes, qui sont rédigés à la troisième personne. Les tests statistiques confirment bien ici un effet du type de textes sur les pourcentages d'erreurs d'accord sur les verbes (ANOVA : $F_{(3,248)}=17{,}830$; $p<0{,}0001$), avec une valeur supérieure dans les mails (Mails-Dictées et Mails-Textes FLE : $p<0{,}0001$; Mails-Résumés : $p=0{,}001$).

Ce type d'erreurs est sans doute lié à l'homophonie des formes conjuguées, très fréquente en français. Dans la plupart des cas, les trois personnes du singulier sont homophones. Cette homophonie s'étend également à la troisième personne du pluriel pour les verbes du premier groupe, qui constituent une grande majorité des verbes du français [Nouveau, 2007].

Pour les autres sous-catégories, les occurrences d'erreurs étant souvent très peu nombreuses

voire nulles, nous n'avons pas effectué de tests statistiques. Nous observons tout de même que seules les dictées présentent des erreurs d'accords du sujet avec un participe passé précédé de l'auxiliaire « avoir ». Nous remarquons également une quasi égalité au niveau du corpus entre les pourcentages moyens d'erreurs d'accord des verbes (16,4%) et des participes passés après un verbe d'état (16,1%), qui laisserait penser que ces deux types d'erreurs sont aussi fréquents l'un que l'autre. Cela ne semble pourtant pas être le cas si nous regardons les données de chaque sous-corpus. Nous voyons un déséquilibre important entre les deux pourcentages moyens, avec des erreurs plus nombreuses sur les verbes que sur les participes passés dans les résumés, les mails et les textes FLE, et l'inverse dans les dictées.

Syntagmes nominaux

Dans les syntagmes nominaux, les ANOVA réalisées sur chaque sous-catégorie montrent qu'il existe un effet du sous-corpus pour les déterminants ($F_{(3,248)}$=3,925 ; p=0,009). Nous voyons dans le tableau 6.7 que le pourcentage de ce type d'erreurs dans les textes FLE (38%) est nettement supérieur aux autres types de textes (5,7% à 11%). Cependant, la différence entre ce sous-corpus et les trois autres n'est significative qu'avec les dictées (p=0,002), et tendancielle avec les mails (p=0,082) et les résumés (p=0,086).

Catégorie Lexique

Les erreurs au sein de la catégorie `Lexique` (voir tableau A.5 p. 245 en annexes) se répartissent surtout entre les sous-catégories `lemme` et `union`, la première à hauteur de 66,3% en moyenne dans le corpus entier, et la seconde de 29,4%. Les erreurs de substitution d'un lemme par un autre constituent une large majorité dans cette catégorie, avec 163 occurrences, contre seulement trois occurrences d'erreurs de coréférence au total, et quatre d'euphonie. Dans nos tests statistiques, nous ne tenons pas compte de ces deux dernières sous-catégories car elles sont trop faiblement représentées.

Dans les textes FLE, les erreurs de lemme représentent un pourcentage moyen très important (86,5%) des erreurs de la catégorie `Lexique`, pourcentage également élevé dans les autres sous-corpus (entre 46,4% et 72,2%). L'analyse de variance réalisée montre que le type de textes à un effet sur ces proportions ($F_{(3,248)}$=5,013 ; p=0,002). Les textes FLE présentent un pourcentage d'erreurs de lemme plus élevé que les autres textes (Textes FLE-Dictées : p=0,010 ; Textes FLE-Résumés : p=0,032 ; Textes FLE-Mails : p<0,0001). Les mails au contraire présente le pourcentage moyen d'erreurs de lemme le plus faible. La différence de pourcentage des mails est significative avec les dictées (p=0,019) et les textes FLE (p<0,0001), et tendancielle avec les résumés (p=0,077).

Catégorie Syntaxe

Dans cette catégorie (voir tableau A.6 p. 245 en annexes), les occurrences d'erreurs sont faibles et se concentrent principalement dans la sous-catégorie `oubli` (69,6%), puis dans `ajout` (29,2%). La sous-catégorie `ordre` ne contient qu'une seule erreur, dans le sous-corpus Textes FLE.

Catégorie Verbe

Le sous-corpus Textes FLE contient l'unique occurrence d'erreur de temps dans la catégorie `Verbe` (voir tableau A.7 p. 245 en annexes). L'intégralité des autres erreurs sur les verbes constitue

la sous-catégorie `mode`. Nous y retrouvons principalement les confusions entre les formes infinitives, les formes conjuguées et les participes passés. L'observation des erreurs correspondantes permet par exemple de dégager 63% d'erreurs de mode où un infinitif ou un verbe conjugué se substituent à un participe passé, et de manière plus générale 87,7% de confusion entre ces trois formes au sein des erreurs de mode.

Catégorie Orthographe

Pour l'orthographe (voir tableau A.9 p. 246 en annexes), les sous-catégories `Graphie`, `Diacritique` et `Majuscule` contiennent la quasi totalité des erreurs, avec près de la moitié (48,1%) d'erreurs de graphie. Les pourcentages moyens d'erreurs calculés dans ces trois sous-catégories semblent proches entre les dictées (`Graphie` : 45,7%, `Diacritique` : 20,2% et `Majuscule` : 30,1%), les mails (39%, 21,7% et 28%) et les textes FLE (43,3%, 12,1% et 29%), mais les résumés donnent des proportions assez différentes (67,7%, 4,1% et 11,1%). Les proportions dans ces sous-catégories semblent dépendre du type de textes, ce que confirment les ANOVA effectuées (`Graphie` : $F_{(3,248)}$=5,872 ; p=0,001. `Diacritique` $F_{(3,248)}$=14,793 ; p<0,0001. `Majuscule` $F_{(3,248)}$=10,401 ; p<0,0001).

Les tests PLSD de Fisher attestent également de la différence entre les résumés et les autres textes pour la graphie et les majuscules. Les résumés présentent un pourcentage moyen d'erreurs de graphie nettement supérieur (67,7%) à celui des dictées (45,7% ; p<0,0001), des mails (39% ; p<0,0001) et des textes FLE (43,3% ; p=0,005). Les résumés contiennent également moins d'erreurs de majuscule (11,1%) que le reste du corpus (Résumés-Dictées et Résumés-Mails : p<0,0001 ; Résumés-Textes FLE : p=0,006) où les valeurs moyennes varient entre 28% (Mails) et 30,1% (Dictées).

En revanche, pour les diacritiques, les résumés ne se détachent pas du reste du corpus. Ce sont tous les types de textes qui se distinguent les uns des autres, avec néanmoins des différences seulement tendancielles entre les dictées et les mails (p=0,081), et entre les résumés et les textes FLE (p=0,092).

Les résumés sont donc globalement différents des autres types de textes au niveau des proportions d'erreurs dans la catégorie `Orthographe`, de même que nous avons vu précédemment qu'ils se distinguaient sur la taille significativement plus importante de cette catégorie d'erreurs par rapport aux autres sous-corpus.

Catégorie Ponctuation

La dernière catégorie que nous commentons, `Ponctuation`, est constituée des sous-catégories `espace` (ajout/oubli d'espace) et `signe` (substitution de signes). La très grande majorité (88%) sont des erreurs d'espaces, dont l'observation révèle qu'il s'agit dans la plupart des cas de la mauvaise utilisation des espaces avant et après les différents signes de ponctuation. Ces erreurs sont sans doute dues à une méconnaissance des règles typographiques en français ou à une confusion avec les règles typographiques d'autres langues.

De cette observation des répartitions d'erreurs intra-catégorie, il ressort principalement que :

- Les erreurs d'accord se situent majoritairement dans les syntagmes nominaux pour tous les sous-corpus sauf les mails.

- Les erreurs d'accord des verbes sont en proportion beaucoup plus importante dans les mails.
- Les erreurs de lemme sont les principales erreurs de lexique et sont en proportion plus élevée dans les textes FLE et plus faible dans les mails.
- Les erreurs de graphie sont les principales erreurs d'orthographe et sont en proportion plus élevée dans les résumés.

d) Focus sur les erreurs grammaticales

Nous avons effectué un tour d'horizon de la répartition des erreurs entre les catégories et sous-catégories de notre typologie, mais nous nous intéressons prioritairement aux erreurs de grammaire. Dans cette section, nous laissons donc volontairement de côté les erreurs d'orthographe et de ponctuation pour nous concentrer sur les catégories `Accord`, `Lexique`, `Syntaxe` et `Verbe` qui regroupent l'ensemble des erreurs grammaticales. Nous avons recalculé les pourcentages d'erreurs dans ces quatre catégories en nous fondant sur l'effectif total d'erreurs de grammaire dans chaque texte, et non plus sur l'effectif total d'erreurs par texte, selon le modèle suivant :

$$Moyenne\left(\frac{Nombre\ total\ d'erreurs\ de\ la\ catégorie\ dans\ le\ texte}{Nombre\ total\ d'erreurs\ \textbf{\textit{de grammaire}}\ du\ texte} \times 100\right)$$

Le tableau 6.8 présente le pourcentage moyen de chaque catégorie parmi les seules erreurs de grammaire, qui représentent en moyenne 44,3% des erreurs du corpus. Nous perdons ainsi beaucoup en effectif puisque de 2096 erreurs au total nous descendons à 782 erreurs de grammaire.

Catégorie Accord

Nous remarquons en premier lieu que les erreurs d'accord représentent la catégorie la plus importante dans tous les sous-corpus, avec une part de plus de 50% (de 52,1% à 59,2%), à l'exception de Textes FLE où elle n'est que de 35%. Il n'y a cependant pas d'effet du type de textes (ANOVA : $F_{(3,248)}$=2,069 ; p=0,105), contrairement à ce qu'avait montré notre première analyse prenant en compte l'ensemble des catégories d'erreurs (*cf.* p.118). Bien que les mails aient toujours une proportion d'erreurs d'accord (53,2%) plus importante que les autres textes (entre 35% et 52,4%), lorsque nous considérons uniquement les erreurs de grammaire, aucune différence significative ne les distinguent.

Catégorie Lexique

La seconde catégorie la plus importante d'après le tableau 6.8 est `Lexique`. Tout comme dans la première partie de cette section (*cf.* p.118), nous constatons que la proportion des erreurs dans les mails est environ de moitié plus petite que dans les autres types de textes, mais les tests statistiques ne révèlent toujours pas d'effet du sous-corpus sur cette catégorie.

Catégorie Syntaxe

Pour `Syntaxe`, les effectifs constituent à présent au moins 5% des erreurs de grammaire dans chaque sous-corpus et nous permettent donc d'appliquer les tests que nous n'avons pas effectués précédemment. Nous avions constaté que les mails et les textes FLE présentaient un pourcentage moyen d'erreurs de syntaxe plus élevé que les dictées et les résumés, dans un rapport allant jusqu'à 6 entre les dictées et les textes FLE. Cet écart augmente si nous nous en tenons

aux seules erreurs de grammaire, avec une proportion moyenne d'erreurs de syntaxe qui apparaît bien plus élevée dans les textes FLE (27,6%) que dans le reste du corpus (entre 4% et 10,7%). Si l'ANOVA montre effectivement un effet du type de textes ici ($F_{(3,248)}$=7,123 ; p<0,0001), d'après le test PLSD de Fisher, les textes FLE ne sont différents que des dictées (p=0,001) et des résumés (p=0,038). Par ailleurs, les dictées se distinguent également des mails (p<0,0001) en plus des textes FLE.

Catégories d'erreurs	Dictées	Résumés	Mails	Textes FLE	CORPUS
ACCORD	52,4% (36,6)	52,1% (41,7)	59,2% (43,5)	35% (30,2)	52,4% (38,6)
LEXIQUE	33,3% (35,1)	35,3% (39,7)	16,3% (30,5)	32,5% (29,6)	29,9% (34,9)
SYNTAXE	4% (12,1)	8,7% (26,2)	10,7% (27)	27,6% (31)	8% (21)
VERBE	10,2% (22,6)	3,9% (12,8)	13,8% (30,4)	4,9% (9,3)	9,7% (22,9)
Toutes erreurs grammaticales	41,9% (24,4)	37,2% (21,4)	56,4% (34,3)	41,6% (27)	44,3% (27,3)

Tableau 6.8 : Proportion moyenne (ET) de chaque catégorie d'erreurs de grammaire dans chaque type de textes

Catégorie Verbe

Enfin, pour `Verbe`, nous observons toujours une proportion d'erreurs plus importante dans les mails et plus faible dans les résumés, trop faible pour que nous la prenions en compte dans nos tests. Nous avons donc réalisé une analyse de variance sur les trois autres sous-corpus uniquement. L'ANOVA révèle que le pourcentage d'erreurs de verbe dépend ici du type de textes ($F_{(6,244)}$=4,800 ; p=0,002), avec une différence significative entre les mails et les textes FLE (p=0,001).

Synthèse sur les proportions moyennes d'erreurs

Le fait de ne pas prendre en compte les erreurs d'orthographe et de ponctuation influe peu sur l'effet du type de textes sur les pourcentages moyens d'erreurs mis en évidence précédemment (*cf.* p.118).

Comme nous pouvions nous y attendre, les textes FLE ressortent comme plus souvent différents du reste du corpus. Ils présentent notamment une plus forte proportion d'erreurs de lemme, dont l'explication est sans doute à rechercher dans le fait que les scripteurs sont non francophones natifs et donc plus enclins à confondre des mots qu'ils maîtrisent mal.

De même, une moins bonne maîtrise de la syntaxe de la part des apprenants est sans doute à l'origine d'un plus fort taux d'erreurs de cette catégorie dans les textes FLE. Ce taux est élevé aussi dans les mails, dans lesquels nous avons observé des tournures typiques de l'oral. Il est au contraire plus faible dans les dictées, ce qui semble peu surprenant du fait de la nature-même de l'écrit.

Les mails apparaissent également différents au niveau des erreurs d'accord. Nous y avons

observé une sur-représentation des erreurs d'accords verbaux que nous avons expliquée par l'utilisation propre à ce sous-corpus de la deuxième personne du singulier, dont les accords verbaux apparaissent souvent erronés.

Il découle de ces différences que la répartition des erreurs dans les catégories varie en fonction des sous-corpus. La variation ne se fait pas tant sur l'ordre d'importance des catégories que sur leurs proportions. En effet, `Accord` occupe toujours la plus grande part, toujours suivie de `Lexique`, puis alternativement `Syntaxe` et `Verbe`, ou `Verbe` et `Syntaxe`.

Des tests t pour échantillons appariés mis en œuvre au sein de chacun des sous-corpus ont permis de comparer deux à deux les pourcentages moyens des quatre catégories d'erreurs grammaticales, et de mettre en évidence des comportements différents. Pour les dictées, ces tests ont fait ressortir une hétérogénéité des proportions d'erreurs, toutes différentes les unes des autres ($p<0{,}0001$ pour toutes les paires comparées sauf `Syntaxe` et `Verbe` : $p=0{,}005$). Dans les résumés, les tests font émerger deux groupes qui s'opposent, avec `Accord` et `Lexique` d'une part, qui rassemblent la majorité des erreurs grammaticales de ce type de textes (87,4%), et `Syntaxe` et `Verbe` d'autre part (`Acc-Syn`, `Acc-Ver` et `Lex-Ver` : $p<0{,}0001$; `Lex-Syn` : $p=0{,}003$). Du côté des mails, le fort pourcentage d'erreurs d'accord se différencie des trois autres catégories qui sont plutôt homogènes (`Acc-Lex`, `Acc-Syn` et `Acc-Ver` : $p<0{,}0001$), et enfin pour les textes FLE, ce sont les catégories `Accord`, `Lexique` et `Syntaxe` qui sont homogènes face au faible pourcentage d'erreurs de la catégorie `Verbe` (`Ver-Acc` : $p<0{,}0001$; `Ver-Lex` : $p=0{,}001$; `Ver-Syn` : $p=0{,}008$).

Nous pouvons alors synthétiser les résultats que nous avons faits émerger sur les proportions moyennes d'erreurs :

- Catégorie `Accord` : proportion moyenne plus élevée dans les mails et répartition syntagmes nominaux / syntagmes verbaux différente dans les mails (erreurs dans les syntagmes verbaux plus importantes dans les mails).
- Catégorie `Lexique` : proportion moyenne d'erreurs de lemme plus élevée dans les textes FLE.
- Catégorie `Syntaxe` : proportion moyenne plus élevée dans les textes FLE que dans les dictées et résumés, et plus faible dans les dictées que dans les mails et textes FLE.
- Catégorie `Orthographe` : proportion moyenne plus élevée dans les résumés.
- Catégorie `Ponctuation` : proportion moyenne plus faible dans les résumés.

e) Fréquence d'apparition des erreurs

Les différences de distribution des erreurs que nous avons constatées et mises en évidence par des tests statistiques nous confirment que le type de textes a une influence sur les proportions d'erreurs commises dans telle ou telle catégorie relativement aux autres catégories. Mais cela ne permet pas de conclure sur des différences de fréquence d'apparition des erreurs selon les textes. Par exemple, nous verrons que le fait que les textes FLE présentent la proportion d'erreurs d'accord la plus faible du corpus ne signifie pas que ces erreurs apparaissent de manière moins fréquente dans ce type de textes. Afin d'observer et comparer les densités, ou fréquences d'erreurs des différentes catégories dans les quatre sous-corpus, nous avons poursuivi notre analyse statistique en nous intéressant aux densités moyennes d'erreurs préalablement calculées (*cf.* § 6.2.1 p. 109) et présentées dans le tableau 6.9 p. 125.

Les analyses de variance[9] réalisées révèlent que les densités moyennes d'erreurs dans les catégories `Accord`, `Lexique`, `Orthographe` et `Ponctuation` sont dépendantes du sous-corpus. Les calculs n'ont pas été effectués pour `Syntaxe` et `Verbe` dans lesquels les occurrences d'erreurs sont trop peu nombreuses. Après l'application du test *post hoc* PLSD de Fisher, il ressort que les textes FLE présentent des densités d'erreurs généralement plus élevées que les autres textes du corpus, comme nous le verrons dans la suite de ce chapitre.

Catégories d'erreurs	Dictées	Résumés	Mails	Textes FLE	CORPUS
ACCORD	75,8	99	74,1	37,3	72,5
LEXIQUE	140,8	196,1	204,1	37,5	126,6
SYNTAXE	769,2	666,7	384,6	52,6	333,3
VERBE	416,7	1666,7	333,3	333,3	434,8
ORTHOGRAPHE	35,3	36	52,4	16,7	34,7
PONCTUATION	91,7	285,7	81,3	20	75,2
Toutes catégories	**15,8**	**20,6**	**18,3**	**5,4**	**14,5**

Rappelons que les valeurs que nous donnons ici sont les inverses des valeurs utilisées pour les calculs de densité. Nous omettons volontairement d'indiquer les écarts-types correspondant à ces valeurs inverses car ils ne pourraient être calculés qu'approximativement.
Pour information, les densités calculées ($\frac{nb\ d'erreurs}{nb\ de\ mots}$) et leurs écart-types sont indiqués dans le tableau A.10 p. 247 en annexe.

Tableau 6.9 : Densité d'erreurs dans chaque catégorie et chaque type de texte

Catégorie Accord

La catégorie `Accord` est concernée par cette différence entre les densités d'erreurs des textes FLE et des autres sous-corpus (PLSD de Fisher : Textes FLE-Dictées et Textes FLE-Résumés : $p<0{,}0001$; Textes FLE-Mails : $p=0{,}010$). Dans les textes FLE, la densité d'erreur est d'environ une erreur tous les 37,3 mots, soit au moins le double de la densité d'erreurs des autres types de textes, où la densité est d'une erreur tous les 75,8 à 99 mots. Or, nous avons vu précédemment que la proportion d'erreurs d'accord était la plus faible dans les textes FLE. Ce résultat ne signifiait donc pas que les apprenants de FLE font moins d'erreurs d'accord, mais simplement que ces erreurs se trouvent en proportion moindre par rapport aux autres types d'erreurs. Ils font beaucoup plus d'erreurs de manière générale et les accords leur posent peut-être moins de problème qu'aux locuteurs francophones, contrairement au lexique et surtout à la syntaxe. Nous notons également que les mails présentent une densité d'erreurs similaire aux dictées et aux résumés, contrairement à ce que nous avions observé lors de l'examen des proportions d'erreurs.

Au sein de cette catégorie `Accord`, pour la sous-catégorie `Déterminants`, nous obtenons pour les densités un résultat comparable à celui présenté précédemment pour les proportions d'erreurs

9. ANOVA : `Accord` : $F_{(3,248)}=5{,}910$; $p=0{,}001$. `Lexique` : $F_{(3,248)}=14{,}869$; $p<0{,}0001$. `Orthographe` : $F_{(3,248)}=7{,}593$; $p<0{,}0001$. `Ponctuation` : $F_{(3,248)}=23{,}026$; $p<0{,}0001$.

(*cf.* § *Catégorie Accord* p. 120). Nous avions mis en évidence un effet du type de textes sur le pourcentage d'erreurs de déterminant, avec un pourcentage plus élevé pour les textes FLE. Le type de sous-corpus a ainsi également un effet sur les densités d'erreurs d'accord de déterminant (ANOVA : $F_{(3,248)}$=38,301 ; p<0,0001). Avec une erreur tous les 105,3 mots, les textes FLE sont différents des autres types de textes (PLSD de Fisher : p<0,0001 pour les trois paires comparées) dans lesquels le nombre de mots pour une erreur s'étend de 1000 (Résumés) à 1666,7 (Dictées).

Le fait que des textes de francophones non natifs se différencient des autres sur ce type d'erreurs peut laisser penser que c'est précisément le critère de la langue maternelle qui a un effet. Nous pouvons émettre l'hypothèse qu'elles résultent d'une difficulté liée au genre des noms en français qui ne sont pas bien maîtrisés par les scripteurs étrangers, conduisant ces derniers à ne pas choisir le déterminant adéquat pour accompagner un nom. Granger *et al.* [2001] précisent d'ailleurs que « l'analyse [de leur] corpus révèle qu'il y a beaucoup plus d'erreurs de genre (74%) que de nombre (26%) ». Après observation des erreurs d'accord de déterminants commises par des apprenants dans notre corpus, nous constatons en effet que la majorité concernent le genre.

Nous avions également fait ressortir précédemment la particularité des mails concernant les erreurs d'accord sur les verbes, en proportion plus grande dans ce sous-corpus. Les tests statistiques montrent à nouveau un effet du type de textes sur la densité de ces erreurs (ANOVA : $F_{(3,248)}$=37,362 ; p<0,0001), cependant les mails ne se différencient pas de manière aussi nette. Ils présentent effectivement la densité d'erreurs la plus grande (153,8 mots pour une erreur) dans cette catégorie, mais d'après le test PLSD de Fisher, ils ne sont pas différents des textes FLE (256,4 mots pour une erreur ; p=0,258).

Catégorie Lexique

Outre la catégorie `Accord`, les textes FLE sont également différents des autres textes pour la catégorie `Lexique` (PLSD de Fisher : p<0,0001 pour les trois paires comparées), avec une densité d'erreurs comparable à la catégorie `Accord`, soit en moyenne une erreur pour 37,5 mots. Mais pour la catégorie `Lexique`, la différence avec les autres sous-corpus est cette fois-ci beaucoup plus importante. Les moyennes s'échelonnent d'une erreur tous les 140,8 mots dans les dictées, à une erreur tous les 204,1 mots dans les mails. L'analyse des proportions moyennes d'erreurs (*cf.* § *Catégorie Lexique* p. 120) n'avait pas montré de différence des textes FLE par rapport au reste du corpus au niveau de l'ensemble de la catégorie `Lexique`. Il y avait en revanche une différence au niveau de la sous-catégorie `lemme`, avec un pourcentage plus élevé pour les textes FLE. Nous retrouvons cette différence en étudiant les densités. Le type de textes a un effet sur la densité d'erreurs de lemmes (ANOVA : $F_{(3,248)}$=19,240 ; p<0,0001), et les textes FLE se détachent une fois encore du reste du corpus (PLSD de Fisher : p<0,0001 pour les trois paires comparées).

Catégorie Syntaxe

Au niveau de la catégorie `Syntaxe`, seuls les textes FLE offrent les effectifs suffisants pour calculer la densité d'erreurs. Nous observons encore une fois une différence avec le reste du corpus, avec une fréquence d'une erreur pour 52,6 mots (soit 14,5 fois plus que les dictées). Les mails, avec leurs tournures syntaxiques empruntées à l'oral (*cf.* § *Catégories Syntaxe et Verbe* p. 116), présentent la deuxième densité la plus élevée, avec une erreur pour 384,6 mots, soit deux fois plus que les dictées dans lesquelles la fréquence est la plus basse avec seulement une erreur tous les 769,2 mots. Nous retrouvons ainsi pour cette catégorie des résultats proches de ceux obtenus lors de l'analyse des proportions d'erreurs, qui indiquaient une proportion moyenne plus élevée dans les textes FLE, et plus faible dans les dictées.

Catégorie Verbe

En ce qui concerne la catégorie `Verbe`, les densités calculées sont trop faibles pour pouvoir effectuer des comparaisons fiables. Nous observons simplement dans le tableau 6.9 p. 125 que les résumés affichent une densité d'erreurs beaucoup plus faible (une erreur pour 1666,7 mots) que les trois autres types de textes, qui paraissent homogènes (une erreur pour 333,3 à 416,7 mots). Les mails et les textes FLE présentent même des densités identiques. Nous avons observé précédemment (*cf.* § *Catégorie Verbe* p. 123) que les mails possédaient la plus forte proportion d'erreurs dans la catégorie `Verbe`. Avec l'analyse des densités d'erreurs, nous voyons qu'ils ont également, à égalité avec les textes FLE, la plus haute fréquence d'erreurs dans cette catégorie.

Catégories Orthographe et Ponctuation

Enfin, pour les catégories `Orthographe` et `Ponctuation`, les textes FLE sont encore différents des autres types de textes, avec des fréquences d'erreurs plus importantes (PLSD de Fisher : `Orthographe` : Textes FLE-Dictées et Textes FLE-Mails : $p<0{,}0001$; Textes FLE-Résumés : $p=0{,}003$; `Ponctuation` : $p<0{,}0001$ pour les trois paires comparées). L'étendue des densités est cependant moins importante que celles que nous avons pu observer dans d'autres catégories. Pour l'orthographe par exemple, il y a entre 16,7 mots (Textes FLE) et 52,4 mots (Mails) pour une erreur. L'écart est plus grand pour la ponctuation entre les textes FLE et les résumés, avec des valeurs allant respectivement de 20 à 285,7 mots pour une erreur. Les résumés apparaissent significativement différents des autres sous-corpus avec la densité la plus faible (PLSD de Fisher : Résumés-Dictées : $p=0{,}010$; Résumés-Mails : $p=0{,}002$; Résumés-Textes FLE : $p<0{,}0001$), ce qui rejoint le résultat obtenu précédemment (*cf.* § *Catégorie Ponctuation* p. 118) qui donnait les résumés comme ayant la plus faible proportion d'erreurs de ponctuation.

Erreurs grammaticales

Si nous nous focalisons sur les erreurs grammaticales, comme nous l'avons fait pour l'étude des pourcentages moyens d'erreurs (*cf.* § *Focus sur les erreurs grammaticales* p. 122), le nombre de mots entre deux erreurs est multiplié par 2,5 en moyenne [10] dans tous les sous-corpus. Ainsi les densités passent de 15,8 à 41,7 mots pour une erreur dans les dictées, de 20,6 à 57,7 mots dans les résumés, de 18,3 à 41,7 mots dans les mails et de 14,5 à 5,4 mots dans les textes FLE. La différence entre les textes FLE et le reste du corpus est toujours visible et attestée par l'analyse de variance et les tests *post hoc* (ANOVA : $F_{(3,248)}=21{,}565$; $p<0{,}0001$. PLSD de Fisher : $p<0{,}0001$ pour les trois paires comparées)

Synthèse sur les fréquences d'erreurs

L'analyse des densités d'erreurs nous apprend ainsi qu'en moyenne, dans notre corpus, une erreur est commise tous les 14,5 mots, et une erreur grammaticale tous les 36,8 mots. Les textes FLE présentent dans toutes les catégories les fréquences d'erreurs les plus élevées, avec une nuance pour la catégorie `Verbe`, pour laquelle nous n'avons pas pu effectuer de test mais où les textes FLE sont *ex æquo* avec les mails. Ces résultats ne sont pas surprenants au regard des densités d'erreurs globales pour chaque sous-corpus, observées dans le tableau 6.4 p. 113 : les Textes FLE présentent une erreur tous les 5,4 mots, alors que dans les autres sous-corpus la densité n'est que d'une erreur tous les 15,8 à 20,6 mots.

Nous remarquons ainsi une scission entre les textes d'apprenants d'une part, avec une densité moyenne d'une erreur pour 5,4 mots, et les dictées, résumés et mails d'autre part, avec une

10. La densité calculée est divisée par 2,5.

densité moyenne plus faible d'une erreur pour 16,9 mots. Contrairement à ce que nous avions pu observer au niveau des proportions moyennes (*cf.* § *Synthèse sur les proportions moyennes d'erreurs* p. 123), les trois sous-corpus de scripteurs francophones natifs forment un ensemble plutôt homogène en ce qui concerne les fréquences d'erreurs dans les catégories `Accord`, `Lexique`, `Syntaxe` et `Orthographe`. Il se pourrait alors que ces fréquences présentent une certaine stabilité de manière générale, en dépit de contextes de scription différents.

Pour la catégorie `Ponctuation`, les résumés se démarquent de l'ensemble des textes de francophones natifs, ce qui était déjà le cas pour les proportions moyennes d'erreurs. En revanche, pour la catégorie `Verbe`, l'absence d'analyse due au nombre d'occurrences d'erreurs trop faible ne nous permet pas de mettre en évidence une différence ou une similarité entre les sous-corpus.

f) L'homophonie dans les erreurs

Une donnée importante que nous avons évoquée mais que nous n'avons pas encore traitée dans nos analyses statistiques concerne le caractère homophone des erreurs relevées. Au moment de l'annotation de cette information, nous avons choisi de considérer les erreurs portant sur la graphie des mots (`Accord`, `Lexique`, `Verbe` et `Orthographe`) ainsi que sur la syntaxe, mais d'exclure les erreurs de ponctuation, ainsi que les erreurs dont il était difficile de déterminer la forme correcte, et par conséquent le caractère homophone ou non des deux formes [11].

Catégories d'erreurs	Dictées	Résumés	Mails	Textes FLE	CORPUS
ACCORD	91,3% (19,5)	88,3% (26,2)	94,3% (23,6)	58,1% (43,9)	88,7% (25,7)
LEXIQUE	87,3% (29,6)	47,2% (49,9)	78,6% (42,6)	49% (46,7)	76,8% (39,6)
SYNTAXE	0	0	0	0	0
VERBE	88,4% (30,3)	100%	100%	70% (44,7)	89,5% (28,6)
ORTHOGRAPHE	70,4% (33,8)	38,7% (32,8)	65,5% (42,8)	65,3% (28,3)	63,5% (36,5)
Total	**83,9% (19,1)**	**54,8% (28,4)**	**82,3% (30,2)**	**59,5% (29,4)**	**77,5% (26,5)**

Tableau 6.10 : Proportions moyennes des erreurs à caractère homophone

Le premier résultat que nous pouvons observer (tableau 6.10) au niveau du corpus est qu'en moyenne plus des trois quarts des erreurs (77,5%) sont homophones des formes correctes attendues. Cette proportion diffère cependant d'un sous-corpus à l'autre. Dans les résumés, elle représente 54,8% des erreurs, 59,5% dans les textes FLE, 82,3% dans les mails et 83,9% dans les dictées.

11. Par exemple, comment corriger *party* et *alimente* dans la phrase suivante ? : **la groupe B1.2 a organisé une party d'alimente*

Nous remarquons que les résumés et les textes FLE ont des valeurs assez proches, de même que les dictées et les mails. La mise en œuvre d'une analyse de variance sur les proportions moyennes des erreurs à caractère homophone, parmi l'ensemble des erreurs considérées (tout sauf la ponctuation), fait ressortir un effet du type de textes (ANOVA : $F_{(3,241)}=18{,}982$; $p<0{,}0001$). En outre, les tests *post hoc* sont significatifs lorsqu'ils concernent les résumés ou les textes FLE (Textes FLE-Dictées, Résumés-Dictées et Résumés-Mails : $p<0{,}0001$; Textes FLE-Mails : $p=0{,}004$). Il n'y a ainsi pas de différence entre les résumés et les textes FLE, ni entre les dictées et les mails, mais les deux groupes sont bien différents l'un de l'autre.

Cette séparation en deux groupes de textes ne semble pas toujours effective si nous observons le détail des résultats par catégorie, toujours dans le tableau 6.10.

Catégorie Accord

Pour la catégorie `Accord`, le pourcentage d'erreurs concerné par l'homophonie est bien inférieur dans les textes FLE. L'effet du type de sous-corpus est confirmé par l'ANOVA ($F_{(3,169)}=9{,}742$; $p<0{,}0001$), et les textes FLE se démarquent effectivement des autres textes (PLSD de Fisher : $p<0{,}0001$ pour toutes les paires comparées) avec une moyenne de seulement 58,1% d'erreurs homophones, contre au minimum 88,3% dans les textes de scripteurs francophones natifs. Il est probable qu'il s'agisse de la conséquence d'une mauvaise maîtrise des genres chez les apprenants, menant à des erreurs plus souvent non homophones (*`Parce que il-y-a des bonnes acteurs et c'est un histoire ca m'interesse.`).

Catégorie Lexique

Dans les erreurs de lexique, nous observons des pourcentages d'erreurs homophones un peu plus faibles, mais qui dépendent toujours des sous-corpus (ANOVA : $F_{(3,124)}=9{,}046$; $p<0{,}0001$). Nous voyons dans le tableau 6.10 que les résumés et les textes FLE ont des taux moyens proches, légèrement inférieurs à 50%, ce qui signifie qu'un peu moins de la moitié des erreurs de lexique dans ces textes présente un caractère homophone avec les formes correctes. Ce taux augmente nettement dans les mails (78,6%), qui se distinguent par rapport aux résumés (PLSD de Fisher : Mails-Résumés : $p=0{,}031$). Il augmente encore davantage dans les dictées (87,3%), qui se distinguent à la fois des résumés et des textes FLE (Dictées-Résumés : $p<0{,}0001$; Dictées-Textes FLE : $p=0{,}001$).

Catégorie Verbe

La catégorie `Verbe` présente des pourcentages d'erreurs homophones élevés dans tous les sous-corpus, allant de 70% à 100%, et aucun effet du type de textes n'est décelé par l'ANOVA ($F_{(3,50)}=1{,}242$; $p=0{,}304$).

Catégorie Orthographe

Enfin, nous constatons pour les erreurs d'orthographe une différence non négligeable entre les résumés d'une part, avec seulement 38,7% d'erreurs homophones en moyenne, et les autres sous-corpus d'autre part, dont les pourcentages varient entre 65,3% pour les textes FLE et 70,4% pour les dictées. L'ANOVA révèle un effet du type de textes pour cette catégorie d'erreurs ($F_{(3,185)}=8{,}405$; $p<0{,}0001$), avec, sans surprise, une différence significative entre les résumés et le reste du corpus (PLSD de Fisher : Résumés-Dictées : $p<0{,}0001$; Résumés-Mails : $p=0{,}002$; Résumés-Textes FLE : $p=0{,}005$).

Nos analyses montrent que les résumés et les textes FLE ont un comportement différent des autres sous-corpus concernant l'homophonie. Le faible taux d'erreurs homophones des textes FLE pourrait s'expliquer par une prononciation qui serait beaucoup plus souvent erronée chez les apprenants, par rapport aux francophones natifs, conduisant ainsi à la transcription de sons erronés, et donc à des formes non homophones des formes correctes.

> « Les sonorités de la langue [étrangère] étudiée sont [...] perçues sur la base du système de référence phonologique constitué par la langue maternelle. Certains sons sont donc mal entendus, mal interprétés et par conséquent mal reproduits. »
>
> [Billières, 1988, p. 7]

Concernant les résumés, nous nous sommes intéressée aux erreurs de frappe, qui nous ont semblé nombreuses dans ces textes. Nous pensons que ce type d'erreurs conduit plus souvent que les autres erreurs à des formes non homophones et non existantes. Un plus grand nombre de ces erreurs expliquerait alors ce faible taux d'erreurs homophones.

Afin de vérifier cette hypothèse, nous avons identifié et comptabilisé toutes les erreurs résultant très certainement d'un problème à la saisie. Nous avons ensuite recherché un lien entre l'homophonie et le type d'erreurs (de frappe ou non). Nous avons utilisé un test du χ^2 d'indépendance, qui permet de déterminer si deux variables nominales sont indépendantes. Dans notre cas, le χ^2 montre qu'il existe bien un lien entre le type d'erreur (erreur de frappe ou non), et l'homophonie ($\chi^2_{(1)}$=310,640, p<0,0001). Précisément, 22,3% seulement des erreurs de frappe sont homophones des formes correctes, contre 77,1% pour les autres erreurs.

Faisant également l'hypothèse que les erreurs de frappe mènent plus souvent à des mots inexistants qu'à des mots existants, nous avons comparé la densité des erreurs de frappe de la catégorie `Orthographe` (catégorie rassemblant les erreurs de mots inconnus) avec l'ensemble des autres catégories. Un test t pour échantillons appariés confirme notre hypothèse en montrant une différence significative (p<0,0001) de densité d'erreurs de frappe entre la catégorie `Orthographe` (1 erreur pour 145 mots en moyenne) et les autres catégories (1 erreur pour 1000 mots en moyenne). Les erreurs de frappe se concentrent effectivement principalement dans les erreurs d'orthographe.

Enfin, une analyse de variance nous permet de prouver que le type de textes a un effet sur les moyennes de densités d'erreurs de frappe dans la catégorie `Orthographe` (ANOVA : $F_{(3,248)}$=17,51 ; p<0,0001). Les résumés, avec une erreur de frappe tous les 48 mots dans la catégorie `Orthographe`, sont significativement différents des trois autres sous-corpus (PLSD de Fisher : p<0,0001 pour les trois paires comparées) dans lesquelles les densités s'échelonnent de 135 (Textes FLE) à 238 (Mails) mots pour une erreur de frappe.

Notre hypothèse semble vérifiée. D'une part, les résumés présentent une densité d'erreurs de frappe bien supérieure aux autres types de textes, d'autre part, ces erreurs sont souvent non homophones des formes correctes et créent généralement des mots inexistants. Il est ainsi normal que le pourcentage d'homophonie dans les erreurs d'orthographe des résumés soit plus faible que dans le reste du corpus.

Erreurs grammaticales

En nous en tenant aux seules erreurs de grammaire, le pourcentage moyen d'homophonie n'augmente que très légèrement pour atteindre 78,2%. Nous n'effectuons donc pas d'analyses détaillées se focalisant sur ces erreurs.

g) Calculs des distances

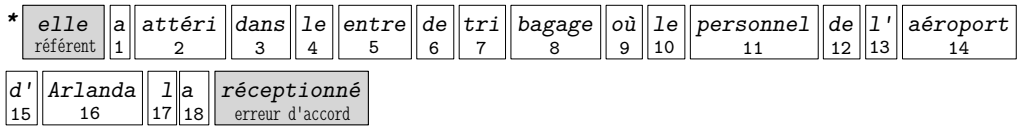
Afin de répondre à notre interrogation sur la part d'erreurs d'accord dans les contextes locaux par rapport aux contextes distants (*cf.* § b).0 p. 115), mais aussi parce que nous avons vu que le scripteur expert peut être sujet à des erreurs de proximité (*cf.* § 5.1.1 p. 77), il nous a semblé intéressant d'étudier, de manière plus générale, l'impact sur les erreurs de la distance séparant deux éléments dont l'un détermine la forme de l'autre (comme par exemple un groupe nominal sujet influe sur la personne à laquelle sera conjuguée le verbe principal). Nous avons donc complété l'annotation de notre corpus en indiquant la distance entre les deux unités concernées dans l'erreur, à savoir l'erreur et son référent.

Distances en mots

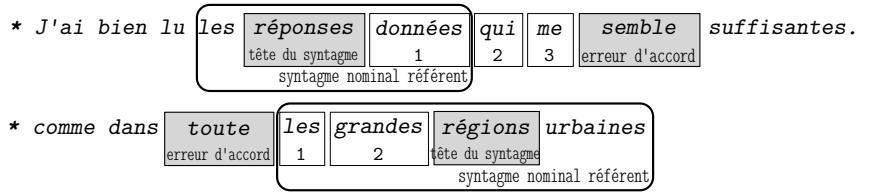
La première annotation que nous avons effectuée concerne le nombre de mots intercalés entre une erreur et son référent. Pour compter les mots, nous avons pris comme base la version erronée de la phrase, sauf en présence de formes mal segmentées, auquel cas nous avons considéré la segmentation correcte pour le décompte. Reprenons par exemple l'énoncé (9) donné p. 102, l'erreur d'accord touchant le participe passé ***réceptionné***, et le pronom ***elle*** que nous avons identifié comme le référent de l'erreur.

(9) *__*elle* a attéri dans le entre de tri bagage où le personnel de l'aéroport d'Arlanda la réceptionné__*

Nous avons indiqué que dans cette phrase, le pronom objet *l'* et l'auxiliaire avoir **a** on été fusionnés de manière erronée en la forme **la**. En considérant la phrase telle qu'elle a été écrite, c'est-à-dire avec la fusion de ces deux mots, nous compterions 17 mots séparant le référent de l'erreur. Cependant, dans ce cas comme dans d'autres cas similaires de mauvaise segmentation, nous avons décidé de compter les éléments comme s'ils étaient séparés. La distance calculée est alors de 18 mots pour notre énoncé, tel que nous le représentons ci-dessous. Nous avons mis en évidence l'erreur et son référent encadrés isolément sur fond gris, en numérotant chaque mot les séparant :



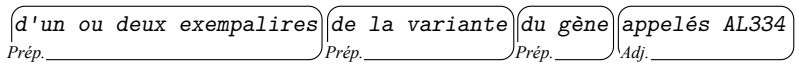
Dans les cas où le référent est constitué de plusieurs mots, comme par exemple un syntagme nominal, nous avons comptabilisé les mots non pas à partir des limites gauche ou droite du syntagme, mais à partir de sa tête lexicale. Ainsi, par exemple, dans les deux extraits ci-dessous où le référent est un syntagme nominal, tous les mots qui se trouvent entre le nom (tête lexicale) et la forme erronée, et non pas uniquement ceux qui sont à l'extérieur du syntagme nominal, sont pris en compte pour déterminer la distance.



Distances en syntagmes

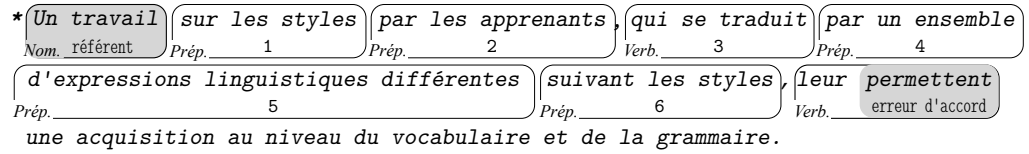
Lorsque les distances sont d'au moins deux mots, nous avons constaté que ces mots forment souvent des unités syntaxiques, du type syntagmes nominaux ou verbaux, propositions relatives, subordonnées, complément du nom, etc. Il nous a alors paru intéressant de calculer la distance entre une erreur et son référent non plus en termes de mots, mais en prenant ces groupes comme unité de mesure.

Nous avons donc également annoté les distances en syntagmes. L'unité que nous considérons pour cette annotation est le syntagme minimal, qui est non récursif. Par exemple, le long syntagme prépositionnel ci-dessous est lui-même constitué de quatre syntagmes minimaux :



Pour alléger notre texte, nous parlerons simplement de syntagmes pour désigner les syntagmes minimaux dans la suite de cette section.

Le calcul des distances en syntagmes s'avère différent de celui des distances en mots. Nous avons effectué les mesures comme illustré dans l'énoncé ci-après :



L'erreur porte sur l'accord du verbe *permettent*, dont le sujet et référent est le syntagme nominal *Un travail*. Nous représentons toujours l'erreur et son référent sur fond gris, mais ce sont cette fois-ci les syntagmes qui sont encadrés et numérotés. Entre l'erreur et le référent, nous avons donc délimité six syntagmes (5 prépositionnels et 1 verbal). Notons que le verbe porteur de l'erreur appartient lui-même à un syntagme verbal que nous n'avons pas inclus à notre calcul de distance. En effet, nous ne prenons pas en compte les syntagmes dont font partie le référent et l'erreur.

Distances non calculées

La dernière ligne des tableaux 6.11 et 6.12 p. 133 montre que dans une grande majorité des cas (77% des erreurs), incluant notamment toutes les erreurs d'orthographe et de ponctuation, il n'était pas pertinent de calculer une distance entre l'erreur et le référent, du fait même de l'absence de référent.

Nous constatons sans surprise qu'il y a peu de distances non calculées dans la catégorie `Accord` (voir tableau 6.11 p. 133). Par définition, pour qu'il y ait accord, il faut au moins deux éléments - dans notre cas le mot erroné et son référent - dont l'un reçoit des marques de genre, de nombre ou de personne données par l'autre. Il peut alors au contraire sembler surprenant de trouver des distances non calculées pour cette catégorie. Il s'agit en grande majorité d'erreurs sur des noms occupant une fonction de complément du nom, après la préposition « de », et pour lesquels un mauvais choix de nombre a été fait. L'extrait suivant, où *compagnie* a été mis au pluriel, en est un exemple :

**[...] John McCain possède déjà 24 animaux de compagnies dont quatre chiens.*

Contrairement à la catégorie `Accord`, la catégorie `Lexique` présente une majorité d'erreurs où la distance n'a pas été calculée. Il n'y a pas de référent pour les erreurs de traits d'union, ni pour la plupart des erreurs de lemmes. Ces dernières sont identifiables principalement sur des critères sémantiques, comme pour *voies* dans l'exemple suivant :

**[...] lors des primaires le caniche a gagné d'un cheveux avec seulement quelques centaines de voies d'avance [...].*

Pour les erreurs de syntaxe, la répartition entre les distances calculées et les non calculées est plus homogène. Parmi les 41% de non calculés se trouvent à nouveaux des cas faisant intervenir non pas un référent mais la dimension sémantique. Pour d'autres erreurs, la structure entière de la phrase est nécessaire pour déceler l'erreur. Dans l'énoncé suivant par exemple, il faut lire la phrase en entier pour conclure à l'oubli probable de *ce sont* (ou une construction équivalente) après *que*. Il se pourrait également que *des combattants* soit le sujet d'un verbe qui manquerait en fin de phrase, mais l'hypothèse précédente nous semble plus vraisemblable. Difficile donc ici de déterminer un référent pour cette erreur. D'autant plus qu'il manque également un complément d'objet au verbe *aimant*.

**Je pense que des combattants qu'on doit respecter en aimant.*

Distances en mots	ACCORD	LEXIQUE	SYNTAXE	VERBE	**Total toutes catégories**
nulle	249 (66,9%)	17 (77,3%)	28 (77,8%)	38 (71,7%)	332 (68,7%)
de 1 à 4 mot(s)	62 (16,7%)	3 (13,6%)	7 (19,4%)	9 (17%)	80 (16,8%)
plus de 4 mots	61 (16,4%)	2 (9,1%)	1 (2,8%)	6 (11,3%)	71 (14,5%)
Total distances calculées	372 (100%)	22 (100%)	36 (100%)	53 (100%)	483 (100%)
calculées	372 (87,3%)	22 (9,6%)	36 (59%)	53 (80,3%)	483 (23%)
non calculées	54 (12,7%)	207 (90,4%)	25 (41%)	13 (19,7%)	1613 (77%)

Tableau 6.11 : Effectifs d'erreurs par distances en mots en fonction de la catégorie d'erreur

Distances en syntagmes	ACCORD	LEXIQUE	SYNTAXE	VERBE	**Total toutes catégories**
même syntagme	224 (60,2%)	17 (77,3%)	31 (86,1%)	41 (77,4%)	313 (64,8%)
nulle	61 (16,4%)	1 (4,5%)	4 (11,1%)	6 (11,3%)	72 (14,9%)
1 syntagme	30 (8,1%)	2 (9,1%)	1 (2,8%)	3 (5,7%)	35 (7,2%)
2 synt. et plus	57 (15,3%)	2 (9,1%)	0	3 (5,7%)	63 (13%)
Total distances calculées	372 (100%)	22 (100%)	36 (100%)	53 (100%)	483 (100%)
calculées	372 (87,3%)	22 (9,6%)	36 (59%)	53 (80,3%)	483 (23%)
non calculées	54 (12,7%)	207 (90,4%)	25 (41%)	13 (19,7%)	1613 (77%)

Tableau 6.12 : Effectifs d'erreurs par distances en syntagmes en fonction de la catégorie d'erreur

Enfin, les cas de distance non annotée dans la catégorie `Verbe` correspondent presque tous

(11/13) à une même erreur, pour laquelle nous n'avons pas identifié de référent permettant à lui seul de déterminer la présence de l'erreur.

Nous laissons de côté toutes ces erreurs où la distance n'a pas été calculée et nous focalisons sur les 23% restants.

Observation des distances calculées

Pour faciliter leur observation, nous avons regroupé les erreurs dans trois ensembles de distances en mots dans le tableau 6.11 p. 133 : distance nulle, distance de 1 à 4 mots, et distance de plus de 4 mots. Dans le tableau 6.12 p. 133, nous les avons regroupé en quatre groupes de distances en syntagmes : même syntagme, nulle, 1 syntagme, 2 syntagmes et plus.

Le premier tableau révèle une forte proportion de distances en mots nulles. En effet, pour environ deux tiers (68,7%) des erreurs, le mot référent précède ou suit directement le mot erroné. Les distances non nulles s'échelonnent ensuite de 1 à 33 mots pour le corpus, avec une médiane à 4. Nous constatons d'ailleurs une répartition assez équilibrée des distances comprises entre 1 et 4 inclus, et des distances supérieures à 4. Ainsi, pour près de la moitié des erreurs ayant une distance non nulle avec leur référent, il y a au moins cinq mots qui séparent les deux éléments. Plus précisément, 85,5% des erreurs sont distantes de quatre mots maximum de leur référent, et seulement 14,5% en sont séparées par au moins cinq mots.

Sans surprise, eu égard au pourcentage élevé de distances en mots nulles, les cas où erreur et référent se situent dans un même syntagme, cumulés aux cas de distances en syntagmes nulles, représentent près de 80% des effectifs calculés. Il ne peut en effet pas y avoir une distance en syntagmes supérieure à zéro pour des erreurs directement voisines de leur référent. Plus précisément, dans 87% de ce cas de figure, l'erreur et le référent font partie du même syntagme.

Les distances en syntagmes nulles correspondent en majorité (59,7%) à des distances en mots également nulles, à une distance d'un mot dans presque un quart des cas (23,6%), et jusqu'à quatre mots. Lorsqu'un seul syntagme sépare ceux de l'erreur et du référent, les distances correspondantes en mots sont principalement de 2 (28,6%), 3 (25,7%) et 4 mots (22,9%), mais s'échelonnent jusqu'à 7.

Quant aux distances de 2 syntagmes et plus, elles concernent presque exclusivement (96,8%) les erreurs éloignées de plus de 4 mots de leur référent. Les valeurs de distances relevées vont de 2 à 6 syntagmes, puis font un bond à 12. Il s'agit d'une valeur extrême correspondant à une erreur présente dans sept dictées où la distance en mots est de 33.

Les tableaux 6.11 et 6.12 p. 133 présentent également la répartition des erreurs dans les différents groupes de distance en fonction de leur catégorie.

Concernant la catégorie `Accord`, nous avions relevé précédemment (*cf.* § *Distribution dans le corpus* p. 115) que le cumul des pourcentages moyens d'erreurs de cette catégorie et de la catégorie `Orthographe` donnait un résultat similaire au taux de 60% d'erreurs détectées par les vérificateurs grammaticaux de seconde génération. Nous avions cependant mentionné le fait que seules les erreurs d'accords locaux étaient détectables par ces outils, mais que nos données ne prenaient pas en compte le critère de distance relative à un accord. Disposant à présent de cette information, il nous est possible de refaire un test portant uniquement sur les erreurs d'orthographe et d'accords locaux. Nous avons pour cela pris comme seuil une distance de quatre mots, considérant ainsi comme local tout accord réalisé entre deux éléments distants de quatre mots au maximum. Nous obtenons alors un pourcentage moyen d'erreurs d'accords locaux de

22,2% (contre 24,9% en incluant les accords distants), que nous ajoutons aux 36,1% d'erreurs d'orthographe. Un test t sur échantillon unique nous permet de comparer la nouvelle valeur obtenue, 58,3%, avec les 60% mentionnés ci-dessus, et nous indique qu'il n'y a toujours pas de différence entre les deux valeurs (p=0,363).

Pour la catégorie `Accord` également, nous avons observé dans nos données que les erreurs des syntagmes nominaux se trouvent majoritairement (89,2%) dans le même syntagme que le référent ou dans deux syntagmes contigus. Il n'est pas surprenant non plus d'observer un grand nombre d'erreurs des syntagmes verbaux (43,8%) avec une distance d'un syntagme ou plus.

Dans la catégorie `Lexique`, les distances n'ont été calculées que pour 9,6% des occurrences, composés des quelques erreurs de coréférence et d'euphonie, ainsi que d'un petit nombre d'erreurs de lemmes. Ces erreurs sont en majorité (77,3%) contiguës à leur référent et dans le même syntagme.

La catégorie `Syntaxe` présente également une majorité d'erreurs dont la distance en mots est nulle (77,8%). Ce sont même 86,1% des erreurs qui sont dans le même syntagme que leur référent.

La catégorie `Verbe` enfin contient un pourcentage élevé de distances calculées (80,3%), parmi lesquelles, encore une fois, la plupart (71,7%) sont des distances nulles. Dans 77,4% des cas, l'erreur et le référent sont dans le même syntagme.

Résumés sur les distances

L'annotation des distances nous a permis de mettre en évidence qu'environ un quart des erreurs est concerné dans notre corpus. Il s'agit des cas où un référent permet de détecter une incohérence grammaticale, principalement des erreurs d'accord et de verbe (respectivement 87,3% et 80,3%), puis des erreurs de syntaxe (60%) et en faible quantité des erreurs de lexique (10%). Les erreurs de ponctuation et d'orthographe n'ont pas de référent et ne sont donc pas concernées.

Le référent est souvent contigu à l'erreur, environ deux fois sur trois, et il n'en est éloigné de plus de quatre mots que dans 14,5% des occurrences. Lorsque cette distance est d'au moins deux mots, ce qui est le cas pour une erreur sur quatre, ces mots forment généralement (76% des cas) entre un et 6 syntagmes, qui s'intercalent entre le syntagme de l'erreur et celui du référent. Il se peut également, avec plusieurs mots de distance, que l'erreur et son référent appartiennent à un même syntagme ou à deux syntagmes contigus. Cette configuration est cependant beaucoup plus fréquente lorsque l'erreur et le référent sont côte à côte ou éloignés d'un mot seulement. Ainsi, lorsque l'erreur et le référent sont contigus, ils appartiennent presque systématiquement (87%) à un même syntagme. Lorsqu'ils sont éloignés d'un mot, la moitié appartient au même syntagme, l'autre moitié appartient à deux syntagmes contigus.

Nous pouvons retenir finalement que l'erreur et son référent sont dans un même syntagme dans deux tiers des cas, et qu'ils sont le plus souvent contigus. Cette prédominance des erreurs dans un contexte local ne préfigure en rien que les scripteurs commettent proportionnellement plus d'erreurs lorsque le référent est proche que lorsqu'il est distant. Il faudrait, pour nous en assurer, connaître le taux d'erreurs par rapport aux formes correctes dans les contextes locaux aussi bien que distants, afin de voir si la distance entre une forme et son référent influe sur le risque de commettre une erreur.Par contre, le fait que les erreurs et leur référent soient fréquemment contigus et dans le même syntagme est une information importante pour la détection d'incohérences grammaticales, qui pourra privilégier les contextes locaux pour mettre en relation erreur et référent.

6.3 Résumé des principaux résultats

Dans ce chapitre concernant l'annotation et l'analyse des erreurs, nous avons tout d'abord défini notre typologie d'erreurs afin d'annoter notre corpus. Nous nous sommes inspirée des travaux de Granger [2007] à partir desquels nous avons construit une typologie adaptée à notre corpus et à notre objectif : nous appuyer sur l'analyse des erreurs du corpus pour modéliser un outil capable de les détecter. L'une des caractéristiques de notre typologie, composée de 6 catégories et 25 sous-catégories, est de faire une distinction entre les erreurs menant à des graphies existantes mais inadaptées au contexte, et les erreurs constituant des mots inexistants et inconnus des lexiques des systèmes informatiques. D'un point de vue informatique, ces dernières erreurs relèvent de la fonction de correction orthographique, qui n'est pas l'objet de notre travail, axé sur la détection d'incohérences grammaticales. Ces erreurs sont rassemblées dans la catégorie `Orthographe` de notre typologie. Nous avons également défini une catégorie propre à la ponctuation, et les catégorie `Accord`, `Lexique`, `Syntaxe` et `Verbe` concernent toutes les erreurs de grammaire.

Une fois la typologie d'erreurs définie, nous avons annoté notre corpus en attribuant une ou plusieurs étiquettes de sous-catégories d'erreurs à chaque incohérence relevée. Notre corpus a alors pu jouer son rôle d'apport (*cf.* § a) p. 64) en nous permettant de procéder à une analyse quantitative. Ces premières analyses des erreurs nous donnent une vue d'ensemble de la répartition et de la fréquence des catégories d'erreurs dans le corpus. Elles nous permettent en outre de dégager quelques spécificités propres à un type d'écrit ou de scripteur.

L'étude des erreurs a fait ressortir une prédominance des erreurs d'orthographe (36%) dans le corpus, et un taux d'erreurs de grammaire d'environ 44%, dont plus de la moitié sont des erreurs d'accord (25% du corpus), majoritairement dans les syntagmes nominaux. Nous retrouvons ici, avec les pourcentages d'erreurs d'orthographe et d'accord, les taux de détection des outils de première et seconde générations, détectant les erreurs d'orthographe pour le premier, complétées des erreurs d'accord pour le second.

Les erreurs touchant au lexique sont également nombreuses ($\approx$ 30%) et sont principalement composées de confusion de lemmes. Les erreurs syntaxiques sont globalement peu nombreuses ($<$10%), de même que les erreurs sur les verbes, dont presque 90% sont des substitutions à caractère homophone entre les formes infinitives, les formes conjuguées et les participes passés.

De manière générale, l'homophonie concerne environ 78% des erreurs de grammaire (77,5% de l'ensemble des erreurs). Cela signifie que près de quatre fois sur cinq, la forme erronée se prononce exactement de la même manière que la forme juste. Les accords, généralement non marqués phonologiquement, sont les premiers concernés par les erreurs à caractère homophone, mais les substitutions de lemmes, en particulier les confusions de mots grammaticaux, et les confusions des finales verbales en /E/ sont également fortement concernées par cette homophonie.

Les moyennes observées pour le corpus sont généralement accompagnées d'écarts-types élevés qui révèlent une hétérogénéité entre les sous-corpus et au sein-même des sous-corpus. Nous avons ainsi relevé des résultats parfois très différents d'un sous-corpus à l'autre concernant les proportions moyennes d'erreurs, à l'exception des dictées qui ne se démarquent jamais de l'ensemble des autres sous-corpus.

En termes de densité d'erreurs, le corpus compte en moyenne une erreur tous les 14,5 mots. Cette densité chute à une erreur tous les 36,8 mots pour les seules erreurs de grammaire, mais augmente dans les écrits d'apprenants. La concentration en erreurs dans les textes FLE est trois

fois plus élevée (une erreur pour 6,4 mots) que dans les textes de francophones natifs, ce qui confirme les résultats de Granger [2007]. Les textes FLE contiennent donc non seulement des erreurs plus nombreuses, mais aussi des erreurs différentes de celles relevées chez les scripteurs natifs. Nous avons relevé, en proportion et densité plus élevées que dans le reste du corpus, des erreurs d'accord de déterminants, des erreurs de lemmes et des erreurs de syntaxe. Les premières s'expliquent sans doute par un défaut de connaissance des genres des mots, les secondes par un manque de vocabulaire et les troisièmes par une acquisition encore imparfaite des règles de construction syntaxique en français. Cette différence des apprenants en FLE pose la question, dans le cadre de la vérification grammaticale automatique, de la prise en compte de la langue maternelle du scripteur, ou de son profil en général (par ex. langue, niveau d'étude) pour personnaliser, dans une certaine mesure, la détection des incohérences et les rétroactions dans un outil de vérification grammaticale.

Les mails présentent également des différences, avec des erreurs d'accords des verbes et d'oublis de mots plus nombreuses que dans les autres sous-corpus. Ces différences trouvent leur explication dans deux spécificités des mails : ils sont d'une part pour beaucoup rédigés à la deuxième personne du singulier avec un oubli fréquent de la flexion de conjugaison correspondante, quand les autres types de textes emploient la troisième personne, et ils présentent d'autre part des tournures syntaxiques typiques de l'oral, avec notamment l'omission du *ne* dans les phrases négatives. Ce dernier point pose la question du rapport à la norme dépendant des situations de scription, avec une certaine liberté qui peut être prise avec le « bon usage » dans certaines situations, notamment les moins formelles. Est-il pertinent de détecter et corriger certaines erreurs dans tous les contextes de rédaction ? Il pourrait ainsi être souhaitable de prendre en compte de l'environnement de la tâche (par ex. type d'écrit, thème, destinataire, etc.) dans un outil de détection d'incohérences.

Les résumés quant à eux présentent des erreurs de frappe en densité beaucoup plus importante que dans le reste du corpus, avec pour conséquence une plus forte proportion d'erreurs d'orthographe, les erreurs de frappe générant plus souvent des mots inexistants. Une autre conséquence se situe au niveau de l'homophonie qui peut exister entre une erreur et la forme correcte correspondante, et que nous avons également annotée dans notre corpus. Dans les résumés, du fait du grand nombre d'erreurs de frappe, qui sont plus souvent non homophones des formes correctes, nous avons observé un taux d'homophonie plus faible que dans les autres types de textes. Ce taux élevé d'erreurs de frappe dans ces textes en particulier peut laisser penser qu'il a pour origine le fait que ces textes n'ont pas été rédigés en vue d'être lus par un destinataire (les résumés étaient destinés à une analyse automatique par un logiciel (*cf.* § 5.1.2 p. 79)), contrairement aux autres textes du corpus, d'où une moins grande importance accordée au « soin » de la saisie et/ou à la correction des coquilles.

Pour finir, une analyse des distances entre les erreurs et leur référent a fait ressortir que deux fois sur trois les deux éléments sont contigus, et qu'ils appartiennent dans ce cas presque systématiquement au même syntagme. Ceci pourrait être pris en compte lors de la recherche d'erreurs d'accord, en considérant prioritairement comme référent possible du mot vérifié les mots appartenant au même syntagme

Ces résultats nous donnent un aperçu de l'importance de certains types d'erreurs en fonction des types de textes, mais ils ne permettent pas à eux seuls d'expliquer la plupart des erreurs ni la manière dont l'humain les détecte. Il est nécessaire pour cela de compléter l'annotation, et notamment de recourir à une analyse interprétative et plus seulement descriptive. Nous avons pour cela élaboré un certains nombre d'hypothèses, en nous appuyant sur les apports de la

psychologie cognitive sur les processus cognitifs mis en œuvre par l'humain dans la production du langage écrit et sa révision.

Troisième partie

Modélisation de la vérification grammaticale

Chapitre 7

Modélisation de la production et de la détection humaine des erreurs

« *Writing is generally viewed as a complex and resource-consuming activity involving various cognitive processes from conceptual planning to the graphic transcription of a written text. A consequence of this complexity is that a writer's attention is regularly shared between the different aspects of the production and in such conditions writers make mistakes of different types.* [1] »

[Largy *et al.*, 2004b, p. 534]

Sommaire

7.1 La production du langage écrit 142
7.1.1 Les processus cognitifs mis en œuvre 142
7.1.2 La production d'erreurs dans le corpus 147
7.2 Révision du langage écrit 160
7.2.1 Le processus de révision 160
7.2.2 Hypothèses sur la manière de détecter une erreur 163
7.3 Conclusion 170

Le chapitre précédent a été l'occasion d'étudier sur un plan essentiellement quantitatif les phénomènes relevés dans notre corpus d'erreurs. Il nous a permis de dégager certaines spécificités des sous-corpus, mais il ne nous renseigne pas sur ce qui a pu conduire à commettre les erreurs, ni sur la manière dont l'humain les détectent après coup, ce qui peut pourtant s'avérer pertinent pour tenter d'améliorer la détection automatique des erreurs. Nous allons donc nous intéresser ici aux processus cognitifs mis en œuvre par le scripteur lors de la production écrite et de sa révision. Nous verrons dans quelle mesure ces processus sollicitent la mémoire de travail et quelles stratégies peuvent être mises en œuvre pour gérer au mieux leur déroulement. Nous émettrons également un certain nombre d'hypothèses quant aux causes possibles des erreurs de notre corpus, liées entre autres au fonctionnement cognitif humain, ainsi que des hypothèses sur la manière dont

1. La rédaction est généralement perçue comme une activité complexe consommatrice de ressources, impliquant des processus cognitifs variés allant de la planification conceptuelle à la transcription graphique d'un texte écrit. Une conséquence de cette complexité est que l'attention du scripteur est régulièrement partagée entre les différents aspects de la production et que dans ces conditions il commet des erreurs de différents types.

les erreurs sont identifiées. Nous verrons enfin que sur certains aspects, un outil de vérification grammaticale automatique pourrait s'inspirer des processus cognitifs impliqués dans la rédaction et la révision.

7.1 La production du langage écrit

La production du langage écrit est une activité cognitivement complexe qui met en jeu différents traitements et processus et sollicite intensément la mémoire de travail. Comme le rappellent Heurley & Garnier [2002] et Heurley [2006], la rédaction d'un texte est un processus stratégique, récursif, dirigé par une hiérarchie de buts et de sous-buts. Nous allons voir qu'elle est décomposable en plusieurs sous-processus (par ex. planification, mise en texte, révision, lecture et compréhension, etc.) interagissant selon une certaine dynamique. Ces processus mobilisent des connaissances déclaratives et procédurales stockées en mémoire à long terme. Ils sont coûteux en ressources cognitives et attentionnelles et dépendent de la capacité de la mémoire de travail du rédacteur. La surcharge cognitive est d'ailleurs l'une des principales sources d'erreurs d'écriture, même si d'autres causes possibles, d'après nos hypothèses, peuvent également entrer en jeu.

7.1.1 Les processus cognitifs mis en œuvre

a) Une production en quatre temps

Le processus de production de texte est généralement décrit, dans la tradition psycholinguistique, comme impliquant quatre niveaux de traitements [Negro, 2003 ; Olive, 2004]. Nous en avons une représentation dans le modèle de Hayes & Flower [1980], précurseurs dans la modélisation de la production du langage écrit (voir figure 7.1 p. 143).

Le premier niveau, de *planification*, consiste à définir les buts de la production et à récupérer les informations en mémoire à long terme et les organiser pour construire un plan, élaborer un message pré-verbal, ou conceptuel.

Les second et troisième niveaux sont rassemblés en un seul niveau dans le modèle de Hayes & Flower [1980], la *mise en texte*, qui permet de transformer le message pré-verbal en contenu linguistique puis graphique. Le second niveau consiste en l'encodage grammatical du message pré-verbal. Dans un premier temps, à un niveau fonctionnel, les représentations abstraites des mots sont récupérées dans le lexique mental, associées à des caractéristiques morphosyntaxiques (fonction syntaxique, flexions de genre, nombre et personne), et organisées dans une structure hiérarchique déterminant les relations entre les mots. Dans un second temps, à un niveau positionnel, les mots sont encodés phonologiquement et positionnés linéairement dans la phrase. Le troisième niveau correspond à la transcription graphique du message.

Enfin, le quatrième niveau est celui de la *révision* du message produit, en comparant le texte aux intentions initiales de l'auteur, afin d'effectuer des corrections ou améliorations, aux niveaux conceptuel et/ou linguistique.

Les processus cognitifs, au centre du modèle de la production d'écrit de Hayes & Flower [1980], interagissent entre eux, mais aussi avec la composante *environnement de la tâche* (voir figure 7.1), constituée des facteurs extérieurs au scripteur. Ces facteurs sont le texte déjà produit et les caractéristiques de la tâche de rédaction, comme le thème de l'écrit ou les destinataires

visés. La tâche de rédaction d'un courrier à un proche sera par exemple très différente de la tâche de rédaction d'un article scientifique.

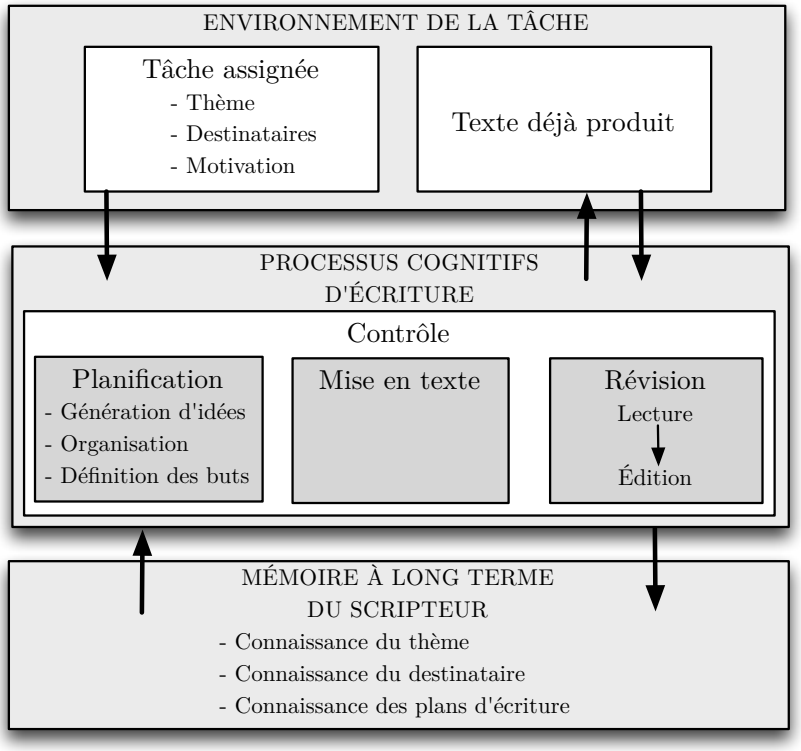


FIGURE 7.1 : Modèle de production d'écrit selon Hayes & Flower [1980], dans sa version clarifiée [Hayes, 1996]

Les processus cognitifs de la rédaction interagissent également avec la *mémoire à long terme* du scripteur dans laquelle se trouvent les différents types de connaissances nécessaires à la production d'écrit, comme les connaissances du sujet traité, des destinataires, mais aussi les connaissances linguistiques.

Ces multiples interactions et le fonctionnement-même des différents sous-processus font de la rédaction un processus jugé complexe et coûteux en ressources attentionnelles. D'après Chanquoy & Alamargot [2002, p. 365], « L'un des problèmes fondamentaux de la production du langage écrit consiste alors en la possibilité de gérer parallèlement et/ou sériellement différents types de traitements, en tenant compte de la capacité limitée de la mémoire de travail ». Il faut alors considérer cette activité plus généralement au sein d'un système cognitif qui comprend notamment la mémoire à long terme et la mémoire de travail [Chanquoy & Alamargot, 2002].

Le modèle de Hayes & Flower [1980], s'il intègre effectivement la mémoire à long terme, ne représente pas la place de la mémoire de travail. Il a cependant par la suite inspiré nombre de modèles, certains décrivant le processus de rédaction de manière générale (pour une revue complète, voir Alamargot & Chanquoy [2001]), d'autres s'attachant à un sous-processus en particulier, notamment le sous-processus de révision (*cf.* § a) p. 160).

Parmi les modèles de la production du langage écrit donnant sa place à la mémoire de travail, nous citerons celui de Kellogg [1996] que nous présentons avec la mémoire de travail dans le point

suivant.

b) Le rôle de la mémoire de travail

Les composants de la mémoire de travail

Les activités complexes telles que la production de texte écrit bien sûr, mais aussi la lecture, la résolution de problèmes ou encore l'apprentissage, dépendent fonctionnellement de la mémoire de travail [Piolat, 2004].

> « *Working memory is aimed at explaining how information is temporarily stored and processed during the realization of cognitive activities, and in particular in complex cognitive activities such as learning, and language production and comprehension.* [2] »
> [Olive, 2004, p. 34]

Les modèles de la production d'écrits intégrant la mémoire de travail s'appuient généralement sur le modèle élaboré par Baddeley, qui décrit la mémoire de travail comme « *a system for the temporary holding and manipulation of information during the performance of a range of cognitive tasks such as comprehension, learning, and reasoning.* [3] » [Baddeley, 1986, p. 34]

Selon ce modèle (voir figure 7.2 p. 145), la mémoire de travail est composée d'un administrateur central et de sous-systèmes esclaves. L'administrateur central est considéré comme la composante attentionnelle de la mémoire de travail. Il organise la récupération d'informations en mémoire à long terme et l'attribution des ressources cognitives aux traitements en cours [Piolat, 2004]. Il coordonne également les systèmes esclaves : la boucle phonologique, le calepin visuo-spatial et le *buffer* épisodique, ajouté par [Baddeley, 2000] afin de répondre aux limites de son premier modèle.

La boucle phonologique et le calepin visuo-spatial ont pour rôle de maintenir temporairement en mémoire « les informations récupérées en mémoire à long terme ou dans l'environnement, mais aussi des représentations transitoires issues des traitements en cours » [Olive & Piolat, 2005, p. 376]. La boucle phonologique permet le stockage et le traitement des informations verbales. Elle est composée d'une unité de stockage passif et d'un système actif de « contrôle articulatoire ». La première conserve des informations verbales, sous forme phonologique, pendant deux secondes au maximum. Le contrôle articulatoire permet alors, d'une part, de rafraîchir les informations pour qu'elles soient conservées plus longtemps, via un processus de répétitions subvocales, aussi appelé « langage intérieur » [Baddeley, 1992], d'autre part d'envoyer de nouvelles informations dans l'unité de stockage, en procédant à leur encodage phonologique lors de la lecture.

Le calepin visuo-spatial, comme son nom l'indique, est spécialisé dans le maintien et le traitement d'informations visuelles et spatiales, et fonctionnerait sur un principe proche de la boucle phonologique. Une unité de stockage conserverait un court moment des informations visuelles et spatiales, tandis qu'un système de rafraîchissement de ces informations permettrait de les maintenir plus longuement. Cette composante de la mémoire de travail est cependant moins étudiée que la boucle phonologique, et son fonctionnement reste par conséquent encore flou.

2. La mémoire de travail vise à expliquer comment l'information est temporairement stockée et traitée durant la réalisation d'activités cognitives, et en particulier au cours d'activités cognitives complexes telles que l'apprentissage, la production et la compréhension du langage.

3. un système pour le maintien temporaire et la manipulation d'informations pendant la réalisation d'un ensemble de tâches cognitives comme la compréhension, l'apprentissage et le raisonnement.

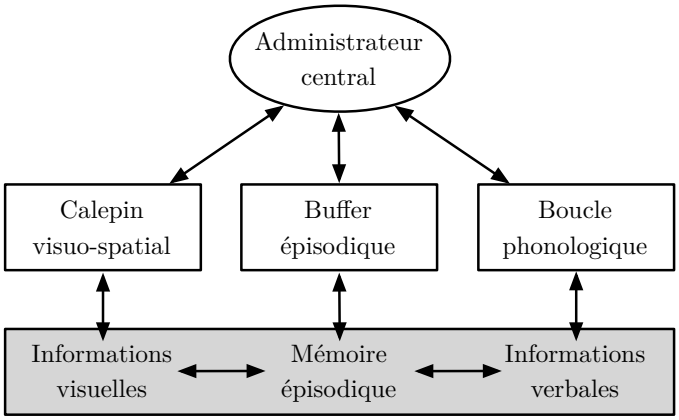


FIGURE 7.2 : Les composants de la mémoire de travail d'après Baddeley [2000]. (En gris : les composants « cristallisés » de la mémoire à long terme ; en blanc : les composants flexibles de la mémoire de travail)

Enfin, le *buffer* épisodique constitue lui aussi un système de stockage temporaire à capacité limitée. Il permet de lier des informations multimodales en provenance de la mémoire à long terme et des autres composants, et de les fédérer en une représentation unitaire [Baddeley, 2000]. Ce *buffer* épisodique constitue ainsi une interface temporaire entre les deux systèmes esclaves et la mémoire à long terme.

S'il existe d'autres conceptions de la mémoire de travail (pour une revue voir [Gaonac'h & Larigauderie, 2000]), le principe-même d'un système cognitif à capacité limitée n'est pas remis en cause. Il est généralement admis que les performances en lecture ou en production d'écrit, pour ne citer que ces activités, sont dépendantes des limites des capacités cognitives du lecteur ou scripteur, et donc de la capacité de la mémoire de travail, quelle que soit sa représentation.

Les composants impliqués dans la production d'écrit

L'un des modèles le plus souvent présenté dans la littérature à avoir formalisé les relations entre la mémoire de travail et le processus d'écriture est le modèle de la production d'écrit de Kellogg [1996] (voir figure 7.3 p. 146). Il définit les relations entre les différents sous-processus impliqués dans la rédaction et les différents composants de la mémoire de travail [4] . La *formulation*, qui correspond aux deux premiers niveaux de traitements dans la rédaction (planification et encodage), requiert les trois composants de la mémoire de travail. Plus exactement, la *planification* et la *traduction* (encodage) font toutes deux appel à l'administrateur central, en plus du calepin visuo-spatial pour la première (les idées pouvant être récupérées sous forme d'images mentales), et de la boucle phonologique pour la seconde. Le processus d'*exécution* (transcription graphique) nécessite l'administrateur central pour la programmation, mais ne solliciterait pas la mémoire de travail pour l'exécution graphique, lorqsque cette dernière est automatisée. Enfin, le sous-processus de lecture du système de contrôle (révision) fait appel à l'administrateur central et à la boucle phonologique, et le sous-processus d'édition requiert l'administrateur central uniquement [Chanquoy & Alamargot, 2002].

4. Le modèle de Kellogg [1996] est antérieur au modèle de Baddeley [2000] qui intègre le *buffer* épisodique. Il ne représente donc pas le rôle de ce composant dans la production d'écrit.

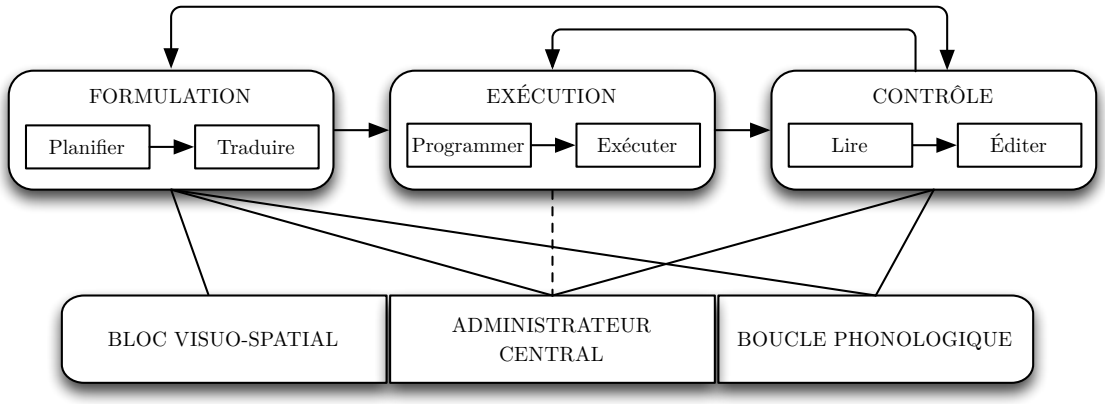


FIGURE 7.3 : Modèle de production d'écrit d'après Kellogg [1996] (notre traduction)

L'automatisation

Afin de soulager la mémoire de travail et pallier ses capacités limitées, l'humain a recours à l'automatisation d'un certain nombre de procédures, celles-ci ne requérant dès lors plus ou très peu de ressources attentionnelles, qui peuvent alors être allouées à d'autres processus en cours. Ces automatismes s'acquièrent avec l'expertise en rédaction et en lecture. La réalisation des accords est l'une de ces procédures expertes. Elle a notamment été étudiée par Fayol & Largy [1992] ou Fayol *et al.* [1994]. Selon eux, les adultes expérimentés réaliseraient les accords sujet-verbe de manière automatique, sur un principe de proximité, en accordant le verbe avec le nom ou pronom le précédant directement. Cette procédure d'accord très rapide et peu coûteuse, qui fonctionne dans la plupart des cas, a en contrepartie l'inconvénient de conduire parfois à des erreurs d'accord de proximité, dans des configurations du type « Nom1 de Nom2 Verbe » ou « Pronom1 Pronom2 Verbe », où les deux noms ou pronoms possèdent un nombre différent, et où le verbe est alors accordé à tort avec le Nom2 ou le Pronom2. Le fonctionnement cognitif peut ainsi le conduire à commettre des erreurs, appelées erreurs d'experts, bien qu'il n'explique pas à lui seul toutes les erreurs.

c) En résumé

L'étude des processus cognitifs impliqués dans la rédaction nous a permis d'envisager quelques pistes de causes d'erreurs possibles. Du modèle de la production d'écrits de Hayes & Flower [1980], nous retenons que l'activité de rédaction est guidée et influencée par des facteurs extérieurs au scripteur, dans l'environnement de la tâche. Ces facteurs peuvent engendrer des erreurs, de même que les différentes connaissances du scripteur sur lesquelles repose l'activité.

Nous retenons également que la mémoire de travail joue un rôle central dans la production de textes, mais que ses ressources sont limitées et partagées entre les différents sous-processus d'écriture (planification, encodage, transcription), ce qui la rend propice à des erreurs d'inattention ou de surcharge cognitive. L'automatisation de certaines procédures (comme les accords) contribue à préserver des ressources utiles à d'autres procédures, mais contribuent également à la production d'erreurs dites « d'expert ».

Nous disposons donc ici de premières hypothèses quant à l'origine des erreurs d'écriture, que

nous pouvons à présent tenter de compléter par l'observation des données de notre corpus.

7.1.2 La production d'erreurs dans le corpus

À partir de ces grandes lignes sur les mécanismes cognitifs et les composants en jeu dans la production du langage écrit, nous avons voulu compléter l'analyse des erreurs de notre corpus et tenter de déterminer le rôle du système cognitif et d'autres facteurs externes au scripteur dans la production d'erreurs. Dans ce but nous avons complété l'annotation du corpus en ajoutant à chaque erreur un attribut spécifiant la cause hypothétique de sa survenue. Il n'est bien sûr pas possible de déterminer de façon certaine pourquoi telle erreur a été commise, mais des hypothèses sur son origine peuvent être inférées du contexte syntaxique, de la situation de scription, de la langue maternelle du scripteur, et même de la configuration du clavier.

Par exemple, l'erreur de genre de l'extrait *`dans le vidéo` ayant été commise par un étudiant de FLE, elle est vraisemblablement due à l'ignorance du genre du nom `vidéo` par l'étudiant. En revanche, un francophone natif qui écrit *`déposer se valise` au lieu de `déposer sa valise`, dans le cadre d'une dictée, a sans doute été victime d'une coquille du fait du temps de rédaction limité et de la proximité des touches « A » et « E ».

Pour déterminer de façon plus précise pourquoi telles ou telles erreurs ont été commises, il aurait été intéressant de pouvoir interroger les auteurs des erreurs eux-mêmes. C'est la démarche adoptée par Jaffré [2003] par le biais des entretiens métagraphiques, qui « fournissent des indices externes du raisonnement métalinguistique produit lors de l'écriture » [Jaffré, 2003, p. 63]. Cette méthode dite de « protocole verbal » est répandue mais nous n'avons trouvé aucune étude menée chez les scripteurs adultes au sujet de leurs choix graphiques. Nous pouvons par ailleurs difficilement nous référer aux études sur les apprenants pour appuyer ou guider nos hypothèses sur les causes d'erreurs. En effet, les travaux sur les apprentis scripteurs et l'acquisition de la morphographie flexionnelle par exemple (voir entre autres Fayol & Got [1991] pour les accords sujet-verbe, Largy *et al.* [2005] pour les accords en nombre) montrent que les enfants ne mobilisent pas toujours les mêmes procédures pour orthographier, et notamment réaliser les accords, que celles utilisées en fin d'acquisition.

Nous nous appuierons donc pour beaucoup sur notre étude des modèles de production d'écrits pour élaborer nos hypothèses sur les causes des erreurs, hypothèses qui pourront, dans une certaine mesure puisqu'il ne s'agit que d'hypothèses, contribuer à la modélisation d'un outil pour repérer lesdites erreurs.

a) Hypothèses sur les causes possibles d'erreurs

Sans les présenter clairement de la sorte, nous avons d'ores et déjà abordé des facteurs pouvant potentiellement être la source d'erreurs. Les types de documents et les conditions dans lesquelles ils sont rédigés sont un premier exemple, qui se rapporte à la composante « environnement de la tâche » dans le modèle de Hayes & Flower [1980]. Nous avons ainsi relevé des erreurs de syntaxe spécifiques de notre sous-corpus de mails, que nous avons rapproché d'un mode de conversation écrite avec des tournures typiques de l'oral (*cf.* p. 117).

Les caractéristiques des scripteurs, leurs connaissances ou compétences sont un second exemple de causes possibles d'erreurs. Nous venons de voir en effet que le fonctionnement cognitif du scripteur, et notamment l'automatisation de certains processus, pouvait le conduire à commettre des

erreurs.

Les erreurs peuvent enfin trouver leur origine directement dans le texte. La distance séparant deux éléments liés morphologiquement, ou encore l'homophonie (*cf.* § f) p. 128) sont ainsi un troisième exemple de causes possibles d'erreurs. Ces causes pourraient être situées, dans le modèle de Hayes & Flower [1980], au niveau du texte produit dans l'environnement de la tâche, ou bien au niveau du sous-processus de mise en texte, et plus précisément de l'encodage.

Nous appuyant sur ces exemples, nous avons organisé nos hypothèses sur les causes d'erreurs selon trois ensembles :

1. Erreurs liées à la situation de scription : pour les dictées notamment, le temps limité pour écrire ou la manière dont le texte a été dicté ou entendu peuvent entrer en ligne de compte. Cet ensemble englobe au sens large tout l'environnement de rédaction, incluant le clavier et les possibles fautes de frappe par exemple.
2. Erreurs liées au scripteur : il peut s'agir de facteurs attentionnels, de ressources cognitives limitées, ou encore de facteurs liés aux connaissances que le scripteur a de la langue.
3. Erreurs liées au texte : le texte en lui-même, par la structure d'une phrase, le vocabulaire employé, etc. peut être source d'erreurs. Nous avons rassemblé ces causes en trois groupes avec comme critères l'homophonie, la position ou l'identification du référent, et l'influence d'un élément du texte proche de l'erreur.

Nous avons ainsi établi un classement en ensembles et sous-ensembles de causes d'erreurs, présenté dans le tableau 7.1 p. 149.

Dans notre annotation, nous avons essayé de nous en tenir aux principales raisons de la survenue de chaque erreur, telles que définies dans notre classement, tout en gardant à l'esprit que les erreurs sont souvent le résultat d'une conjonction de facteurs. Par exemple, dans le cas de deux lemmes homophones confondus (par ex. *`*en therme d'ecologie`), nous avons mis en cause le texte, et plus précisément le lexique, mais il s'agit d'une prise de position discutable puisque le lexique est généralement (à l'exception des dictées dans notre corpus) choisi par le scripteur, qui pourrait alors être considéré comme une cause de ce type d'erreurs.

Nous pourrions également discuter du fait que les erreurs sont forcément toutes causées par le scripteur puisque c'est lui qui les écrit. Ce n'est pas tout à fait vrai dans le cas de rédaction via un clavier, sur un ordinateur. Le périphérique peut être source d'erreurs, avec une touche qui fonctionnerait mal par exemple, ou encore l'utilisation d'un clavier dans une configuration de touches non familière (par ex. AZERTY *vs.* QWERTY), qui pourrait concerner notamment les scripteurs étrangers. Il en va de même pour le système informatique, dont un dysfonctionnement peut déboucher sur une erreur. Nous n'avons pas identifié d'erreur potentiellement causée par le système. En revanche, nous avons rencontré beaucoup d'erreurs sans doute consécutives à l'utilisation du clavier dans la catégorie `Orthographe`, erreurs que nous n'abordons pas dans cette section. Nous ne traitons que les erreurs de grammaire, parmi lesquelles nous avons souvent considéré le scripteur comme cause implicite de l'erreur, ne l'annotant donc pas comme tel. Nous avons privilégié des causes contextuelles qui ne dépendent pas du scripteur, mais qui influent sur sa performance.

De même, l'homophonie concerne en moyenne 78,2% des erreurs (*cf.* § f) p. 128) de grammaire, mais nous n'en avons fait une cause d'erreurs que dans certains cas, excluant les erreurs d'accord. Nous avons réservé cette annotation aux occurrences impliquant une substitution de lemme, ou bien un changement de mode verbal, et pour lesquelles nous pensons que le critère de l'homophonie joue un rôle plus marqué que pour les erreurs d'accord. Pour ces dernières en effet,

			Description	**Exemple**
Situation				
	Dictée		Texte mal dicté ou mal compris	**un jugement qui tendrait a affirmer dire que non.* **ces bananes folatées* [frelatées]
	Limite de temps		Écriture dans la précipitation	**pendantr un an __ recu __ methadone* **va retouvué sa liberté aptres 3an de desintoxication*
	Conversation écrite		Utilisation délibérée de tournure typique de l'oral dans les mails	**Ca __ marche pas* **j'__ai pas d'idée immédiate*
	Clavier		Erreurs de frappe	**Les enseignants de lagues qui laborent des solutions didactiques*
Humain				
	Connaissances	Méconnaissances de l'orthographe	Méconnaissance de la graphie d'un mot dans un certain contexte	**dommage__ et intérêt__* **quarante deux milles personnes*
		Méconnaissances du lexique	Méconnaissance du genre d'un nom, choix du lexique, etc.	**Les effluves échappées* **C'était une party très amuzante*
		Maîtrise de la langue	Méconnaissance de règle syntaxique ou morphologique	**On fesait la bataille* **les soldats se quittent leurs familles*
		Omission	Omission a priori volontaire d'un accent, d'un accord, etc.	**certains hommes sont sujet_ a des relation_ ouleuse_* **Les extensions sont peut_etre un phenomene de mode*
	Surcharge cognitive et/ou inattention	Automatisation	Application automatique de règle à mauvais escient	**les américains ont déjà choisis* **elle a atterrit*
		Oubli	Oubli d'un accord	**tu filtre__ sur la colonne 1* **présente chez 4 homme__ sur 10*
		Coquille	Oubli ou doublement d'un mot, substitution de lettre	**qui permet de d'unifier les sorties* **au sein d'uns plate-forme*
		Sémantique	Non prise en compte du sens ou redondance sémantique du pluriel	**les maîtres de chien dit «Dangereux»* **des dizaines de stand__*
Texte				
	Homophonie	Homophonie grammaticale	Confusion de mots grammaticaux ou de formes verbales homophones	**L'oral est l'écrit sont concernés* **pour le dompté*
		Homophonie lexicale	Confusion de lemmes homophones	**seulement quelques centaines de voies d'avance,* **avoir un coup important*
		Homophonie de syntagmes	Confusion de deux groupes de mots homophones	**un jugement qui tendrait a affirmé que non...* **Les fondateurs [...] on présentait*
	Position du référent	Distant	Référent éloigné du mot à accorder	**ce sont les oiseaux occupants un position dominante qui dorlote_ leur subalternes*
		Postposé	Référent postposé au mot à accorder	**la potentiel__ futur__ première famille* **C'est à celà que serve__ d'ailleurs les sommes md5*
		Ambigu	Référent non facilement identifiable, plusieurs possibles	**les hommes porteurs d'un ou deux exempalires de la variante du gène appelés AL 334*
	Influence d'un élément du texte	Accord de proximité	Accord influencé par un mot plus proche	**le langage des signes américains* **ceux qui ont misé sur l'animal domestique prospère__*
		Fréquence de la graphie	Influence d'une graphie plus fréquente que celle attendue	**La personne âgées* **après l'éléction présidentielles*
		Influence autre	Influence d'un genre, nombre, d'une forme verbale, etc.	**je ne me souviens plus avoir créer* **qui ont déposé une plainte en référée*

Tableau 7.1 : Classement des causes possibles d'erreurs de grammaire

ne pas accorder l'élément cible revient à ne pas lui transmettre les traits de genre, de nombre et/ou de personne de l'élément source (ou référent), et ainsi à ne pas marquer la relation entre les deux items. En revanche, lorsqu'un lemme ou un mode verbal sont remplacés par un homophone (ou un mot phonologiquement très proche), plus qu'une question de transmission de traits et de relation entre deux éléments, c'est une différence conceptuelle entre le terme attendu et son homophone qui résulte de la substitution, et l'analyse morphosyntaxique de la phrase ou bien son sens s'en trouvent altérés, ce qui n'est pas le cas, ou alors dans une moindre mesure, lors d'un accord erroné.

S'agissant des erreurs d'accord, plusieurs causes s'appliquent à ce type d'erreurs dans notre classement. Nous avons vu qu'il s'agit des erreurs de grammaire les plus fréquentes (*cf.* § d) p. 122) et les recherches sur ce sujet sont nombreuses.

Sur les accords sujet-verbe, les expériences menées par Fayol & Largy [1992] avec des enfants et des adultes montrent qu'il existe effectivement plusieurs raisons possibles à la survenue de mêmes erreurs. Les auteurs font état de quatre niveaux potentiels d'erreurs [Fayol & Largy, 1992, p. 83] :

1. « la méconnaissance complète (déclarative et procédurale) d'une règle. »
2. « la seule connaissance déclarative d'un énoncé sans capacité à mettre en œuvre la procédure correspondante. »
3. « une gestion cognitivement trop coûteuse pour que le sujet parvienne à appliquer systématiquement et régulièrement une procédure ».
4. « l'application automatique de procédures d'accord alors même que certaines conditions restrictives sont remplies, qui devraient conduire à une gestion contrôlée. »

Les deux premiers niveaux correspondent, selon Fayol & Largy [1992], aux erreurs d'apprenants et « conduisent à la mise en œuvre de deux méthodes didactiques : faire mémoriser les règles, pour assurer la connaissance déclarative ; faire réaliser des exercices pour induire la maîtrise procédurale. » Nous avons dans notre classement des causes d'erreurs de cet ordre, liées aux connaissances, mais pas uniquement pour les accords. Nos scripteurs étant adultes, nous considérons qu'ils disposent des connaissances déclaratives et procédurales de réalisation des accords, à l'exception des apprenants de FLE, pour lesquels nous supposons que ces connaissances sont encore incomplètes. Nous avons d'ailleurs défini, pour ces scripteurs non francophones natifs, une cause d'erreurs impliquant la non maîtrise de plusieurs aspects de la langue (le genre des noms, la construction syntaxique, la morphologie verbale, etc.) propre à ce type de scripteurs. Pour les francophones natifs, nous avons imputé des erreurs à un manque de connaissances, mais non procédurales. Il s'agit davantage de connaissances liées aux mots en eux-mêmes, à la manière dont ils s'orthographient selon le contexte, dans une locution par exemple (**dommage et intérêt*), indépendamment des questions d'accord à proprement parler. Il peut s'agir également de la méconnaissance du genre d'un mot peu fréquent, etc.

Le troisième niveau potentiel d'erreurs de Fayol & Largy [1992] concerne les adultes aussi bien que les enfants possédant une connaissance procédurale des règles d'accord sujet-verbe à appliquer. Quant au dernier niveau, il s'agit de l'automatisation des procédures d'accord propre aux experts, réduisant le temps et l'attention consacrés à l'accord, mais conduisant à des erreurs difficilement évitables. Nous retrouvons ces deux niveaux dans notre classement, avec le facteur de surcharge cognitive, parfois induit par un temps limité pour la rédaction, et conduisant à des automatisations génératrices d'erreurs d'accord par exemple, ou à des erreurs d'inattention. Dans de nombreux cas cependant, il est difficile de déterminer si une erreur est provoquée par

un problème d'inattention ou de surcharge cognitive. Nous avons donc préféré en faire un sous-ensemble unique d'erreurs.

Les expériences de Fayol & Largy [1992] abordent également la question des accords de proximité, dans les configurations « Nom1 de Nom2 Verbe » ou « Pronom1 Pronom2 Verbe ». Notre corpus contient des erreurs de ce type :

*`Très souvent, la correction des exercices se limitent à comparer la forme de la réponse attendue avec celle de la réponse donnée.`

Selon Fayol & Got [1991], les erreurs d'accord verbal de proximité seraient favorisées par le non-marquage de l'accord à l'oral, privant le scripteur d'indices phonologiques pour la transcription graphémique, ainsi que par la surcharge cognitive inhérente à la conduite d'une seconde tâche en parallèle. En ce qui concerne les résumés et les dictées de notre corpus, il est probable que la surcharge cognitive résultant des contextes de rédaction ait joué un rôle important dans la survenue d'erreurs. Pour les dictées, d'une part le scripteur était limité dans le temps, d'autre part il a souvent dû écouter le nouvel extrait à transcrire tandis qu'il était en train de finir de saisir le précédent. Pour les résumés, c'est le fait de devoir faire des retours sur le texte à résumer, par une récupération en mémoire ou par plusieurs relectures, qui constitue sans doute une seconde tâche en parallèle. Nous pensons que cette surcharge cognitive, associée à l'absence fréquente de marquage phonologique des accords, favorise non seulement les erreurs d'accord de proximité [Fayol & Got, 1991], mais aussi les erreurs d'oubli de l'accord ou encore de généralisation de l'accord du participe passé après l'auxiliaire *avoir*.

Sur l'omission de la marque graphique du pluriel dans les accords, que nous rencontrons fréquemment dans notre corpus, Lucci & Millet [1994, p. 76] émettent l'hypothèse dans certains cas d'un « abandon d'une redondance de type « sémantique » [...] le pluriel étant clairement véhiculé par des éléments du lexique ». Nous avons inclus ce facteur de redondance sémantique dans notre classement, pour des erreurs telles que par exemple : *`beaucoup d'accident`, *`des gammes de produit de beauté`.

Concernant les erreurs d'accord avec le sujet du participe passé après l'auxiliaire *avoir*, les études en acquisition de l'orthographe avancent plusieurs hypothèses [Brissaud *et al.*, 2006 ; Fayol & Pacton, 2006], comme la mise en oeuvre délibérée d'une procédure d'accord à mauvais escient, la récupération en mémoire de séquences fléchies fréquemment rencontrées, ou encore la mise en place d'un schéma général du type pluriel-pluriel. Ces hypothèses ne concernent cependant que les enfants et ne peuvent être étendues aux adultes ayant terminé la période d'apprentissage de l'écrit. Nous pensons néanmoins, dans le cadre de notre corpus, que les accords entre le sujet et le participe passé après *avoir* sont réalisés à tort, comme nous l'avons évoqué plus haut, du fait d'une surcharge cognitive liée notamment au temps limité (l'intégralité de ces erreurs se trouve dans les dictées), induisant une application automatique de l'accord selon un schéma pluriel-pluriel, la procédure étant favorisée par l'absence de marquage phonologique de l'accord.

Franck & Hupet [2002] abordent également la question des erreurs de proximité, appelées aussi erreurs d'attraction, pour les accords grammaticaux, mais ils présentent également une seconde cause d'erreur.

> « Le second type d'erreurs consiste à considérer, pour accorder l'élément cible, non pas la marque grammaticale de l'élément source, mais bien la représentation conceptuelle qu'en a le locuteur. Les exemples 2) et 3) illustrent ces accords dits sylleptiques (Grevisse, 1993) ou plus simplement conceptuels.
> *2) *La proportions d'étudiants étrangers sont en hausse*

*3) *Quelle sera l'ordre de grandeur de cette augmentation ?* »*

[Franck & Hupet, 2002, p. 62]

Nous n'avons relevé aucune erreur de ce type dans notre corpus. Cependant nous avons constaté que les accords pouvaient être influencés par d'autres causes que la proximité d'un élément distracteur de genre et/ou de nombre différent(s) de l'élément référent de l'accord. Largy *et al.* [2005, p. 348] avancent par exemple que « la sélection du morphème flexionnel verbal -nt [est] sensible à la survenue d'un item pluriel juste avant le verbe mais aussi aux dimensions sémantiques et phonologiques de la production écrite ». Ils indiquent également que « les performances de l'expert sont hautement sensibles à la fréquence en langue des items ou des configurations sur lesquels il doit opérer ». Ce sont autant de facteurs d'erreurs dont nous pensons avoir trouvé des manifestations dans notre corpus et dont nous avons donc tenu compte dans notre classement. Nous avons en effet déjà indiqué que nous avions de nombreuses occurrences d'accord de proximité (ex. 10), mais nous avons également quelques occurrences d'influence de la dimension phonologique (ex. 11) ou encore d'influence de la fréquence d'une graphie *`L'arrêté municipale`, fréquence que nous avons pu vérifier grâce à la base Lexique [New, 2006]. Nous avons également relevé d'autres causes d'interférence avec la réalisation correcte d'un accord, comme par exemple la prédominance du pluriel dans l'environnement du mot à accorder au singulier (ex. 12), ou l'influence d'une graphie en -s évocatrice du pluriel (ex. 13 et 14).

(10) *`Et chacune de ces discplines doivent pouvoir travailelr ensemble.`

(11) *`[...] proposer une plate-forme qui permettent de construire des systèmes didactqiues`

(12) *`repérer les erreurs usuelles des apprenants pour proposer une corrections des erreurs corantes`

(13) *`Tarbes étaient d'ailleurs composé de plusieurs bourgs [...]`

(14) *`le populaire caniche a devancé d'un museau le moins connus terrier irlandais`

Enfin, toujours sur la thématique des accords, nous avons considéré comme cause d'erreurs potentielle la position et/ou l'identification du référent. Nous avons notamment émis l'hypothèse que la distance entre le référent et le mot à accorder pouvait susciter des erreurs. Nous pensons en effet que plus le référent est éloigné du mot à accorder, plus il nécessite de ressources cognitives pour ne pas être oublié pour réaliser l'accord, et plus il risque également d'être court-circuité par un élément plus proche avec lequel l'accord sera effectué par proximité. Il peut en outre y avoir une ambiguïté dans l'identification du référent, parfois concomitante au facteur d'éloignement, lorsque plusieurs items candidats se trouvent dans l'environnement du mot cible. Enfin, une autre cause que nous avons identifiée concernant les référents est relative plus spécifiquement à la position postposée, comme dans l'exemple suivant où `révélé` devrait être accordé avec `l'identité` :

*`Les trois plaignants [...] demandaient en particulier à ce que soit révélé l'identité de l'auteur.`

b) Observation des causes possibles d'erreurs

Nous appuyant sur notre classement en trois ensembles (tableau 7.1 p. 149), établi en nous fondant sur des travaux de recherche en psychologie cognitive et en acquisition, et à partir de l'observation des erreurs de notre corpus, nous avons procédé de manière systématique à

l'annotation des causes d'erreurs de grammaire. Dans de nombreux cas, nous avons identifié et annoté plusieurs causes possibles pour une erreur. Les différentes causes appartiennent parfois au même ensemble (16,7% des erreurs), comme dans l'exemple (15) ci-dessous, où l'accord erroné de *situés* est selon nous dû au texte, pour trois raisons possibles : d'une part le référent (*l'institut*) est à la fois éloigné et ambigu car quatre noms, de genre et nombre différents, pourraient être référent (*fondateurs*, *institut*, *recherche* et *chimpanzés*), d'autre part l'accord peut avoir été réalisé par proximité avec le masculin pluriel de *les chimpanzés*.

(15) **Les fondateurs de l'institut de recherche sur les chimpanzés, situés sur le campus de la centrale Washington University [...]*

Pour 23,6% des erreurs, nous avons mis en cause un ou des facteurs de plusieurs ensembles : texte et/ou humain et/ou situation de scription. Par exemple, dans l'énoncé (16), nous avons assimilé l'erreur d'accord sur *dégusté* à un simple oubli de l'accord de la part du scripteur, ou encore à la présence du complément du nom masculin *de chili* favorisant un accord de proximité :

(16) **[...] une pâte de chili souvent dégusté avec des biscuits salés aux crevettes.*

Dans les figures de cette section (figure 7.4 p. 154, diagrammes 7.5 p. 155 et 7.6 p. 157), une erreur qui se trouve dans plusieurs ensembles est représentée dans les effectifs et les pourcentages de chacun des ensembles. Ceci explique que la somme des pourcentages soit supérieure à 100%. Par conséquent, nous n'avons pas pu utiliser les tests statistiques mis en œuvre jusqu'à présent pour appuyer nos observations dans la suite de cette section.

Tendances selon les ensembles de causes d'erreurs

Sur la figure 7.4 p. 154, nous avons représenté à la fois la proportion d'erreurs dans les ensembles *Humain* (H), *Texte* (T) et *Situation* (S), ainsi que la proportion d'erreurs affectées à deux ou trois de ces ensembles. Nous voyons que les causes de l'ensemble *Humain* sont les plus représentées, suivies par l'ensemble *Texte* et enfin *Situation* dont la proportion est la plus faible. Nous voyons également que les erreurs causées à la fois par la situation et par le texte et/ou l'humain (HS, TS, HTS) sont peu nombreuses en comparaison aux erreurs causées conjointement par le texte et l'humain (HT).

Causes liées à la situation de scription

Les erreurs pour lesquelles la situation de scription est selon nous en cause sont peu nombreuses et ne représentent que 8,3% des erreurs de grammaire (voir figure 7.4 p. 154). La principale cause est *Dictée*. Elle concerne essentiellement des erreurs de syntaxe, principalement des oublis de mots dans les dictées, et des substitutions de mots, généralement proches phonologiquement. Il est difficile dans la plupart des cas de déterminer si le scripteur a alors mal entendu le texte énoncé ou si c'est le « dicteur » qui a mal prononcé, comme dans l'extrait (17) où *bêtement* se substitue à *vêtements*, phonologiquement très proche :

(17) *[...] des difficultés croissantes pour écouler bêtement, chaussures et sacs*

Parmi les autres causes possibles dans l'ensemble *Situation*, le clavier et les erreurs de frappe expliqueraient seulement 2,2% des erreurs grammaticales, principalement des oublis ou des substitutions de lettres conduisant à d'autres mots que ceux attendus.

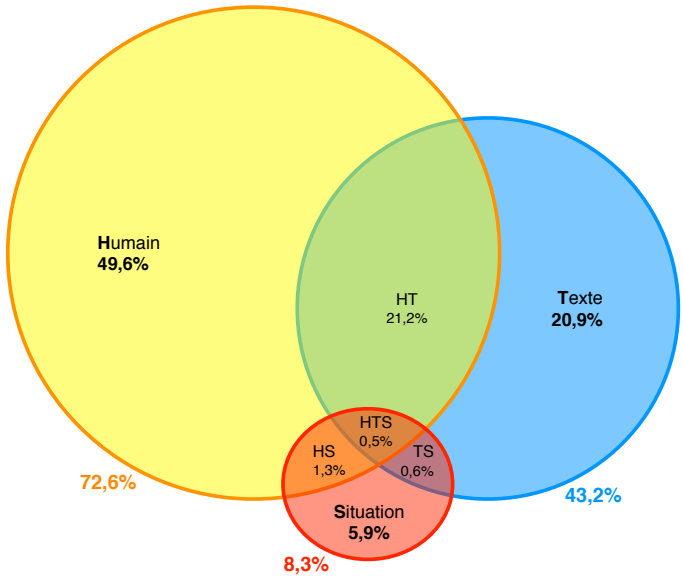


FIGURE 7.4 : Répartition des erreurs selon les causes *Humain*, *Texte* et *Situation*

Nous n'avons mis en cause la limite de temps que pour à peine plus de 1% des erreurs de grammaire, ce qui peut sembler peu compte tenu de la quantité d'erreurs relevées dans les dictées, qui sont les principales concernées par cette cause d'erreurs potentielle, par rapport aux autres sous-corpus. Nous avons en fait considéré le temps comme facteur d'erreurs uniquement dans les cas où des segments de phrases manquent, manifestement parce que le scripteur a décroché et n'a pas eu le temps de les écrire. Comme pour le facteur d'homophonie que nous estimons implicite dans la plupart des erreurs d'accord, et que nous n'avons donc pas annoté comme cause d'erreurs dans ces cas-là, nous considérons que la contrainte de temps, par la surcharge cognitive qu'elle implique, est une cause sous-jacente à la plupart des erreurs dans les dictées. Nous l'avons annotée indirectement par le biais des facteurs d'erreurs liés à la surcharge cognitive, et directement dans les cas cités précédemment d'absence de segments de phrase.

La dernière cause d'erreurs que nous avons identifiée en rapport avec la situation de scription concerne une quantité infime d'occurrences (moins de 1% des erreurs de grammaire) et apparaît dans les mails uniquement. Elle consiste à utiliser à l'écrit des tournures généralement réservées à l'oral, comme l'omission de la particule *ne* de la négation.

Causes liées au scripteur

Les erreurs dont nous faisons l'hypothèse qu'elles sont liées au scripteur sont les plus nombreuses et concernent 72,6% des cas (voir figure 7.4 p. 154). Les deux tiers (68,4%) d'entre elles ne sont causés que par l'humain et pas conjointement par le texte ou la situation.

Les connaissances du scripteur (ou l'application à mauvais escient de procédures automatisées) expliqueraient 27,9% des erreurs de grammaire du corpus (voir figure 7.5 p. 155), parmi lesquelles un quart (24,1%) résulte de méconnaissances de certaines graphies des mots relativement au contexte (nombre des mots dans les locutions ou mots composés, utilisation des traits d'union, etc.). Des connaissances manquantes sur le vocabulaire, une maîtrise partielle de la langue, ou encore l'omission délibérée d'accents, d'accords et même de mots ne sont pas des causes très représentées. Elles concernent à elles trois seulement 10,9% des erreurs de grammaire.

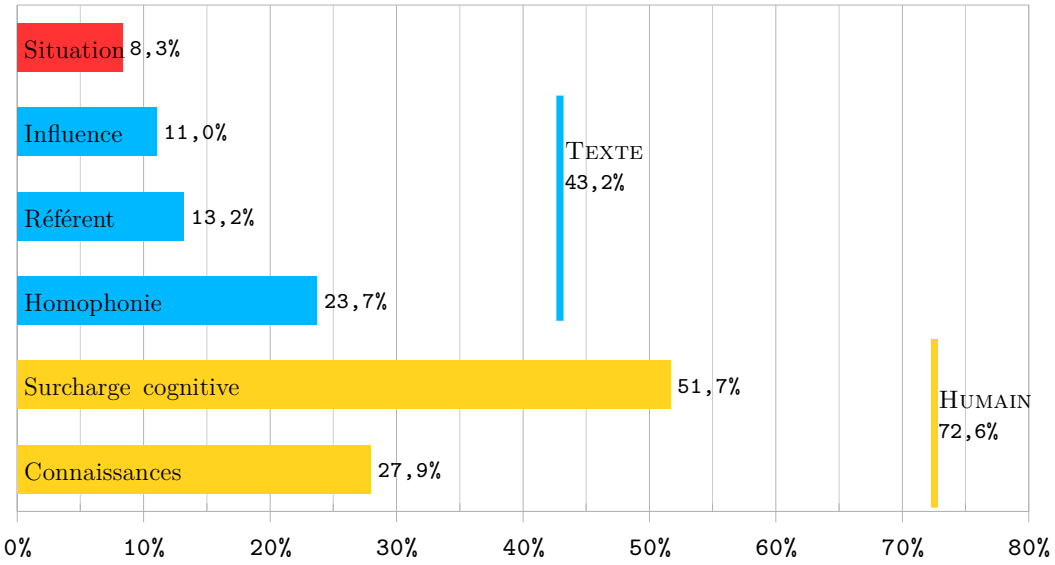


Figure 7.5 : Pourcentages d'erreurs dans les ensembles et sous-ensembles de causes

En revanche, les causes que nous avons regroupées autour de la surcharge cognitive touchent plus d'une erreur de grammaire sur deux (51,7%) (voir figure 7.5 p. 155). La moitié de ces occurrences (52,8%), qui représente 27,3% des erreurs de grammaire, est *a priori* le résultat d'un oubli de marques du féminin ou du pluriel. Lorsqu'une autre cause est présente, dans 65,7% des cas, il se peut que l'oubli soit le résultat, en plus d'une surcharge cognitive, de l'application, consciente ou non, d'une mauvaise procédure. Par exemple, dans l'extrait (18) suivant, l'absence d'accord du participe passé après *avoir* avec l'objet antéposé *les tentatives d'attentats terroristes* peut provenir d'un simple oubli, mais aussi de l'omission volontaire de l'accord, en généralisant la règle de non-accord du participe passé avec le sujet lorsqu'il est précédé de l'auxiliaire *avoir*. De plus, la configuration inhabituelle de la phrase, avec le sujet postposé et l'objet antéposé, favorise sans doute une confusion par le scripteur des fonctions de ces deux éléments, et par conséquent le non repérage de l'antéposition de l'objet.

(18) **les tentatives d'attentats terroristes qu'a connu Londres*

L'oubli de l'accord est également parfois associé, dans 12,7% des cas, au facteur de redondance sémantique, conduisant à omettre la marque du pluriel dans les contextes où la pluralité est déjà marquée sémantiquement. Nous en avons deux exemples dans l'extrait (19), *stand* et *produit* :

(19) **des dizaines de stand proposant des gammes de produit de beauté*

L'automatisation de procédures dont nous parlons plus haut (*cf.* p. 146), favorisée en condition de surcharge cognitive, a été identifiée comme cause d'erreurs pour seulement 11,7% des erreurs de grammaire. Nous retrouvons ici en majorité les accords de proximité et des participes passés après l'auxiliaire *avoir*, mais nous avons également inclus les erreurs de confusions de finales verbales en -i et -u, transformant un participe passé en une forme conjuguée. En effet, les travaux de Cordary [2010] sur l'orthographe des participes passés ont mis au jour, via des entretiens métagraphiques chez des élèves de seconde, une différence de comportement avec les participes passés selon qu'ils sont en -é ou en -i/-u. La finale en -é favoriserait l'accord en genre et en nombre avec le sujet, tandis que les finales en -i ou -u favoriseraient non pas l'accord mais la conjugaison de la forme. « Ainsi il y a les formes en [e] que l'on accorde, et les formes en [i] et [y] que l'on conjugue »

[Cordary, 2010, p. 83]. Les erreurs que nous avons effectivement relevées sur les participes passés en -i et -u semblent concorder, pour la moitié d'entre elles au moins, avec cette analyse selon laquelle ces formes verbales ont tendance à être perçues comme des formes pleines et à être conjuguées. Nous reprenons pour ces erreurs une hypothèse de l'auteure selon laquelle cette procédure de conjugaison du participe passé serait automatisée.

Enfin, une charge cognitive élevée peut rendre propice la survenue de ce que nous appelons des coquilles, matérialisées par des oublis ou des doublements de mots, ou encore par des substitutions de lettres (*pas*/*par*, *croissances*/*croissantes*). Ce sont 8% des erreurs de grammaire qui sont dans ce cas.

Causes liées au texte

Pour ce qui est des erreurs que nous supposons dues au texte, elles représentent 43,2% des occurrences (voir figure 7.4 p. 154), dont un peu plus de la moitié (54,9%) ayant pour cause l'homophonie, ce qui concerne 23,7% des erreurs de grammaire (voir figure 7.5 p. 155). L'homophonie grammaticale est dominante et consiste principalement en des confusions de mots grammaticaux (*est*/*et*, *à*/*a*, *sait*/*c'est*, etc.), mais également beaucoup d'erreurs de mode (*était*/*été*, *parût*/*paru*, *serve*/*servent*, etc.).

Les erreurs dues à l'homophonie, tout comme celles dues à l'influence d'un élément du texte, n'ont en général pas d'autres causes que le texte. Pour chacun de ces deux sous-ensembles de causes, environ 30% des erreurs seulement ont également une autre cause possible dans l'ensemble *Humain* ou *Situation*. Au contraire, 75,7% des erreurs causées selon nous par un problème d'identification du référent sont également potentiellement causées par un autre facteur dans l'ensemble *Humain* ou *Situation*. Plus précisément, ces erreurs sont en grande majorité le résultat, sans doute concomitant, d'un référent mal identifié (postposé, distant et/ou ambigu) et d'un oubli de l'accord par inattention ou surcharge cognitive.

Les causes impliquant les référents n'expliquent néanmoins que 13,2% des erreurs de grammaire. Quant à l'influence d'un élément ou d'un paramètre dans le texte, comme la fréquence d'une graphie mémorisée, la prédominance du pluriel dans la phrase, la présence d'un mot interférant avec le référent dans les accords de proximité, cela ne concerne que 11% des erreurs grammaticales (voir figure 7.5 p. 155).

Tendances selon les catégories d'erreurs

Si nous regardons plus précisément les facteurs d'erreurs mis en cause en fonction des catégories, nous observons des comportements différents. Le diagramme 7.6 p. 157 illustre ces différences en représentant, dans les quatre catégories d'erreurs grammaticales, le pourcentage d'erreurs entraînées par chacun des trois ensembles de causes ou par deux ou trois ensembles simultanément (en gris).

Catégorie Accord

Une grande majorité des erreurs d'accord est due au scripteur (88,9%), mais près de 30% a d'autres causes possible dans l'ensemble *Texte*. Nous voyons ainsi sur le graphique que 35,8% des erreurs d'accord sont *a priori* causées par le texte. L'ensemble des erreurs dues au scripteur doit sa prédominance à une proportion importante d'oubli des accords (environ 61,5% des erreurs d'accord), par inattention ou charge cognitive élevée. De plus, plus de la moitié de ces erreurs

possèdent d'autres causes potentielles, en particulier la position et l'identification du référent dans le texte.

Ainsi, 22,4% des erreurs d'accord sont concernées par un mauvais repérage du référent du mot à accorder (référent distant, ambigu, postposé). La postposition du référent est la cause la plus représentée (9,7%), suivie de l'éloignement du référent. Celui-ci ne serait responsable que de 7,8% des accords erronés, ce qui est relativement peu élevé. La distance ne semble donc pas être un facteur d'erreurs important.

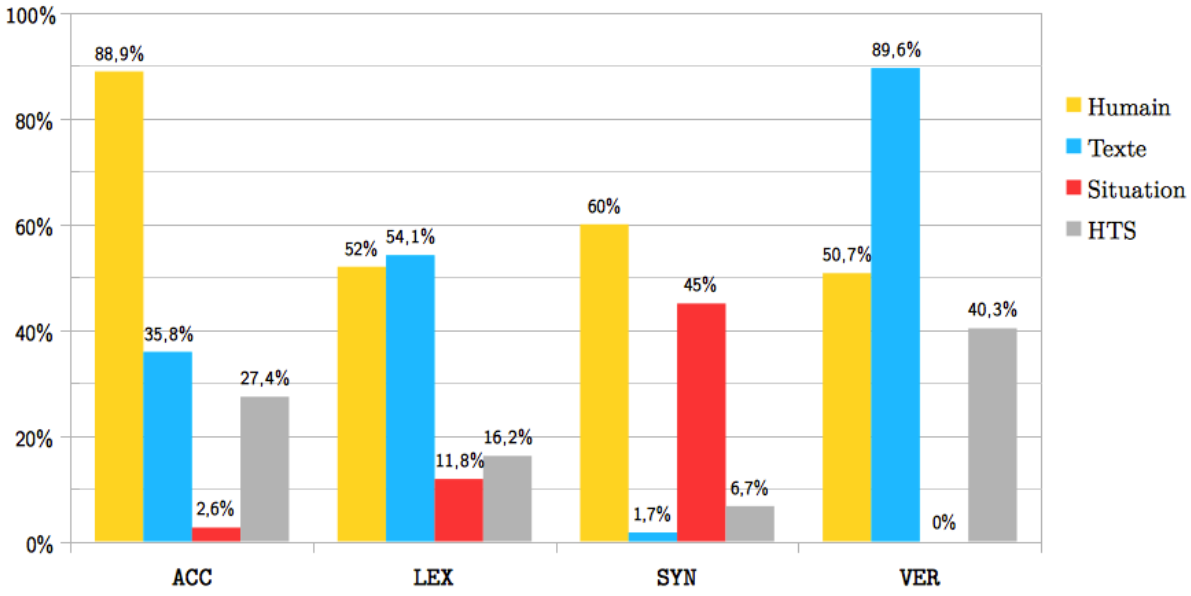


FIGURE 7.6 : Répartition des catégories d'erreurs selon les causes *Humain*, *Texte* et *Situation*

Catégorie Lexique

Contrairement aux erreurs d'accord où les causes humaines sont largement prédominantes, les erreurs de lexique sont causées en proportion équivalentes par le scripteur (52%) ou par le texte (54,1%) (voir figure 7.6). Dans l'ensemble *Humain*, la principale cause est la connaissance de la graphie des mots (28,3%). Il s'agit plus précisément à 80% de l'utilisation erronée du trait d'union, pour l'écriture des nombres ou de noms composés peu courants (*sous-entendait*, *sud-ouest*) par exemple. L'hypothèse d'une méconnaissance de la graphie va dans le sens des observations de Mout [2013, p. 206] sur la question des traits d'union. L'auteure constate une « méconnaissance généralisée de son usage » et rapporte également les constats de l'omission de ce signe que font Catach [1991] ou Lucci & Millet [1994].

Dans l'ensemble *Texte*, nous trouvons l'homophonie comme cause de la moitié (52,4%) des erreurs de lexique. Il s'agit d'une homophonie grammaticale deux fois sur trois. Nous retrouvons dans ces erreurs les confusions de mots grammaticaux tels que *ou*/*où*, *se*/*ce* ou encore *la*/*l'a*. L'homophonie lexicale est également importante (30% des erreurs causées par l'homophonie) et conduit à la substitution de lemmes homophones ou très proches phonétiquement (*thermes*/*termes*, *gemme*/*gène*).

Catégorie Syntaxe

Pour les erreurs de syntaxe, nous remarquons une prédominance non plus des ensembles *Humain* et *Texte*, mais des ensembles *Humain* et *Situation* (voir figure 7.6 p. 157). Les causes d'origine humaine sont les plus nombreuses. Elles représentent 60% des erreurs de syntaxe, parmi lesquelles la moitié est causée par des problèmes de maîtrise de la langue par les apprenants. Les erreurs de syntaxe sont également potentiellement causées, pour un tiers d'entre elles (33,3%), par de l'inattention ou une surcharge cognitive qui mènent à des coquilles telles que le doublement d'un mot (*`qui permet de d'unifier les sorties`) ou l'oubli d'un mot (*`En même fallait s'y attendre`).

Outre le facteur *Humain*, les causes provenant de la situation de scription sont également nombreuses. La syntaxe est la seule catégorie d'erreurs où ces causes n'apparaissent pas en quantité négligeable par rapport aux autres causes. Même s'il ne s'agit ici que de 26 occurrences, ces occurrences représentent 45% des erreurs de syntaxe, et consistent toutes en un ou plusieurs mots manquants, soit par un manque de temps pour les écrire sous dictée, soit par la non compréhension du texte dicté (noms propres notamment).

Catégorie Verbe

Pour les erreurs sur les verbes, la tendance est encore différentes des autres catégories (voir figure 7.6 p. 157). La grande majorité des erreurs a pour cause le texte (89,6%), mais les causes liées à l'humain représentent tout de même 50,7% des erreurs. Les erreurs de la catégorie `Verbe` sont presque exclusivement des erreurs de mode, pour lesquelles nous avons indiqué plus haut (p. 148) que l'homophonie était un facteur d'erreurs important. 79,1% des occurrences sont ici concernées par l'homophonie grammaticale comme cause d'erreurs, dans l'ensemble *Texte*. Parmi elles, 41,5% ont également une cause liée à l'humain : la connaissance ou l'automatisation de la graphie de certains mots, et plus précisément de la finale des participes passés en -i ou -u (`poursuivit`/`poursuivie`, `parut`/`paru`) qui ont tendance à être conjugués. Ce pourcentage important explique que le taux d'erreurs ayant des causes dans au moins deux des trois ensembles, comme nous l'observons dans le diagramme 7.6 p. 157, est plus élevé pour la catégorie `Verbe` que pour les autres catégories, avec 40,3%, soit plus d'une erreur sur trois.

c) Résumés sur les causes possibles d'erreurs

À défaut de disposer d'entretiens métagraphiques des scripteurs de notre corpus, qui nous donneraient des indications plus fiables sur les raisons d'apparition des erreurs, nous nous sommes appuyée sur des travaux en psychologie cognitive et en acquisition de l'orthographe pour élaborer des hypothèses sur les causes d'erreurs. Nous avons défini trois ensembles de facteurs d'erreurs possibles : 1) causes liées à la situation de scription (*Situation*), 2) causes liées au scripteur (*Humain*) et 3) causes liées au texte (*Texte*). Ces ensembles ne sont pas cloisonnés de façon étanche, et une même erreur peut recevoir plusieurs hypothèses de causes, issues d'un ou plusieurs ensembles, ce qui correspond à 34,6% des erreurs.

Le facteur humain est à l'origine d'au minimum une erreur sur deux dans chaque catégorie d'erreurs grammaticales, et même de quasiment 85% des erreurs d'accord. La surcharge cognitive notamment conduit à la moitié des erreurs de grammaire, et plus précisément à des oublis d'accords pour une erreur sur trois. Le texte est responsable d'environ 90% des erreurs de mode verbal. Le principal facteur d'erreurs est l'homophonie, qui touche un quart des erreurs de gram-

maire. Il s'agit essentiellement d'homophonie grammaticale qui altère les mots grammaticaux et les modes. Enfin la situation de scription génère moins de 10% d'erreurs. Un quart sont des erreurs de frappe, mais nous avons également relevé beaucoup d'erreurs dues à l'activité de dictée, notamment la contrainte de temps et la prononciation ou compréhension du texte, qui conduisent à des substitutions de lemmes, ou à des portions de texte manquantes. La catégorie `Syntaxe` est d'ailleurs celle qui contient le plus d'erreurs (presque la moitié) causées par la situation de scription.

Les erreurs d'accord sont en majorité dues à l'inattention ou à une surcharge cognitive du scripteur qui l'amènent à oublier de marquer un accord. Environ une fois sur trois, la position postposée ou éloignée du référent favorise la survenue de l'erreur. Nous avons également observé de nombreux cas, au contraire, d'ajouts d'accords sur les participes passés après *avoir*, par généralisation, peut-être automatique, de l'accord avec le sujet.

Les erreurs de lexique s'expliqueraient à part égales par l'humain ou par le texte, et plus précisément par l'homophonie dans la moitié des cas. L'homophonie grammaticale en particulier génère un tiers des erreurs de cette catégorie par substitution de lemmes. Un certain nombre d'erreurs également a pour cause la méconnaissance de la graphie d'un mot, mais cela concerne en grande majorité l'usage du trait d'union.

La syntaxe est la seule catégorie d'erreurs pour laquelle la situation de scription joue un rôle important. La compréhension orale et le temps restreint dans les dictées sont sans doute à l'origine d'omissions de mots. L'inattention et la surcharge cognitive, qui peuvent également être induites par le temps limité, expliqueraient aussi des coquilles syntaxiques, à savoir l'oubli ou le doublement de mots. La maîtrise imparfaite de la langue par les apprenants est une dernière cause d'erreurs de syntaxe.

Pour finir, la catégorie `Verbe`, essentiellement constituée d'erreurs de mode, voit la majorité de ses erreurs causées par l'homophonie grammaticale. Pour un grand nombre de ces erreurs, nous avons également mis en cause la méconnaissance qu'a le scripteur de la manière d'orthographier la finale des participes passés en -i ou -u, ou le fait qu'il conjugue automatiquement ces formes.

Les hypothèses sur les causes d'erreurs peuvent aider à la modélisation d'un outil pour détecter ces erreurs, dans le but de concevoir un outil fondé sur le fonctionnement humain. La surcharge cognitive du scripteur, induite par l'activité de rédaction et augmentée dans des situations de scription particulières, les connaissances déclaratives ou procédurales qu'il possède, et l'homophonie sont ainsi *a priori* les principaux facteurs favorisant les erreurs de grammaire. S'il est difficile de prendre en compte la charge cognitive avec un outil de détection d'erreurs, il peut être envisagé de fournir à l'outil, avant la saisie, des données concernant l'environnement de rédaction (saisie libre, contrainte de temps, type de clavier utilisé, etc.), ou encore concernant le scripteur, et en particulier son statut de natif ou non. Ceci permettrait par exemple d'orienter la recherche d'erreurs vers des types d'erreurs favorisés par une configuration donnée. Quant à l'homophonie, autre cause concernant beaucoup d'erreurs, elle pourrait entre autres être mise à contribution lors du processus de rétroactions faisant suite à la détection d'erreurs, en favorisant des propositions de corrections homophones des erreurs.

Ces hypothèses gagnent cependant à être complétées par d'autres hypothèses, sur la manière dont l'humain repère les incohérences grammaticales. Afin de formuler ce second type d'hypothèses, il nous faut préalablement nous intéresser aux processus cognitifs en jeu dans la révision, qui constitue un sous-processus de la production du langage écrit.

7.2 Révision du langage écrit

Ayant approfondi le fonctionnement cognitif de la production d'écrit et les raisons possibles de l'apparition des erreurs, c'est tout naturellement que nous nous intéressons à présent à la façon dont l'humain détecte ces erreurs. Après avoir abordé les processus cognitifs et les traitements réalisés lors de la révision, nous élaborerons des hypothèses quant aux éléments pouvant nous alerter sur la présence d'une erreur, et les confronterons à nos données. Nous raisonnerons alors en terme d'attentes de différentes natures auxquelles le texte ne répond pas lorsqu'il contient une erreur.

7.2.1 Le processus de révision

a) Exemples de modèles de révision

Le modèle de production d'écrit de Hayes & Flower [1980], que nous avons présenté précédemment (*cf.* § a) p. 142), confère à la révision le statut de sous-processus à part entière au sein-même du processus de production. De ce fait, ce modèle a été le point de départ de nombreuses recherches sur la révision, dont il résulte une diversité de modèles et de définitions de ce qu'est la révision. La définition donnée par Fitzgerald & Markham [1987] résume assez bien cette diversité :

> « *Revision means making any changes at any point in the writing process. It is a cognitive problem-solving process in that it involves detection of mismatches between intended and instantiated texts, decisions about how to make desired changes, and making the desired changes.* [5] »
>
> [Fitzgerald & Markham, 1987, p. 4]

Rajoutons que l'activité de révision, telle qu'elle est conçue dans les modèles que nous allons aborder, est une activité récursive, pouvant être déclenchée automatiquement ou délibérément, et qui peut interrompre les autres processus en cours. Elle est réalisée généralement au cours d'une seule lecture en ce qui concerne les révisions de surface, mais nécessite au moins deux lectures pour les erreurs de syntaxe ou pour la cohérence [Roussey & Piolat, 2005].

Les principaux modèles existants [Hayes & Flower, 1980 ; Scardamalia & Bereiter, 1983 ; Flower *et al.*, 1986 ; Hayes *et al.*, 1987 ; Butterfield *et al.*, 1996] ont en commun de fonder la révision sur la comparaison entre la représentation du texte écrit, élaborée avec la lecture, et la représentation mentale du texte à réviser.

C'est dans le modèle de Scardamalia & Bereiter [1983] que le principe de comparaison de deux représentations est présenté le plus explicitement. Le modèle représente le processus *CDO* (*COMPARE*, *DIAGNOSE*, *OPERATE*), qui fonctionne comme une boucle commençant avec une comparaison (sous-processus *COMPARE*) de la représentation du texte attendu et de la représentation du texte réel. La non correspondance entre les deux entraîne l'activation du sous-processus de diagnostic (*DIAGNOSE*) de l'origine du problème détecté. Dans les autres modèles,

5. Réviser signifie effectuer n'importe quelle modification à n'importe quel moment du processus de rédaction. Il s'agit d'un processus cognitif de résolution de problème dans le sens où la révision implique la détection de non-correspondances entre le texte souhaité et le texte effectivement produit, des décisions sur la manière d'effectuer les modifications souhaitées et la mise en oeuvre de ces modifications.

la comparaison n'est pas explicitement représentée mais sous-tend tout de même le processus de révision.

En revanche, la définition de la tâche, qui fixe le cadre de la révision, est représentée dans les différents modèles, sous différentes formes. Dans le modèle de Hayes & Flower [1980], nous retrouvons implicitement la définition de la tâche dans le sous-processus d'édition. L'édition, par un examen systématique et automatique de tout ce qui est mis en texte, détecte différents types de problèmes (conventions d'écriture, inadéquations de sens, inadéquations par rapport aux buts poursuivis, etc.) en se fondant sur des règles de type *condition-action*. La condition, qui cadre la révision, détermine ce qui est acceptable ou non en fonction du type de discours considéré, et l'action permet d'apporter la solution afin de rendre acceptable ce qui ne l'était pas d'après la condition.

Dans le modèle de Flower *et al.* [1986] et Hayes *et al.* [1987], la définition de la tâche apparaît comme un sous-processus fondamental en tout début de l'activité de révision et la conditionne. Il permet de définir comment le réviseur se représente la tâche de révision, les buts de l'activité, le niveau sur lequel elle doit porter et les stratégies pour atteindre les buts définis. Le modèle présente également un sous-processus d'*évaluation* qui repose sur la lecture pour comprendre le texte et l'évaluer, et permet la détection et le diagnostic de problèmes à différents niveaux (orthographe, syntaxe, cohésion, ton, etc.). Il débouche sur une représentation du problème qui est plus ou moins bien défini.

Ce processus de définition et diagnostic du problème rencontré est présent dans les autres modèles. Chez Scardamalia & Bereiter [1983], il constitue un sous-processus à part entière, *DIAGNOSE*. Chez Butterfield *et al.* [1996], il est représenté sous forme de connaissances dans la mémoire à long terme utilisées par la mémoire de travail.

La réflexion sur le problème détecté et son diagnostic conduit ensuite à choisir une stratégie, à prendre une décision quant au problème. Dans le modèle de Scardamalia & Bereiter [1983], le choix d'une stratégie s'effectue dans le sous-processus *OPERATE* et débouche soit sur un échec, soit sur une modification du texte. D'après le modèle de Flower *et al.* [1986] et Hayes *et al.* [1987], un choix entre cinq stratégies peut être opéré par le réviseur au cours du processus *Sélection d'une stratégie*. Le réviseur peut décider de rechercher plus d'informations sur le problème, de l'ignorer, de différer sa résolution, ou de modifier le texte en réécrivant le segment de texte concerné ou en effectuant une ou des modifications préservant le texte au maximum, en utilisant une table de solutions (*moyens-fins*).

À partir du modèle de Flower *et al.* [1986] et Hayes *et al.* [1987], les connaissances du réviseur figurent explicitement dans les modélisations, par une représentation des mémoires à long terme et de travail et de leurs interactions au sein du processus de révision. Sont stockées en mémoire à long terme des connaissances variées quant au thème, à la langue, aux buts de l'écrit et de la révision, aux moyens de résoudre les problèmes, etc. Pour Flower *et al.* [1986] par exemple, les connaissances d'un relecteur expérimenté concernant les écarts aux règles de grammaire, de ponctuation et d'orthographe lui permettent de disposer de règles du type condition-action prêtes à être utilisées en cas de nécessité. Ces règles se déclenchant automatiquement lorsque qu'une condition est remplie, elles épargnent alors au réviseur l'application d'un coûteux processus de diagnostic de l'erreur et de choix d'une stratégie pour la corriger en mettant directement à sa disposition la solution disponible en mémoire à long terme.

Les révisions peuvent donc être réalisées en partie automatiquement grâce à des règles condition-action acquises avec l'expérience, et ne requérir ainsi que peu de ressources attention-

nelles. Ceci est d'autant plus important que l'activité rédactionnelle, et plus particulièrement l'activité de révision, sont des activités très coûteuses en ressources de la mémoire de travail. Cette dernière joue un rôle crucial dans la révision.

b) Le rôle de la mémoire de travail

Les composants impliqués dans la révision

La mémoire de travail, telle que nous l'avons décrite (*cf.* § *Le rôle de la mémoire de travail* p. 144), est constituée d'un administrateur central qui coordonne le calepin visuo-spatial et la boucle phonologique (et le *buffer* épisodique). D'après le modèle de production du langage écrit de Kellogg [1996] (voir figure 7.3 p. 146), le sous-processus de révision ferait appel à la boucle phonologique et à l'administrateur central. La lecture, processus de base du système de contrôle, nécessite en effet l'administrateur central bien sûr, et la boucle phonologique (même en lecture silencieuse) qui permet de coder phonologiquement, de manière automatique, les mots perçus visuellement afin d'accéder à leur signification via leur prononciation mentale [Billières, 1988].

Il semblerait cependant que le calepin visuo-spatial également joue un rôle dans la détection, comme le montrent Dédéyan *et al.* [2006]. Selon eux, une procédure experte de détection d'erreurs d'accord, fondée sur la cooccurrence de flexions récupérées en mémoire à long terme (*cf.* § *Automatisation : le cas des accords* p. 162), s'appuierait sur les indices visuels que constituent les flexions de deux mots contigus. Ces indices visuels seraient alors traités par le calepin visuo-spatial. Les travaux de Dédéyan et Largy, en manipulant les polices d'écriture, montrent également que le réviseur est sensible aux caractéristiques graphoperceptives de la trace écrite [Largy *et al.*, 2005], et donc aux indices visuels.

Une activité coûteuse

Selon Piolat *et al.* [2004], l'activité de lecture dans le but de réviser serait nettement plus coûteuse en ressources cognitives que la « simple » lecture-compréhension. Une personne lisant pour comprendre établit des buts que le texte doit atteindre (véracité, cohérence, etc.) et détecte automatiquement un échec du texte à atteindre ces buts. Elle effectue alors déjà une évaluation du texte. Lors d'une lecture évaluative, le lecteur fixe un plus grand ensemble de contraintes et de critères sur le texte [Flower *et al.*, 1986], et le coût de l'activité augmente en conséquence. La détection d'erreur est alors très sensible aux surcharges cognitives, un échec de révision étant ainsi souvent dû à une surcharge en mémoire de travail. Une des méthodes pour optimiser les performances est alors d'automatiser un certain nombre de processus, comme nous l'avons vu pour la production, afin de réduire leur coût cognitif et libérer des ressources attentionnelles pour les autres traitements en cours, ou pour effectuer plusieurs traitements en parallèle. Parmi les processus automatisés, celui de la réalisation et révision de l'accord est l'un des plus étudiés. Nous avons déjà vu en effet qu'il existerait une procédure automatique de production d'accord chez l'expert, son pendant existant également pour la révision.

Automatisation : le cas des accords

Au stade de l'apprentissage, les connaissances procédurales acquises sont stockées en mémoire à long terme sous forme de règles du type condition(s)-action(s), appelées aussi algorithmes. Pour que la ou les actions, c'est-à-dire la procédure, soient réalisées, il est nécessaire que toutes les conditions soient remplies. En outre, tant que la procédure n'est pas réalisée (par exemple tant que, au cours de la révision, le verbe avec lequel s'accorde le sujet n'est pas apparu) ou annulée

par l'apparition de nouvelles conditions, elle doit être maintenue active en mémoire de travail, ce qui peut s'avérer très coûteux [Fayol & Largy, 1992]. L'application d'un tel algorithme pour réviser, requérant beaucoup de ressources attentionnelles, est caractéristique des apprenants, même si l'expert y a également recours en certaines circonstances. Notons ici le contraste avec les règles condition-action d'après Flower *et al.* [1986] et Hayes *et al.* [1987], évoquées au sujet du processus de révision (*cf.* § *Exemples de modèles de révision* p. 160) et qui, contrairement à celles que nous abordons à présent, sont décrites comme se déroulant de manière automatique et donc ne nécessitant quasiment pas de ressources cognitives pour détecter un problème et trouver sa solution.

En progressant dans l'apprentissage, les apprenants enregistrent petit à petit des informations linguistiques qu'ils rencontrent fréquemment au cours de leurs activités de lecture et écriture. Ces informations peuvent par exemple être des graphies, des séquences de mots, ou encore des cooccurrences de flexions. Il s'agit d'un apprentissage implicite, qui bénéficie ensuite à la détection d'erreurs [Largy & Dédéyan, 2002]. Ces configurations prototypiques stockées en mémoire peuvent ensuite être récupérées plus ou moins automatiquement par l'expert, par une automatisation du déclenchement de la procédure et/ou de la réalisation de la procédure. Ainsi, la révision d'un accord par un expert ne mobilise pas, ou peu, de ressources grâce à un processus automatique, tandis que pour un apprenant, l'application coûteuse de l'algorithme de vérification est nécessaire.

Dans certains cas cependant, incluant les situations ambigües, l'expert est amené à utiliser ce coûteux algorithme pour vérifier un accord. Largy & Dédéyan [2002] avancent alors l'hypothèse de l'existence d'un *monitoring* capable de décider, en cas de doute sur un accord, de déclencher l'exécution de l'algorithme. Le reste du temps, le *monitoring* se fonde sur la reconnaissance immédiate de cooccurrences de flexions pour ne pas déclencher de règle condition-action et préserver ainsi des ressources cognitives nécessaires à d'autres traitements. Il arrive cependant qu'il soit leurré par certaines configurations syntaxiques, telles que « Nom1 de Nom2 Verbe », où le verbe serait accordé par erreur avec le Nom2. Il s'avère alors plus fiable, pour réviser des accords tout du moins, d'utiliser l'algorithme de vérification. Ce dernier implique néanmoins de disposer de temps pour la révision, ce qui n'est pas toujours le cas. En situation de temps limité, la procédure automatique reste alors généralement plus adaptée.

7.2.2 Hypothèses sur la manière de détecter une erreur

Pour confronter nos données aux fonctionnements cognitifs que nous venons de décrire, et pour compléter l'annotation des hypothèses de causes d'erreurs que nous avons exposées plus tôt (*cf.* § *Hypothèses sur les causes possibles d'erreurs* p. 147), nous avons ajouté un second attribut pour chaque erreur de grammaire, prenant pour valeur ce qui nous permet *a priori* de détecter l'erreur.

Nous avons vu que les psychologues cognitivistes considèrent la révision comme une comparaison du texte produit au texte planifié, conduisant à des modifications en cas de non concordance entre les deux représentations. La détection de ces divergences entre le texte souhaité et le texte produit nous fait raisonner en termes d'attentes. Les erreurs de grammaire constituent des divergences entre le texte produit et le texte attendu. Elles consistent en l'apparition d'un élément inattendu, ou au contraire en l'absence d'un élément attendu, à un certain endroit du texte. Nous avons donc, à un moment donné, une attente pour un certain mot, d'une certaine catégorie et/ou avec certains traits morphosyntaxiques. Grunig [1993] nomme « suspension » le fait qu'un

élément soit en attente d'un autre élément.

> « Un élément a est **suspendu** pour un intervalle débutant au temps t si en ce temps t on est amené à le marquer [...] comme étant ***en attente*** pour un traitement complémentaire ultérieur.
> [...] si un adjectif a est suspendu depuis le temps t en attente d'un n et qu'il en rencontre un à un temps t', l'établissement en ce temps t' de la connexion entre a et n peut constituer le signal mettant un terme à la suspension. »
>
> [Grunig, 1993, p. 15]

Ainsi, dans l'exemple (20) ci-après, le déterminant `le` crée une attente pour un nom qui n'est pas satisfaite car une erreur, une faute de frappe vraisemblablement, a substitué une préposition (`entre`) au nom attendu (`centre`). En reprenant la terminologie de Grunig [1993], nous dirions que le déterminant est suspendu, sans qu'il soit mis un terme à cette suspension. Dans l'extrait (21), le nom pluriel `apprenants` ne comble pas non plus l'attente d'un nom singulier provoquée par le déterminant singulier `l'`, inversement le déterminant singulier ne comble pas l'attente d'un déterminant pluriel créée par le nom. Une erreur peut alors être définie en termes d'attentes réciproques non comblées entre deux éléments de la phrase.

(20) *Elle a attéri dans [le]Nms entre de tri bagage

(21) *Ceci permet des scénarios ouverts, adaptés à [l']Ns [apprenants]Dmp

Afin de mettre en évidence visuellement les attentes dans un énoncé, nous avons adopté un modèle de représentation que nous utiliserons dans les exemples suivants. Chaque attente est matérialisée par un petit carré positionné au-dessus de l'élément concerné, lui-même encadré. Lorsqu'une attente n'est pas comblée, le carré reste blanc. Lorsqu'au contraire une attente est satisfaite, le carré correspondant est rempli en noir et une flèche provenant de l'élément comblant l'attente pointe vers lui. Dans le cas d'attente facultative, comblée ou non, nous représentons cette attente par un carré rempli de pointillés.

(22) *Mais c'est [pas]neg grave tu [ne]neg diriges fort heureusement [rien]neg

Dans l'énoncé (22) sont représentées les attentes des particules de négation. À l'écrit, une négation est composée de deux éléments généralement indissociables, la présence de l'un impliquant la présence de l'autre. Nous pouvons voir dans l'exemple que la particule de négation `pas` a une attente non comblée (carré blanc) car il manque en effet l'élément `ne`. Pour la seconde négation en revanche l'attente réciproque entre les deux éléments `ne` et `rien` est bien comblée. `rien` répond à l'attente générée par `ne`, et *vice versa*.

a) Des attentes de différents types

Les travaux sur la révision d'écrits nous ont fourni quelques pistes sur les types d'attentes qu'un texte peut générer et qui permettent au relecteur d'identifier les incohérences. Il a par exemple été montré [Largy & Dédéyan, 2002] que les apprenants et les experts n'effectuent pas la révision des accords verbaux selon la même procédure, de même que nous avons vu qu'il existe une procédure novice et une procédure experte pour la production de ces accords (*cf.* § b).0 p. 146).

Pour détecter une erreur d'accord sujet-verbe, les apprenants ont recours à un algorithme de vérification de l'accord du type « condition-action », d'après leurs connaissances déclaratives, alors que les experts ont recours à une récupération automatique en mémoire de cooccurrences des flexions du nom préverbal et du verbe, du type -s/-nt [Dédéyan *et al.*, 2006]. Largy & Dédéyan [2002] ont également mis en évidence l'existence, chez l'expert, d'un *monitoring* chargé de repérer les combinaisons de flexions ambigües ou potentiellement erronées et de déclencher l'application de l'algorithme de vérification d'accord le cas échéant. Ainsi, par exemple, la présence d'un nom fléchi au pluriel directement suivi d'un verbe conjugué au singulier alerterait l'expert sur la possibilité d'une erreur et le conduirait à vérifier l'accord.

Même si les travaux sur la révision d'écrits portent essentiellement sur les accords sujet-verbe, il est probable que la procédure experte que nous venons de décrire s'applique également à d'autres situations d'accord, comme par exemple les accords nom-adjectif (ex. 23), déterminant-nom (ex. 24), auxiliaire-participe passé (ex. 25), etc. Nous inférons ainsi qu'au sein des syntagmes nominaux, des syntagmes verbaux, et entre les deux pour les accords sujet-verbe, les différentes flexions créent des attentes pour d'autres flexions, et une combinaison inattendue, potentiellement erronée, attire l'attention sur elle et permet la détection d'erreurs. Nous voyons par exemple dans l'extrait (23) que le déterminant `des` crée une attente pour une flexion nominale pluriel, attente qui est comblée par `crises`. Inversement, on s'attend à trouver un déterminant féminin pluriel précédant `crises`. Cette attente est également comblée. En revanche, le nom `crises` crée une attente facultative pour un adjectif, qui doit être au féminin pluriel. L'adjectif est bien présent, mais il est au masculin singulier et ne répond donc pas à l'attente. La combinaison des flexions du nom et de l'adjectif n'est ici pas cohérente et alerte le relecteur sur une erreur potentielle.

(23) **responsable* [Np] [Dfp] [Afp] [Nms] *des crises conjugal*

(24) [Np] [Dfs] **les résolution des problèmes*

(25) **Les divers formats de caractères* [PPp] [Être s ou Avoir] *sont conservé*

Cette hypothèse de procédure experte de révision des accords en général est également avancée par Largy *et al.* [2004a], selon lesquels un apprentissage implicite de l'accord aurait lieu au cours de la lecture chez les enfants et conduirait à un stockage en mémoire à long terme de configurations de morphèmes flexionnels fréquemment rencontrées (article + nom, nom ou pronom + verbe). Combiné à la pratique de l'écrit, ce phénomène serait à l'origine de la procédure experte de récupération en mémoire des configurations attestées.

En prolongement de ce postulat, nous émettons également l'hypothèse qu'une procédure experte semblable pourrait mobiliser un algorithme de vérification dans le cas de configurations autres que les accords. Nous pensons par exemple aux prépositions ou semi-auxiliaires qui requièrent un verbe à l'infinitif (ex. 26), mais également à des configurations non proximales comme l'absence d'une des deux particules de négation (ex. 22 p. 164), la présence à un endroit donné d'un mot d'une catégorie grammaticale inattendue (ex. 27), ou encore une graphie inconnue. Sur ce dernier point, Hayes [2004] avance l'hypothèse selon laquelle les erreurs d'orthographe seraient détectées en présence d'une configuration de lettres ne correspondant à aucune des images de mots stockées dans la mémoire à long terme du relecteur.

(26) *Un éléphant […] va (Vinf) retrouvé (Aux.) la liberté

(27) *une mamie (V3s) peu coutumière des aéroports empreinte (Dfs) le tapis bagage

De manière générale, nous pensons que l'expert, du fait de son expérience en lecture et écriture, possèderait en mémoire un très grand nombre de configurations syntaxiques et morphologiques qu'il récupèrerait lors de la révision de son texte, mettant ensuite en œuvre un algorithme de vérification en cas de doute sur une configuration ou un élément ambigu ou inattendu.

La dimension sémantique nous semble également jouer un rôle important dans la détection d'erreurs. Les résultats des travaux de Dédéyan & Largy [2003] montrent, par exemple, que les apprentis scripteurs seraient sensibles à la plausibilité sémantique du nom préverbal comme sujet pour l'application de l'algorithme de vérification lors de la révision des accords verbaux. En présence d'un nom local qui serait un sujet plausible du verbe, les apprenants commettent davantage de manqués dans la détection d'accord verbal erroné qu'en présence d'un nom local ne pouvant sémantiquement pas être sujet (*« Le compagnon des matelots siffle. » vs. « Le papier des chambres tombe. »* [Dédéyan & Largy, 2003, p. 104]). Les auteurs montrent également que les experts, en détectant ce type d'erreurs grâce à la récupération en mémoire de cooccurrences de flexions, ne seraient pas sensibles à la dimension sémantique dans ce cas. Il est probable en revanche qu'ils le soient, comme les apprenants, lors de l'application de l'algorithme de vérification, déclenchée, comme nous l'avons vu plus haut, en présence d'une combinaison de flexions ambigüe et inattendue. De plus, les résultats obtenus par Lusson [2013] indiquent que les adultes, aussi bien que les enfants, sont sensibles à l'influence de la pluralité conceptuelle portée par la cible de l'accord, à savoir le verbe, dans son expérimentation.

Nous pouvons également émettre l'hypothèse qu'en présence d'une substitution de mots homophones de sens très éloignés, nous sommes alertés sur une possible erreur (par ex. **en therme d'ecologie et d'environnement* ; **nous avon fait pose*), tout comme lorsque nous nous trouvons face à une phrase qui « ne veut rien dire ».

Il nous est également arrivé à tous de penser d'une phrase ou d'une portion de phrase qu'elle « sonne mal ». C'est alors la dimension phonologique qui entre en jeu et qui nous paraît importante à prendre en compte dans la détection d'erreurs. Lorsque nous lisons un texte, nous prononçons mentalement ce qui est écrit. Ce « langage intérieur » permet de transformer en code phonologique les mots perçus visuellement. Ainsi, nous avons une représentation phonologique du texte que nous lisons, et un son inattendu peut être l'indice d'une erreur. Les expérimentations de Lusson [2013] sur l'influence des indices morpho-phonologiques en révision ont d'ailleurs montré que la révision est « facilitée par la présence d'une terminaison phonologiquement audible, qui alerterait davantage les individus sur la présence illicite ou l'absence d'une marque morpho-phonologique sur le verbe » [Lusson, 2013, p. 130].

Toutes ces hypothèses que nous avons formulées, ainsi que l'observation des erreurs grammaticales dans notre corpus, nous ont permis d'identifier plusieurs types d'attentes que nous avons rassemblés selon six catégories, présentées dans le tableau 7.2 page suivante.

	Description	**Exemple**
Syntagme nominal (SN)	Attente pour une flexion de genre ou nombre au sein d'un SN	**[Une formation effectué_] par l'éducateur canin de la ville.* **L'essor du TAL pout l'ALAO date [du débuts] des années 90.*
Syntagme verbal (SV)	Attente pour une flexion de conjugaison, de mode, etc. au sein d'un SV	**[ça me le fais] pour toutes les photos* **c'est la solution [que j'ai utilisé_]*
Syntaxe	Attente pour une catégorie morphosyntaxique d'après la structure de la phrase	**j'_ai pas d'idée immédiate* **L'oral est l'écrit sont concernés*
Lexique	Attente pour une graphie	**Cet système repose sur un modèle de l'apprenant...* **Une femme de soixante_dix_huit ans...*
Sémantique	Attente pour un mot, une catégorie, des traits morphosyntaxiques d'après le sens de la phrase	**Un gemme masculin responsable des crises conjugales...* **en laissant le bouton de la souris enfoncée.*
Phonologie	Attente pour une sonorité	**Ton propos a au moins m'aventage de nous montrer...* **les grands magazins connaissent des difficultés croissances...*

Tableau 7.2 : Classement des types d'attentes

b) Observation des attentes dans la détection d'erreurs

Nous avons annoté chaque erreur de grammaire de notre corpus avec une ou plusieurs attentes auxquelles elle ne répondait pas. Plus précisément, nous avons associé à chaque erreur relevée le ou les types d'attente(s) non comblée(s) qui nous ont alertée sur une possible incohérence, que l'élément erroné soit à l'origine de l'attente (ex. 28) ou qu'au contraire il ne réponde pas à une attente (ex. 29).

(28) *la pachiderme a été envoyé dans un centre
(Nfs, Dms)

(29) *3 individus qui ont deposer une plainte
(PP, Prep ou Vconj/inf)

Tout comme pour les causes d'erreurs précédemment étudiées, le fait qu'une même erreur puisse être détectée grâce à plusieurs attentes (31,4% concernés), et soit ainsi comptabilisée dans les effectifs et les pourcentages de plusieurs catégories d'attentes, nous conduit à ne pas mettre en oeuvre de tests statistiques. Nous réalisons donc uniquement des descriptions des données obtenues suite à l'annotation.

Dans le diagramme 7.7 p. 169, nous avons représenté, pour chaque catégorie d'erreurs, la proportion de chaque type d'attentes permettant de détecter les erreurs de la catégorie. Pour les erreurs d'accord, nous observons trois types d'attentes en proportions équivalentes. Ces erreurs sont relevées grâce à des attentes sémantiques (37,7%), et sans surprise grâce à des attentes de flexions dans les syntagmes nominaux (41%) ainsi que dans les syntagmes verbaux (37,3%). Nous avons vu en effet que les erreurs d'accord peuvent être détectées par les relecteurs experts à l'aide d'une procédure de récupération en mémoire de cooccurrences de flexions fréquemment rencontrées. Lorsque les flexions ne forment pas une combinaison stockée en mémoire, il est possible qu'il y ait une erreur, comme dans les exemples (30) et (31) suivants :

(30) *les résolution des problèmes
(Np, Dfs)

(31) *après l'éléction présidentielles
(Afs, Nfp)

Les attentes impliquant la sémantique concernent les cas où il est nécessaire d'analyser le sens de la phrase, ou d'un segment de la phrase, pour confirmer une anomalie d'accord et la corriger. Il s'agit essentiellement d'une part d'erreurs de nombre sur les noms précédés de la préposition `de`, qui ne permet pas de déterminer le nombre à appliquer au nom qui la suit (ex. 32) et d'autre part, d'erreurs sur des adjectifs ou participes passés dans des configurations syntaxiques contenant un complément du nom (du type « Nom1 de Nom2 Adj » ou « Nom1 de Nom2 Verbe Part.Passé ») et où il existe une ambigüité sur le référent à prendre en compte (Nom1 ou Nom2) pour réaliser l'accord (ex. 33). Dans ces deux cas, la sémantique est nécessaire pour déterminer s'il y a une erreur ou non. Ce type d'attente est par ailleurs souvent accompagné d'autres types. Nous voyons sur le diagramme qu'un quart (24,8%) des erreurs d'accord est détecté via plusieurs attentes, et dans la majorité des cas (78,4%), il s'agit de l'association d'une attente sémantique et d'une attente de flexions (nominales ou verbales).

(32) *le pahciderme a été envoyé dans un centre [Singulier: de protections] pour animaux sauvage*

(33) *une pâte de chili souvent [Féminin: dégusté] avec des biscuits salés*

Pour la catégorie `Lexique`, le diagramme 7.7 montre en toute logique une prédominance des attentes lexicales (57,2%), composées principalement d'attentes pour des graphies données ou pour d'autres mots (ex. 35). Elles rassemblent notamment l'intégralité des erreurs de traits d'union (ex. 34) et près de 42% des substitutions de lemmes.

Les attentes syntaxiques également représentent un pourcentage non négligeable de 30,1% des occurrences d'erreurs de lexique. Une majorité (62,3%) est constituée de confusions de mots grammaticaux, et nous relevons même une proportion élevée (88,4%) de cas de substitutions de mots de catégories morphosyntaxiques différentes (ex. 36).

(34) *[Trait d'union: Devons nous] penser qu'il y a eu un phénomène...*

(35) *un éléphant [accro: accroc] à l'héroine*

(36) *déposer [V: se] [Dfs: valise]*

Pour les erreurs de syntaxe, deux types d'attentes se détachent très nettement (voir figure 7.7 p. 169) : les attentes syntaxiques qui sont à l'origine de la détection de 98,3% des erreurs, et les attentes phonologiques qui concernent 81,7% des erreurs. Nous remarquons sur le diagramme que le taux d'attentes multiples est identique à celui des attentes phonologiques. Cela tient au fait que nous avons associé une attente syntaxique à chaque attente phonologique. Nous avons en effet considéré que nous détections les erreurs de syntaxe à la fois par une anomalie phonologique dans la lecture de la phrase et par la perception d'un mot manquant ou en trop (ex. 37).

(37) *Wikipédia[…] était poursuivie a [Nfs: la] [Nms: d'un] [Dms: article] de fin septembre*

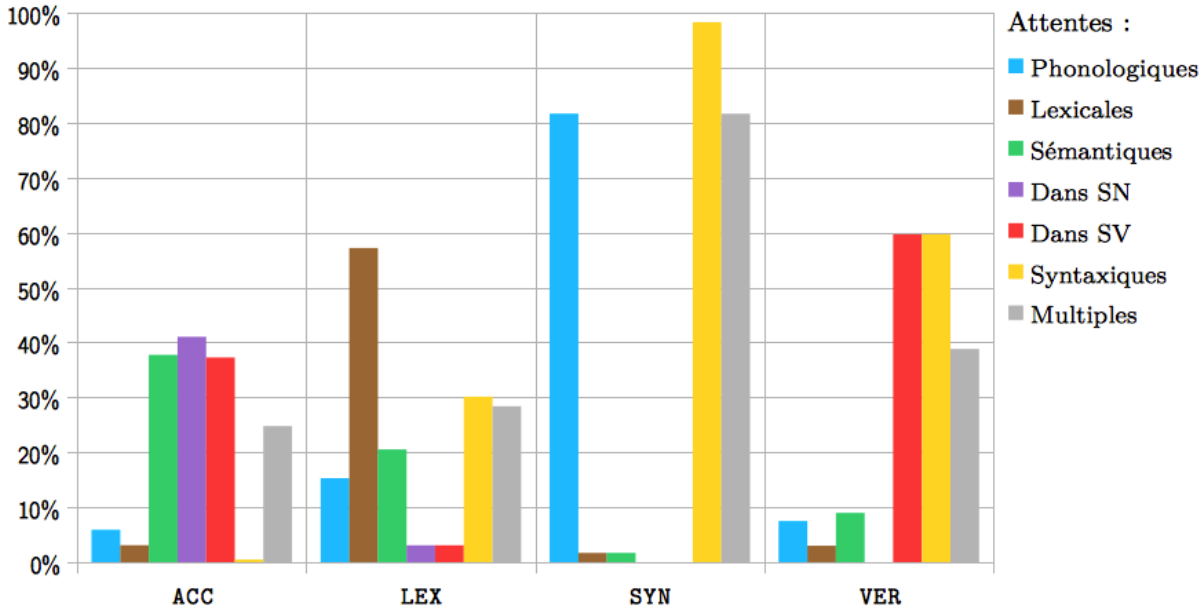


FIGURE 7.7 : Répartition des attentes selon les catégories d'erreurs

Enfin, la catégorie `Verbe` présente un pourcentage identique (59,7%) de détection grâce à des attentes syntaxiques et grâce à des attentes de flexions dans les syntagmes verbaux, les deux types d'attentes étant concomitants dans 42,5% des cas. Les attentes syntaxiques non comblées qui conduisent à identifier une erreur consistent principalement en l'attente d'un certain mode à un endroit de la phrase où il n'apparaît pas. Dans l'extrait (38) par exemple, nous nous attendons à trouver un participe passé (ou un adjectif) pour qualifier `article`, mais c'est une forme conjuguée que nous trouvons. Concernant les flexions de formes verbales attendues mais non présentes, il s'agit dans la moitié des cas d'attentes pour un participe passé (ex. 39), auquel se substitue une forme infinitive ou une forme conjuguée, et dans près d'un tiers des cas d'attentes pour une forme conjuguée à une certaine personne, à laquelle se substitue un participe passé ou encore une flexion de conjugaison à une mauvaise personne (ex. 40).

(38) *poursuivie à la suite [Adj ou PP: d'un article] [SN sujet: parut] fin septembre

(39) *la rétroaction doit [PP: être] [Prep ou Vconj/inf: expliquer]

(40) *[V3s: Elle] [ProSuj 1s: aimerai] pouvoir décaler plusieurs fois la numérotation

De manière générale, les attentes les plus faiblement impliquées dans la détection d'erreurs sont les attentes phonologiques, qui concernent seulement 14,6% des erreurs grammaticales. Ceci concorde avec le fait qu'environ 78% des erreurs de grammaire sont homophones de la forme correcte et ne s'entendent donc pas.

Au contraire, les attentes sémantiques sont les attentes les plus représentées (27,4%). Ce

résultat s'explique par le fait que ce type d'attentes concerne essentiellement les erreurs d'accord, qui sont les erreurs grammaticales les plus nombreuses.

Les attentes de flexions dans les syntagmes verbaux sont les deuxièmes les plus représentées (26%). Outre les erreurs d'accord, elles concernent également de nombreux cas de confusion de flexions infinitives, conjuguées et de participes passés.

Les autres types d'attentes, lexicales, syntaxiques et dans les syntagmes nominaux, permettent de détecter en moyenne un peu plus de 20% d'erreurs chacune (lexicales : 18,8% ; syntaxiques : 21,8% ; dans les SN : 23,3%).

c) Résumé des attentes permettant de détecter les erreurs grammaticales

Nous sommes partie de l'hypothèse selon laquelle, lors de la révision de textes, l'humain est sensible à des attentes de différentes natures, qui sont déclenchées par le texte et qui, lorsqu'aucun élément n'y répond, attirent son attention sur la présence d'une éventuelle erreur. En nous appuyant sur des travaux de recherche sur la révision [Dédéyan & Largy, 2003 ; Dédéyan *et al.*, 2006 ; Largy & Dédéyan, 2002 ; Largy *et al.*, 2004a ; Lusson, 2013], nous avons identifié six catégories d'attentes en observant notre corpus : les attentes de flexions dans les syntagmes nominaux, dans les syntagmes verbaux, les attentes sémantiques, lexicales, syntaxiques et phonologiques.

Les **attentes de flexions dans les syntagmes nominaux et verbaux** concernent presque exclusivement les erreurs d'accords, et touchent la majorité d'entre elles. Les erreurs de cette catégorie sont également détectées grâce à des attentes sémantiques plus d'une fois sur trois. Il s'agit des cas où la syntaxe seule ne permet pas de déterminer la bonne flexion. Les attentes sémantiques concernent d'ailleurs principalement ces cas.

Les **attentes sémantiques** permettent également de détecter environ 20% des erreurs de lexique, mais ces dernières sont surtout identifiées grâce à des **attentes lexicales** pour un certain mot ou une certaine graphie. Les substitutions de lemmes dans cette catégorie `Lexique`, lorsqu'elles changent la catégorie grammaticale, peuvent dans certains cas être aussi repérées par le biais d'attentes syntaxiques.

Les **attentes syntaxiques** sont la source de détection de presque toutes les erreurs de syntaxe, souvent en association avec des **attentes phonologiques**, puisqu'un mot manquant ou un mot en trop « s'entend » lors de la lecture de la phrase. Elles permettent également de détecter les erreurs de mode sur les verbes, dans les mêmes proportions que les attentes pour des flexions dans les syntagmes verbaux.

7.3 Conclusion

Dans ce chapitre nous avons abordé les phénomènes cognitifs liés à la production du langage écrit, et en particulier au sous-processus de révision. Nous avons vu que ces activités impliquent la gestion de plusieurs processus, et sont ainsi très coûteuses en ressources cognitives de la mémoire de travail, qui joue un rôle central. La surcharge cognitive est d'ailleurs source de nombreuses erreurs, notamment des erreurs d'accord. Pour préserver des ressources cognitives pour d'autres tâches plus coûteuses et mener à bien son travail de rédaction et de révision, le scripteur acquiert avec l'expertise des automatismes qui lui permettent de réaliser certaines tâches (comme réaliser un accord) en faisant très peu appel à la mémoire de travail. Toutefois,

les procédures automatisées peuvent elles aussi produire des erreurs, appelées erreurs d'experts (comme les accords de proximité).

> « Il semble donc que l'erreur soit inhérente au fonctionnement même de notre système cognitif et que son occurrence soit liée non seulement à l'existence des connaissances (déclaratives ou procédurales) mais aussi à leur gestion en temps réel dans le cadre d'activités plus ou moins complexes à conduire en parallèle avec d'autres. »
>
> [Fayol & Largy, 1992, p 96]

Les connaissances du scripteur jouent également un rôle important dans la survenue d'erreurs, et pourrait conditionner la vérification grammaticale automatique. Nous pensons par exemple à la prise en compte de la langue maternelle, ou plus précisément de la distinction entre les francophones natifs et les apprenants de FLE. Nous avons déjà indiqué ne pas aspirer à concevoir un outil destiné spécifiquement aux apprenants, mais à destination de tous les publics, qu'ils soient francophones ou non. Or, nous avons vu dans le chapitre précédent que les apprenants de FLE commettent des erreurs plus nombreuses et parfois différentes des natifs. Par exemple, les erreurs d'accord sont plus nombreuses chez ces scripteurs, en particulier les erreurs de genre sur les déterminants (*cf.* § *Catégorie Accord* p. 126), et sont également moins souvent homophones des formes correctes que les chez les francophones natifs (*cf.* § f) p. 128). C'est également le cas pour les erreurs de substitution de lemmes (*cf.* § *Catégorie Lexique* p. 126), qui sont plus nombreuses et moins souvent homophones. Savoir si le scripteur est natif peut présenter un intérêt dans la détection d'erreurs, mais est surtout intéressant au niveau des rétroactions. Un francophone natif sait bien que tel mot est féminin ou masculin et n'aura pas besoin d'une explication sur ce point dans une rétroaction sur une erreur d'accord en genre. Cela peut par contre s'avérer utile pour un scripteur étranger.

Le scripteur est donc le principal pourvoyeur d'erreurs, mais le texte et le contexte de scription y concourent fortement. Nous avons ainsi également relevé que, dans notre corpus, un quart des erreurs de grammaire seraient favorisées par l'homophonie, et en particulier l'homophonie grammaticale qui altère les mots grammaticaux et les flexions verbales. Près d'un quart également des erreurs d'accords proviendraient d'une mauvaise identification du référent, mais l'éloignement de celui-ci n'apparaît pas comme un facteur d'erreur important.

Le facteur distance revêt néanmoins de l'importance du fait de la prédominance des contextes où l'erreur et son référent sont contigus, et généralement dans un même syntagme (*cf.* § g) p. 131). Des études sur la distance dans la situation d'accord du verbe avec le nom sujet mettent en avant deux approches : une approche non syntaxique qui considère la distance linéaire, et une approche syntaxique dans laquelle l'accord est réalisé au niveau de l'encodage grammatical, avant le positionnement des éléments dans la phrase, et donc indépendamment de la distance linéaire [Lusson, 2013].

Dans le cadre de la première approche, selon Jespersen [1924], le maintien en mémoire des traits du sujet jusqu'à l'apparition du verbe entraînerait une surcharge cognitive d'autant plus importante que la distance linéaire est grande entre les deux éléments. Cette approche est également défendue par les travaux de Fayol & Got [1991] ; Fayol & Largy [1992] ; Fayol *et al.* [1994], selon lesquels l'accord du verbe se ferait automatiquement avec le nom le plus proche linéairement pour alléger les besoins en ressources de la mémoire. C'est dans cette première approche que s'insère l'hypothèse que nous avions formulée sur l'influence de la distance sur l'apparition d'erreurs (*cf.* § 14 p. 152).

Selon la seconde approche, l'accord dépend de la position du sujet et du verbe dans la structure syntaxique, et non pas dans la structure linéaire. « Les erreurs d'accord seraient plutôt consécutives à la percolation (ou migration) des traits syntaxiques du nom local dans la structure hiérarchique de la phrase » [Lusson, 2013, p. 26]. Une première conception de cette approche permettrait d'expliquer, par exemple, que les erreurs sont plus fréquentes lorsque le sujet est séparé du verbe par un syntagme prépositionnel plutôt que par une proposition relative. En effet, selon l'hypothèse de *clause packaging* de Bock & Cuting [1992], le nom sujet et le nom local d'un syntagme propositionnel sont encodés en même temps dans la même unité, tandis que le nom d'une proposition relative n'appartient pas à la même unité d'encodage que le nom sujet, ce qui protège ce dernier des interférences avec le nom local. En résumé, une frontière propositionnelle isole le nom local dans une proposition relative et limite les risques d'accord erroné entre ce nom local et le verbe. Effectivement notre corpus présente moins d'erreurs d'accord après une proposition relative qu'après un syntagme prépositionnel. Il existerait en outre un effet de la longueur du syntagme prépositionnel interférant [Bock & Cuting, 1992], avec une augmentation de la fréquence des erreurs d'accord avec les syntagmes longs. Une seconde conception de l'approche syntaxique considère que l'accord sujet-verbe est réalisé au cours de la construction de la structure hiérarchique de la phrase, à partir de la représentation conceptuelle du message et des traits de nombre des différents constituants. Selon Smedt [1990] en particulier, l'accord serait le résultat d'opérations successives d'unification le long de l'arbre syntaxique, depuis le syntagme nominal sujet jusqu'au syntagme verbal. L'unification permet de vérifier que le nom sujet et le verbe partagent le même trait de nombre pour un accord correct. Il s'agit d'un mécanisme fréquemment utilisé en TAL que nous abordons dans le prochain chapitre (*cf.* section 8.2.6 p. 190).

Ce que nous retenons de ces approches, c'est que la distance linéaire et la distance syntaxique sont toutes deux susceptibles d'influencer la réalisation des accords. Nous retenons également que l'unité de mesure de la distance, dans l'approche syntaxique en particulier, est le syntagme, dont la longueur pourrait influer sur les erreurs d'accord. Utiliser la segmentation en syntagmes peut alors être un atout pour l'analyse des textes et la détection d'incohérence. Elle permettrait de porter son attention en priorité sur les mots faisant partie du même syntagme pour rechercher des erreurs accords. Le mécanisme d'unification cité plus haut serait alors parfaitement adapté pour vérifier ces accords au sein d'un syntagme.

Les erreurs étant inévitables, en particulier les erreurs d'experts dues à l'automatisation, il faut généralement recourir à la vérification de ce que nous écrivons, via le processus de révision. Ce processus cognitif, qui peut intervenir au fil de la rédaction ou bien être sollicité en fin de rédaction, fonctionne sur la base d'une comparaison entre la représentation du texte attendu et la représentation du texte produit, élaborée au cours de la lecture. Une absence de correspondance entre ces deux représentations requiert alors de diagnostiquer et définir le problème rencontré, sélectionner une stratégie (modifier le texte ou non) et la mettre en œuvre.

La révision est cependant conditionnée par la définition (ou environnement) de la tâche. Selon les contextes, les mêmes faits seront considérés comme des erreurs ou non. Ainsi, l'écriture d'un mail à un proche n'est pas soumise aux mêmes contraintes que l'écriture d'une lettre de motivation. Il en va de même pour la révision, qui sera beaucoup plus stricte pour la seconde tâche que pour la première. Omettre des accents ou des majuscules dans un mail à un ami sera toléré, mais considéré comme inacceptable dans une lettre de motivation. De manière plus générale, cela pose la question de la notion d'erreurs, du registre de langue et du bien fondé de la détection et la correction des écarts grammaticaux constatés dans certains contextes où les « erreurs »

peuvent être le résultat d'une volonté délibérée du scripteur d'adopter un certain style, comme cela semble être le cas dans les mails avec les « erreurs » de syntaxe (*cf.* § *Catégories Syntaxe et Verbe* p. 117). Nous avons indiqué que ces erreurs trouvaient vraisemblablement leur origine dans l'adoption délibérée par les scripteurs de tournures typiques de l'oral. En effet, les mails en question constituent des échanges entre deux ou plusieurs personnes, et peuvent être assimilés à des conversations écrites. Si les écarts de syntaxe relevés dans les mails sont effectivement des erreurs sur le plan strictement grammatical, ils sont néanmoins conformes à la représentation du texte planifié par le scripteur et ne sont donc pas censés être détectés comme des divergences entre le texte écrit et le texte planifié (*cf.* § *Exemples de modèles de révision* p. 160). Ainsi leur détection, et plus encore leur correction, ne sont sans doute pas souhaitées par leur auteur.

Quelle utilisation est donc faite d'un outil de vérification grammaticale dans des cas tels que la rédaction de mails, de notes personnelles, etc. ? Serait-il vraiment opportun de prendre en compte ces situations particulières dans un outil de détection d'erreurs ? Il nous semble que l'utilisation d'un outil de vérification grammaticale ne prend tout son sens que dans les situations où justement nous ne souhaitons pas laisser d'erreurs, et qu'elle est plutôt indésirable dans les autres situations. Mais seule une étude des usages des utilisateurs permettrait de s'en assurer. Il serait néanmoins peut-être pertinent d'envisager, au sein d'un outil de vérification grammaticale, différents seuils de « tolérance grammaticale » pour s'adapter à ces situations diverses.

Nous assimilons la révision à un phénomène d'attente que nous avons à un endroit donné du texte pour un mot d'une certaine catégorie morphosyntaxique, pour une certaine flexion, pour un certain sens. etc. Nous avons ainsi identifié six différents types d'attentes qui selon nous entrent en jeu lors de la détection d'erreurs : attentes de flexions dans un syntagme nominal, dans un syntagme verbal, attentes lexicales, attentes syntaxiques, attentes sémantiques et attentes phonologiques.

Les attentes sémantiques sont les plus nombreuses à nous avoir alertée sur la présence d'une erreur dans notre corpus parce qu'elles n'étaient pas comblées. La nécessité de recourir au sens pour détecter en particulier un grand nombre d'erreurs d'accord constitue un problème pour un outil de vérification grammaticale automatique qui, même avec des ressources sémantiques, aura du mal à décider du sens à donner à un mot, un syntagme, une phrase. Il sera moins difficile de considérer les attentes syntaxiques ou les attentes de flexions, elles aussi nombreuses, mais faisant intervenir des critères plus concrets de catégories et sous-catégories grammaticales.

Ce chapitre nous a permis de dégager quelques éléments ayant *a priori* un fort potentiel pour la révision de texte écrit : le concept des attentes, la segmentation en syntagmes minimaux, ainsi que l'unification de traits. Ces principes correspondent à des notions couramment utilisées dans le domaine du TAL pour l'analyse ou la génération de textes. Nous les présentons dans le prochain chapitre afin de déterminer dans quelle mesure ils pourraient être utilisés pour la vérification grammaticale automatique.

Chapitre 8

Proposition d'un modèle pour la vérification grammaticale

Sommaire

8.1 Structure du modèle 176
8.1.1 Mécanisme de lecture gauche-droite 176
8.1.2 Étiquetage morphosyntaxique 178
8.1.3 Segmentation en *chunks* 179
8.2 Des attentes aux piles 183
8.2.1 Les valences de Tesnière 183
8.2.2 Les actants de Mel'čuk 186
8.2.3 Des attentes de différents niveaux 186
8.2.4 Un traitement par piles 188
8.2.5 Contenu des piles 189
8.2.6 Portée des attentes 190
8.3 Ressources 192
8.3.1 Règles de segmentation en chunks 193
8.3.2 Ressources pour les attentes 193
8.4 Fonctionnement attendu 197
8.4.1 Exemple de détection par une attente non comblée 197
8.4.2 Exemples de détection par un échec d'unification 200
8.4.3 Rétroactions 201
8.5 Conclusion 203

Les travaux menés en psychologie cognitive sur la production et la révision d'écrits, ainsi que notre étude des causes possibles des erreurs de notre corpus et des moyens de les détecter, nous ont conduite vers différents concepts appropriés à la vérification grammaticale. Chez l'humain, la détection d'une erreur grammaticale reposerait sur une attente du réviseur non comblée. Nous plaçons ce principe d'attente au cœur du modèle que nous proposons dans ce chapitre. Nous le complétons avec les principes d'unification et de segmentation en syntagmes minimaux, couramment utilisés dans le domaine du TAL, qui faciliteront la gestion des attentes du point de vue du traitement numérique. Le premier est particulièrement adapté à la vérification des accords et le second constitue une unité de calcul intermédiaire permettant de définir des bornes simplifiant la recherche d'incohérences grammaticales.

Dans une première section de ce chapitre, nous présentons la structure générale de notre modèle pour la vérification grammaticale et décrivons les différents traitements effectués lors de l'analyse de la phrase.

Dans la seconde section, nous abordons la manière dont sont gérées les attentes afin de permettre la détection d'incohérences grammaticales.

La troisième section s'intéresse à la question des ressources langagières nécessaires au modèle.

Enfin, dans la quatrième section, nous testons la validité de notre modèle sur quelques phrases de notre corpus.

8.1 Structure du modèle

La vérification grammaticale proposée par les différents outils libres que nous avons étudiés (*cf.* section *La vérification grammaticale* p. 27), repose toujours sur un même schéma de traitement. Les phrases sont analysées dans leur ensemble. Elles sont d'abord segmentées en tokens (tokenisation). Ces tokens reçoivent ensuite des étiquettes de traits morphosyntaxiques à partir d'une lexique étiqueté, puis sont parfois regroupés en chunks. Enfin sont appliquées des règles locales pour rechercher des erreurs, selon le principe du *pattern-matching*. Ce principe présente l'inconvénient de nécessiter un très grand nombre de règles (*cf.* § a) p. 43) qui provoquent des détections redondantes et ne peuvent décrire l'infinité des erreurs possibles.

Contrairement à ces outils, le modèle que nous proposons repose sur une analyse de la phrase de gauche à droite, au fur et à mesure de l'écriture ou bien de la lecture après la frappe. Il réalise également de façon concomitante les différents traitements d'étiquetage, de segmentation en syntagmes et de recherche d'incohérence, token après token. Notre modèle se distingue également en proposant non plus une recherche des erreurs par un principe de *pattern-matching*, mais une détection des incohérences en construisant des attentes de différents niveaux auxquelles le modèle réagit lorsqu'elles ne sont pas validées.

Notre modèle est illustré dans la figure 8.1 p. 177. Il est constitué d'un processus d'analyse d'énoncés qui observe chaque mot (token) de l'énoncé l'un après l'autre. Nous présentons ici cette analyse gauche-droite, ainsi que les processus menés en parallèle : l'étiquetage morphosyntaxique des mots et la segmentation de l'énoncé en syntagmes. La détection d'incohérences, également menée en parallèle de ces traitements, fait l'objet d'une présentation dans la section suivante.

8.1.1 Mécanisme de lecture gauche-droite

Pour tenter de reproduire le modèle humain de détection des erreurs et suivre également le processus linéaire de scription et de lecture de gauche à droite, notre modèle se veut capable de prendre en compte linéairement les mots constituant une phrase, au fur et à mesure de leur apparition. Nous réalisons ainsi un traitement de la phrase de gauche à droite, prenant comme entrée le dernier mot ayant été saisi dans la chaîne, afin de déterminer s'il répond aux attentes ou s'il constitue une incohérence à signaler.

Nous parlons de mots ici, mais le terme token est davantage approprié. Comme nous l'avons vu dans la première partie (*cf.* § *Tokenisation* p. 27), token désigne une séquence de caractères du texte délimitée par des blancs ou des ponctuations (ce qui correspond à la définition que nous

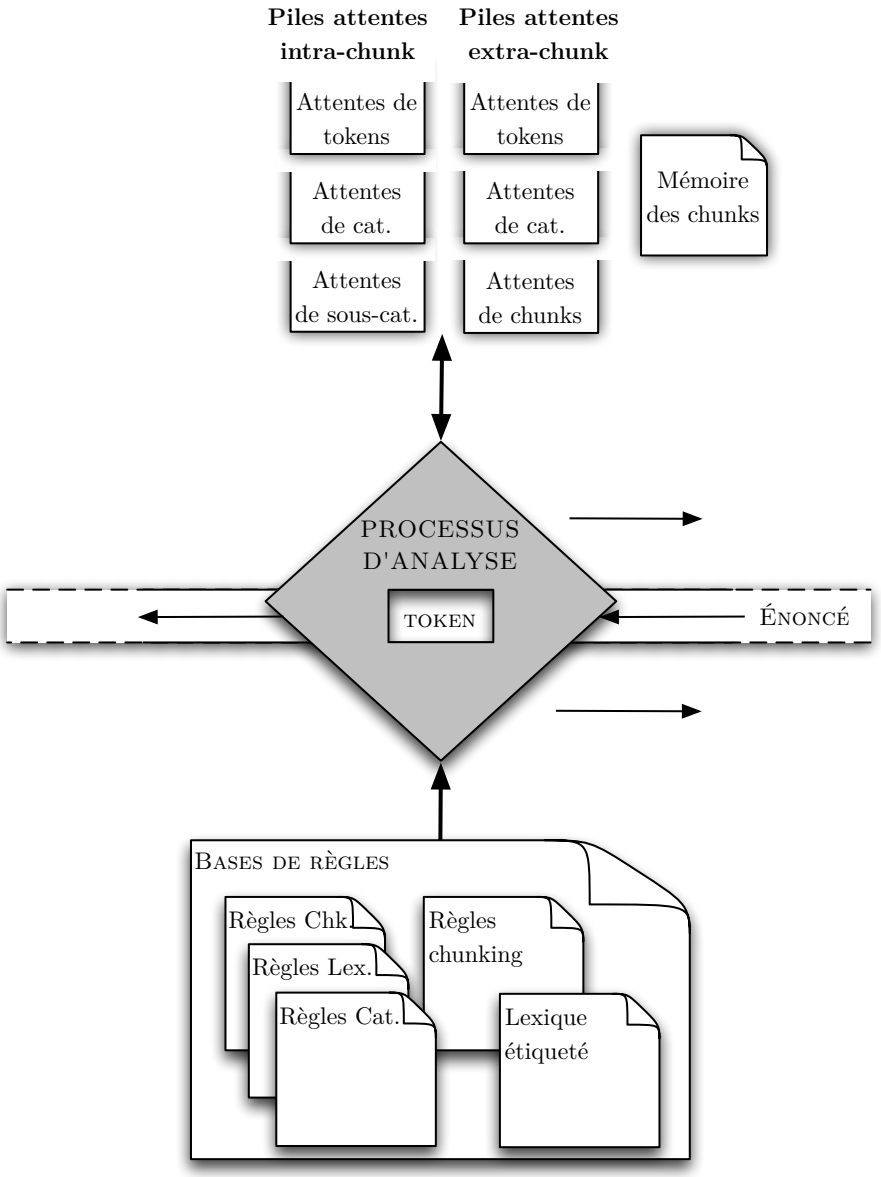


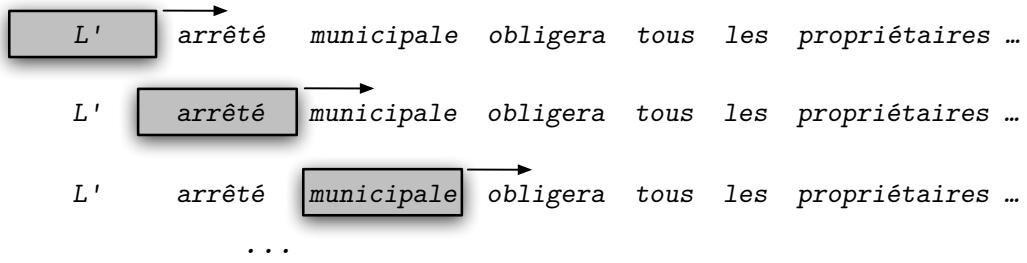
Figure 8.1 : Modèle pour la vérification grammaticale automatique

avons donnée du mot (*cf.* définition 1.4 p. 17), mais peut désigner également une ponctuation.

Afin de nous concentrer sur la détection d'incohérence grammaticale, nous partons du principe que la segmentation en tokens de la chaîne faisant l'objet d'une analyse par le modèle a d'ores et déjà été effectuée. Ceci implique que le modèle de vérification grammaticale prend en entrée, non pas une chaîne de caractères qu'il doit segmenter au fur et à mesure pour délimiter les différents tokens, mais une suite de tokens qui se présentent dans leur ordre linéaire d'apparition dans la phrase.

Il est a noter que cette tokenisation pourrait être effectuée en suivant un principe d'attentes, comme des attentes de rupture de token, ou des attentes facultatives de complément de token pour délimiter des tokens complexes. Ceci permettrait à des suites de caractères comme `lors de` ou encore `bien que` de former un unique token facilitant par la suite l'analyse syntaxique.

Nous représentons ci-dessous le déplacement de la fenêtre d'observation du processus d'analyse d'un token à l'autre, et reprendrons cet exemple de manière filée pour illustrer chacun des traitements par la suite .

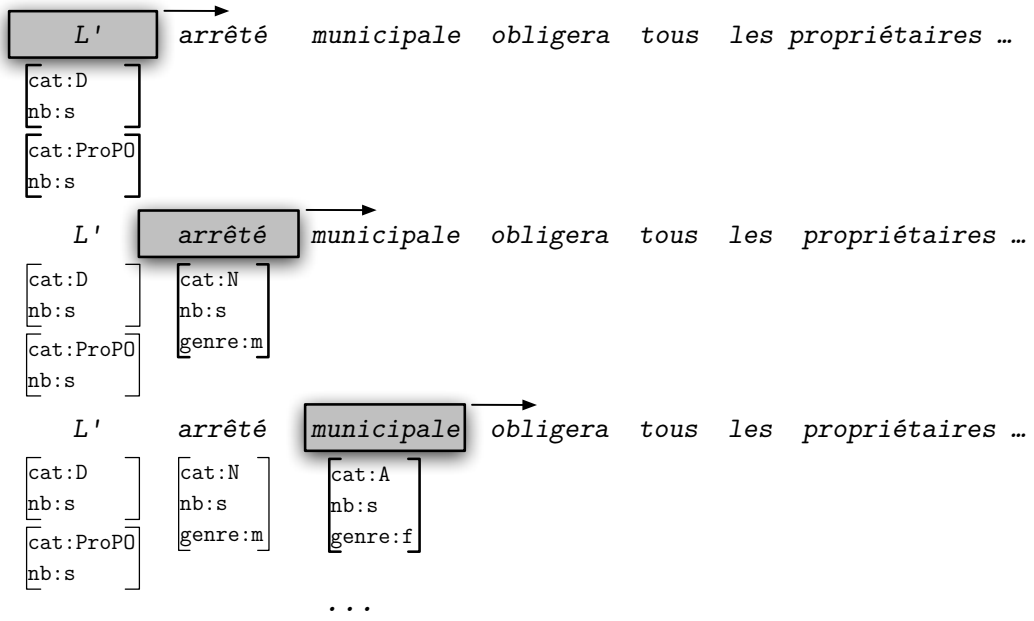


8.1.2 Étiquetage morphosyntaxique

Les tokens lus par le moteur d'analyse doivent être étiquetés morphosyntaxiquement. Cet étiquetage s'effectue de façon classique grâce à un lexique étiqueté (*cf.* § *Étiquetage morphosyntaxique* p. 30). Chaque token est comparé au lexique et reçoit l'étiquette correspondant à sa forme graphique. Ces étiquettes contiennent des informations de catégorisation et de sous-catégorisation.

Les tokens sont parfois ambigus au niveau de leur catégorie ou des traits de sous-catégorisation. Ils reçoivent dans ce cas traditionnellement plusieurs étiquettes qui font par la suite l'objet d'une désambiguïsation, par une méthode probabiliste ou par l'application de règles (*cf.* § *Étiquetage morphosyntaxique* p. 31). Notre principe d'analyse gauche-droite permet de répondre en partie à la question de l'ambiguïté. En effet, le dernier token traité va généralement attendre du suivant qu'il possède telle ou telle catégorie et/ou sous-catégorie. Si ce prochain token correspond à plusieurs étiquettes, seule celle répondant à l'attente générée par le précédent token sera attribuée.

Par exemple, le token `arrêté` peut être un nom ou le participe passé du verbe arrêter. Le mot grammatical `l'` qui le précède, lui-même ambigu (déterminant ou pronom personnel objet), a une attente pour un nom ou un verbe. L'étiquette *Participe passé* n'est par retenue car elle ne valide aucune des attentes, contrairement à l'étiquette *Nom* qui est alors attribuée au token. L'exemple ci-dessous illustre le résultat de l'étiquetage token après token.



Les étiquettes attribuées aux tokens sont utiles aux autres traitements parallèles (segmentation en chunks, gestion des attentes) mais ne constituent ensuite qu'une trace de l'étiquetage qui servira à vérifier le fonctionnement du mécanisme.

L'étiquetage morphosyntaxique est également l'occasion de procéder à une vérification orthographique, puisque les mots dont la graphie est inconnue du lexique ne peuvent être étiquetés. Ceci peut à nouveau être vu comme des attentes de graphies appartenant au lexique. Un token dont la graphie n'est pas dans le lexique doit être vérifié par le scripteur. Nous ne traitons pas cette question avec notre modèle.

8.1.3 Segmentation en *chunks*

Dans le chapitre précédent, nous avons mis en évidence que deux fois sur trois les erreurs se trouvent dans le même syntagme que leur référent (*cf.* § *Calculs des distances* p. 131). Il est donc intéressant d'avoir recours à cette unité pour délimiter des espaces de recherche d'incohérences, c'est-à-dire de gestion d'attentes.

Cette unité serait par ailleurs utilisée cognitivement, notamment lors de la production des accords (*cf.* section 7.3 p. 172), mais également lors de la lecture. Abney [1991, p. 257] la nomme *chunk* :

« I begin with an intuition : when I read a sentence, I read it a chunk at a time ».

a) Une unité syntaxique

Le chunk est une unité de type syntagme, définie par Abney [1991]. Elle fait partie des grammaires robustes du TAL dans lesquelles la notion de syntagme diffère de celle des grammaires chomskyennes. Dans Chomsky [1979], les syntagmes sont récursifs par exemple, ce qui signifie qu'ils peuvent être constitués de plusieurs syntagmes de même type. Ce n'est pas le cas des chunks, que l'on retrouve d'ailleurs dans la littérature française sous des appellations explicites

telles que *syntagme minimal* [Giguet, 1998], *syntagme non récursif* [Vergne & Giguet, 1998] ou *groupe non récursif* [Lebarbé, 2002], *syntagme réduit* [Vergne, 1998], *syntagme simple* [Déjean, 1998], *syntagme noyau* [Trouilleux, 2009], ou encore tout simplement *syntagme*, qui sera le terme que nous emploierons alternativement avec chunk. Dans l'exemple ci-dessous nous voyons la différence entre les trois syntagmes non récursifs (ou chunks), délimités par des [ ], et le syntagme récursif unique qu'ils composent à eux trois.

```
[ Les effluves ] [ d'une décoction ] [ de piments ]
```

Le chunk trouve son équivalent approximatif dans le groupe accentuel à l'oral. Il s'agit d'un segment à un niveau intermédiaire, situé entre le niveau du mot et le niveau de la phrase, qui s'avère par ailleurs « très stable d'une langue à l'autre » [Déjean, 1998, p. 118]. Cependant, il n'est pas reconnu comme une unité linguistique à part entière par toute la communauté de linguistes. Abney [1991, p. 257] en donne la définition suivante :

« The typical chunk consists of a single content word surrounded by a constellation of function words, matching a fixed template ».

Cette définition est en partie reprise en français dans Lebarbé [2002, p. 32] :

« Le chunk est constitué d'un mot lexical (au sens restreint : nominal ou verbal) entouré d'une constellation de mots fonctionnels (au sens large : déterminants, pronoms, adverbes, adjectifs...) ».

Un chunk est donc un groupe de mots contigus, réunis autour d'une tête lexicale dont ils dépendent. Ces relations de dépendance font de la structure interne des chunks une structure relativement figée. Par contraste, les chunks forment au sein des phrases des structures instables aux contraintes plus relâchées, mais liées par des relations de dépendance. L'extrait tiré du Bourgeois Gentilhomme de Molière est parfois utilisé pour illustrer les permutations possibles des chunks dans une phrase, en opposition au positionnement figé des constituants intra-chunks :

« [Belle Marquise], [vos beaux yeux] [me font] [mourir] [d'amour]
[D'amour] [mourir] [me font], [belle Marquise], [vos beaux yeux]
[Vos yeux beaux] [d'amour] [me font], [belle Marquise], [mourir] »

Les têtes lexicales qui caractérisent les chunks sont le plus souvent un nom (ou un pronom tonique) ou un verbe (conjugué, infinitif ou participe), plus rarement un adjectif ou un adverbe. Ces éléments centraux sont ensuite entourés d'éléments périphériques, énumérés par Vergne [1999, p. 6] :

- Dans le chunk nominal : conjonction de coordination et/ou de subordination, préposition, déterminant, adjectif épithète antéposé ou postposé, adverbe antéposé à l'adjectif épithète.
- Dans le chunk verbal : conjonction de coordination et/ou de subordination, préposition, tous les pronoms atones (sujet, objet et autres) antéposés ou postposés, négations, auxiliaire, adverbe le plus souvent postposé, adjectif attribut avec la copule être, adverbe antéposé à l'adjectif attribut.

b) Une unité de calcul

La composition des syntagmes n'est cependant pas la principale caractéristique intéressante de ces unités. La manière dont elles sont délimitées l'est bien davantage. En effet, les syntagmes sont délimités grâce aux mots grammaticaux, à la ponctuation ou aux marques morphologiques, et sont donc relativement faciles à définir, d'autant plus que, n'étant pas récursifs, ils ne peuvent pas s'imbriquer. La fin d'un chunk marque alors le début du chunk suivant. En général, un syntagme commence par un mot grammatical (qui détermine le type de syntagme) ou juste après une ponctuation, et se termine juste avant une ponctuation ou le mot grammatical suivant, qui marquent alors le début d'un nouveau chunk. Ainsi, les syntagmes ne sont pas définis en fonction de leur contenu, mais en fonction de marqueurs de début et de fin.

Le syntagme nominal, qui contient obligatoirement un nom (la tête lexicale), commence généralement par un déterminant. Cependant, ce dernier n'est pas pris en compte s'il est précédé d'une préposition, qui marque alors le début d'un chunk prépositionnel, et non nominal. Le syntagme verbal contient obligatoirement un verbe et commence généralement par celui-ci. Il contient également les pronoms personnels et adverbes de négation car ils sont fortement liés au verbe.

Le fait que les chunks soient calculés en fonction de leurs bornes de début et de fin, et non en fonction de leur contenu (ce qui impliquerait des énumérations de toutes les configurations internes possibles), fait de cette unité un outil intéressant pour les calculs syntaxiques en TAL. Le fait qu'ils présentent en outre une structure interne très contrainte ainsi que des dépendances entre eux font de ces syntagmes des unités pivot pour le calcul de structures syntaxiques.

Vergne & Giguet [1998] présente par exemple le potentiel de l'utilisation des chunks pour l'étiquetage morphosyntaxique. Selon lui, l'étiquette d'un mot (grammatical ou lexical) peut être déduite contextuellement à partir de l'étiquette du mot grammatical qui marque le début du chunk ainsi que son type. L'auteur donne pour exemple le mot grammatical `une`, étiqueté *déterminant*, qui ouvre un chunk de type nominal, et permet alors d'attribuer l'étiquette *nom* au mot lexical de catégorie ambiguë `ferme` :

`une ferme` : `une` = déterminant ⇒ `une ferme` = chunk nominal ⇒ `ferme` = nom

Un second exemple avec le mot grammatical `ne`, étiqueté *négation*, qui ouvre un chunk de type verbal, permet cette fois d'attribuer l'étiquette *verbe* au mot lexical `ferme` :

`ne ferme` : `ne` = négation ⇒ `ne ferme` = chunk verbal ⇒ `ferme` = verbe

Notons que la déduction contextuelle ne fonctionne pas dans toutes les situations, comme c'est le cas avec un troisième exemple que nous proposons et qui ne permet pas de désambiguïser `ferme` :

`la ferme` : `la` = déterminant ou pronom personnel ⇒ `la ferme` = chunk nominal ou verbal ⇒ `ferme` = nom ou verbe

Le relatif figement des éléments intra-chunk présente également un grand intérêt pour la gestion des accords, qui subissent de fortes contraintes. Par exemple, un chunk nominal « est marqué par un genre et un nombre homogènes sur tous ses composants variables en flexion » [Vergne, 1999, p. 6]. Il est alors aisé d'imaginer, dans le cadre de la vérification grammaticale, un contrôle des accords à l'intérieur de chaque chunk, via un mécanisme d'unification de traits (*cf.* § 8.2.6 p. 190)

Enfin, la mobilité des syntagmes dans la phrase ne les empêche pas d'entretenir des relations de dépendance, permettant d'élaborer la structure syntaxique de la phrase. Un syntagme dépend généralement de son prédécesseur, sauf dans le cas d'un chunk sujet (souvent le premier syntagme nominal à gauche). Celui-ci dépend du chunk verbal principal. Ces propriétés de dépendance vont s'avérer très utiles pour vérifier les accords des chunks nominaux sujets et verbaux.

Les chunks constituent donc une unité de calcul particulièrement adaptée à l'analyse syntaxique et au calcul de structures syntaxiques. Un de leur atout réside dans leur délimitation relativement facile en s'appuyant principalement sur les mots grammaticaux, les ponctuations, ou les frontières avec les syntagmes adjacents. Ce processus ne requiert ainsi pas d'énumération coûteuse des configurations possibles du contenu.

Leur structure interne très contrainte est également un atout. Centrée sur une tête lexicale de laquelle dépendent tous les mots fonctionnels inclus au chunk, la structure intra-chunk est relativement figée avec un ordre fixe des constituants et des relations de dépendance très fortes avec la tête. Il en découle un potentiel important de déduction contextuelle des catégories grammaticales des constituants lors de l'analyse syntaxique.

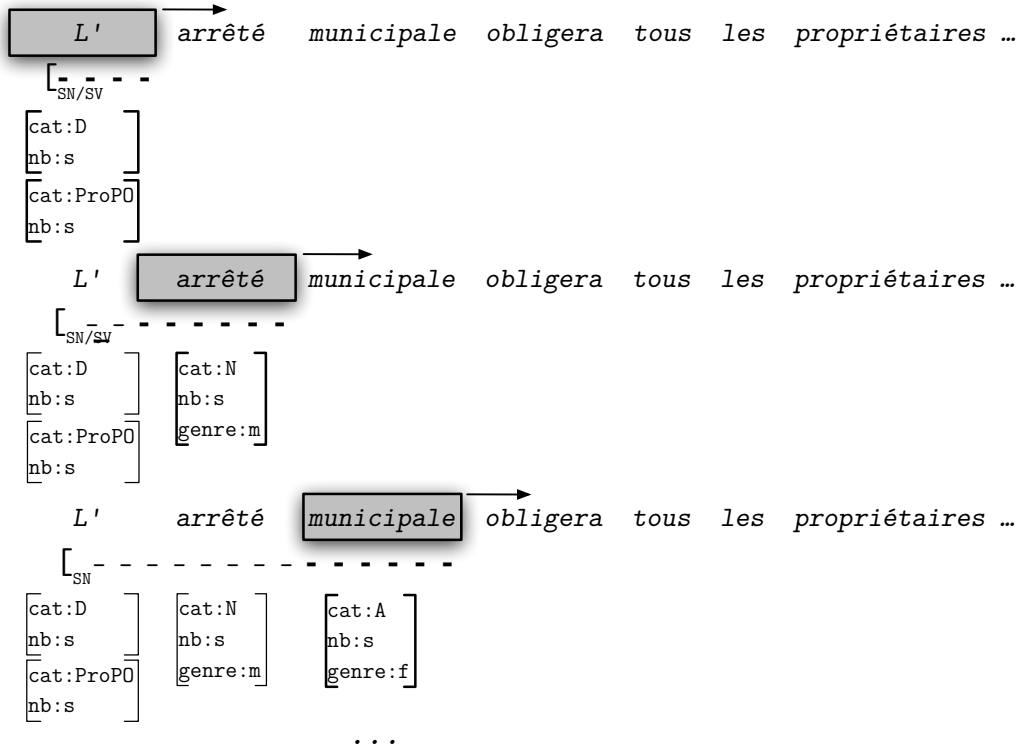
Enfin, les relations de dépendance entre les syntagmes, et notamment entre le syntagme sujet et le syntagme verbal principal, peuvent faciliter la vérification d'accords, via des opérations d'unification, même en situation de relation distante comme c'est souvent le cas entre le sujet et le verbe.

Nous avons donc intégré la segmentation en chunk à notre modèle, ce qui constitue une première distinction avec les vérificateurs libres (*cf.* § *Traitement facultatif : le* chunking p. 32), qui n'ont pas recours au chunk. Le modèle procède à la construction des chunks au fur et à mesure de la lecture des tokens, en se fondant principalement sur l'apparition des ponctuations et des mots grammaticaux pour déterminer les frontières, mais aussi sur les catégories des tokens « autorisées » dans un chunk. Si en cours de construction d'un chunk nominal un verbe apparaît, alors il doit être fermé et un chunk verbal doit être ouvert.

Nous illustrons cette segmentation page suivante, avec le même extrait que précédemment. Un chunk qui peut être nominal ou verbal est ouvert avec le premier token qui possède deux étiquettes. Les tokens suivants permettent de désambiguïser le chunk et lui sont inclus, jusqu'à ce qu'un token marque une rupture avec ce chunk. Ce sera le cas avec le verbe. Le chunk nominal sera alors fermé et un chunk verbal sera ouvert.

Il existe ainsi pour la délimitation des chunks un certain nombre de règles qui peuvent être réunies en une base utilisable par un moteur de segmentation. Les conditions d'application de ces règles sont constituées d'une part, du token lu, d'autre part, des informations stockées dans la *Mémoire des chunks* (voir figure 8.1 p. 177). Cette ressource construite par le processus d'analyse contient les informations concernant l'état du chunk en cours de construction (s'il vient d'être ouvert ou fermé par exemple), son type (nominal, prépositionnel, etc.) défini généralement grâce au token qui marque le début du chunk, et ses traits morphosyntaxiques (genre, nombre, etc.) déterminés d'après les traits de ses constituants. En fonction de ces données, le processus d'analyse décide si le token ouvre un chunk, le ferme, ou y est inclus. Il actualise ensuite la *mémoire des chunks* pour prendre en compte le résultat de la décision.

Notons que lorsqu'un token ouvre un chunk, cela induit généralement la fermeture du chunk précédent, dont les attentes qu'il génère doivent être prises en compte avant celles du nouveau token.



8.2 Des attentes aux piles

Les attentes sont au centre de notre conception de la vérification grammaticale. Nous avons vu que l'humain détecte les erreurs grâce à une comparaison entre le texte attendu et le texte produit (*cf.* § 7.2.1 p. 160). Nous avons donc émis l'hypothèse que l'humain qui révise un texte détecte une erreur lorsqu'un élément de la phrase ne valide pas une attente, au niveau morphologique, lexical, syntaxique, sémantique ou encore phonologique.

Nous retrouvons dans les grammaires de dépendance des notions proches de ces attentes avec les valences de Tesnière [1959] et les actants de Žolkovskij & Mel'čuk [1965].

8.2.1 Les valences de Tesnière

Le chapitre précédent nous a amenée à considérer les phrases en termes d'attentes. À la lecture d'un déterminant, nous nous attendons à rencontrer un nom, à la lecture d'un syntagme nominal singulier vraisemblablement sujet, nous nous attendons à voir apparaître un verbe conjugué à la troisième personne du singulier, etc. Ce phénomène est à rapprocher de la manière dont les grammaires de dépendance représentent la structure syntaxique d'une phrase.

Nous avons brièvement évoqué les grammaires de dépendance dans le premier chapitre (*cf.* § b) p. 10). Si le concept de dépendances apparaît dès l'antiquité (pour un rapide historique, voir [Kahane, 2001]), les travaux de Tesnière sont à l'origine de la première théorie linguistique se fondant sur la notion de dépendance syntaxique. Cette théorie considère que « dans une phrase, la présence de chaque mot (sa nature et sa position) est légitimée par la présence d'un autre

mot (son *gouverneur syntaxique*), à l'exception d'un mot, le mot principal associé au sommet de l'arbre syntaxique. La dépendance syntaxique est donc une dépendance entre mots » [Kahane, 2001, p. 3].

Pour Tesnière, les mots d'une phrase sont liés par des connexions structurales qui « établissent entre les mots des rapports de dépendance » [Tesnière, 1959, p. 13]. Le terme supérieur, ou régissant, commande le terme inférieur, ou subordonné. Les relations entre les mots d'une phrase sont représentées au sein d'arbres de dépendance, appelés « stemma » par Tesnière, au sein desquels des traits verticaux matérialisent les connexions entre régissants et subordonnés. Ces arbres font ressortir l'ordre structural des phrases, quand les arbres syntagmatiques font ressortir l'ordre linéaire. La figure 8.2 illustre la représentation d'un énoncé à l'aide d'un arbre syntagmatique, par opposition à la figure 8.3 qui représente ce même énoncé selon le concept de dépendance.

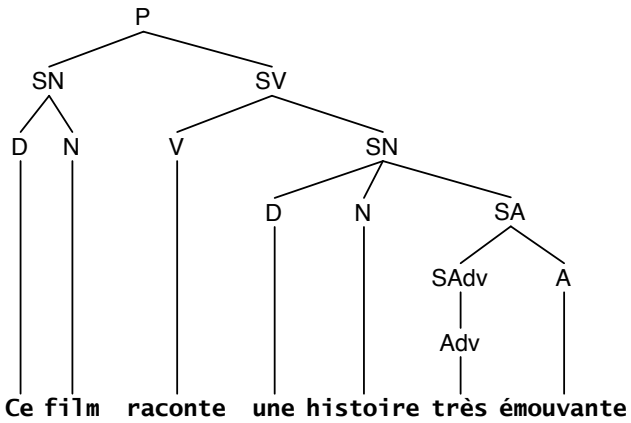


FIGURE 8.2 : Arbre syntagmatique

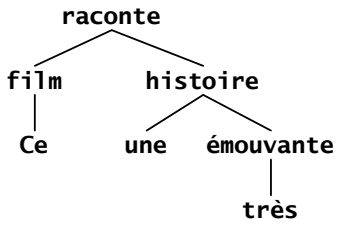


FIGURE 8.3 : Arbre de dépendance

La théorie de Tesnière, sur laquelle nous nous appuyons par la suite, est la principale à se fonder sur les dépendances. Parmi les concepts clé qu'il a apportés à la linguistique, il en est un qui nous intéresse particulièrement. Il s'agit des valences verbales.

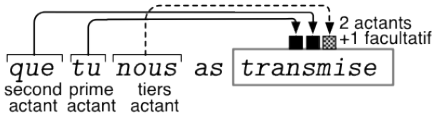
> « On peut ainsi comparer le verbe à une sorte d'**atome crochu** susceptible d'exercer son attraction sur un nombre plus au moins élevé d'actants, selon qu'il comporte un nombre plus ou moins élevé de crochets pour les maintenir dans sa dépendance. Le nombre de crochets que présente un verbe et par conséquent le nombre d'actants qu'il est susceptible de régir, constitue ce que nous appellerons la **valence** du verbe. »
> [Tesnière, 1959, p. 238]

Pour Tesnière [1959, p. 102], « Le **verbe** exprime le **procès**. [...] Les **actants** sont les êtres ou les choses qui, à un titre quelconque et de quelque façon que ce soit, même au titre de simples figurants et de la façon la plus passive, participent au procès ». Un verbe peut avoir jusqu'à trois actants, que Tesnière nomme « prime actant », « second actant » et « tiers actant ». Ils correspondent à ce que la grammaire scolaire désigne respectivement par « sujet », « Complément d'Objet Direct [1] » (COD) et « Complément d'Objet Indirect [2] » (COI). La valence est dite saturée lorsque tous les actants d'un verbe sont présents dans l'énoncé.

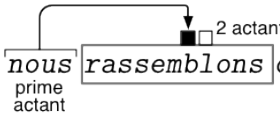
1. appelé simplement « complément direct » avant la réforme de la nomenclature grammaticale de 1910.
2. appelé simplement « complément indirect » avant la réforme de la nomenclature grammaticale de 1910.

Ainsi, dans l'énoncé (41), la valence de « transmettre » est saturée par les trois actants que ce verbe peut avoir au maximum (quelqu'un transmet quelque chose à quelqu'un). Les deux premiers (prime et second actants, soit le sujet et le COD) sont obligatoires pour ce verbe, mais pas le tiers actant qui peut être omis. Nous représentons cet actant facultatif par une flèche en pointillés pointant vers un carré rempli également de pointillés. De manière générale nous utilisons la même représentation [3] que celle utilisée pour représenter les attentes dans le chapitre précédent (*cf.* § a) p. 164). En effet, nous assimilons la notion de valence à la notion d'attentes que nous avons mentionnée à plusieurs reprises. Le fait qu'un verbe régisse un, deux ou trois actants revient à considérer, selon nous, que ce verbe attend un, deux ou trois actants. Autrement dit, en lisant un texte, si nous rencontrons par exemple le verbe « transmettre » de l'exemple (41) précédent, nous nous attendons à trouver dans la phrase un sujet et un ou deux compléments, ni plus ni moins. L'absence d'un de ces éléments va nous interpeler et nous permettre de localiser une erreur. Ainsi, si un actant obligatoire est manquant, c'est-à-dire si une attente obligatoire n'est pas comblée, l'énoncé est alors agrammatical. C'est le cas par exemple dans l'énoncé (42), où le verbe « rassembler », qui se construit avec deux actants, génère une attente obligatoire pour un sujet, qui est comblée (carré noir), et une autre non comblée (carré blanc) pour un COD ou un pronom réfléchi.

(41) *Je ne parle pas bien sûr de l'info* *que* *tu* *nous* *as* *transmise*

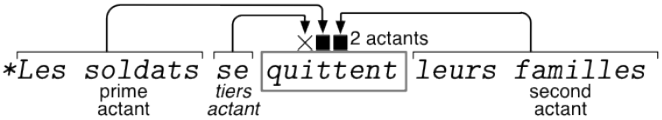


(42) **Puis nous rassemblons dans la pelouse qui est couvrie par la neige.*



Il peut au contraire y avoir des éléments superflus dans un énoncé, qui soit ne répondent à aucune attente, soit répondent à une attente déjà comblée par un autre élément. Nous avons un exemple de ce second cas dans l'énoncé (43), où le verbe « quitter » est construit avec trois actants, alors qu'il ne peut en recevoir que deux. Nous avons représenté le comblement de cette fausse attente, qui constitue une erreur grammaticale, non pas par un carré comme les attentes légitimes, mais par une croix.

(43) **Les soldats se quittent leurs familles*



Les cadres valenciels de plus de 3700 verbes les plus fréquents en français ont été décrits et réunis dans le dictionnaire DICOVALENCE [van den Eynde & Mertens, 2010]. Cette ressource langagière présente environ 8000 entrées pour chacune desquelles sont indiquées les actants et leurs propriétés. Ces informations présentent un grand intérêt pour l'analyse syntaxique et en particulier la détection d'incohérences grammaticales. Elles permettraient de vérifier que les verbes possèdent bien leurs dépendants obligatoires dans la phrase, et qu'ils sont construits avec les prépositions attendues le cas échéant. Nous présentons plus en détail ce dictionnaire dans la section *Ressources* p. 192.

3. Une attente est matérialisée par un petit carré positionné au-dessus de l'élément en attente, lui-même encadré. Lorsqu'une attente n'est pas comblée, le carré reste blanc. Lorsqu'au contraire une attente est satisfaite, le carré correspondant est rempli en noir et une flèche provenant de l'élément comblant l'attente pointe vers lui.

8.2.2 Les actants de Mel'čuk

Les principes de dépendance et de structure actancielle des verbes développés par Tesnière se retrouvent dans les travaux de Mel'čuk. La Théorie Sens-Texte [Žolkovskij & Mel'čuk, 1965], ébauchée par Mel'čuk et ses collègues russes dans les années 1960, est « une théorie linguistique visant la description de la correspondance Sens⇔Texte, au moyen de la construction de modèles formels. » [Polguère, 1998, p. 12]. Ces modèles permettent de générer l'ensemble des paraphrases exprimant un sens donné en entrée, en passant successivement par des représentations sémantiques, syntaxiques (profonde puis de surface), morphologiques (profonde puis de surface) et phonétiques (profonde puis de surface – la théorie considère la langue parlée).

Au centre du Modèle Sens-Texte de Žolkovskij & Mel'čuk [1967] se trouve le Dictionnaire Explicatif et Combinatoire (DEC). La particularité de ce dictionnaire est de contenir pour chaque entrée (« lexie[4] »), des données sémantiques (définition) et des données combinatoires (syntaxiques et lexicales). Ce sont ces dernières qui retiennent notre attention.

La combinatoire syntaxique met en œuvre la notion d'actant telle que nous la trouvons pour les verbes chez Tesnière [1959], mais l'étend aux autres catégories grammaticales. Elle décrit toutes les façons de réaliser les différents actants de la lexie vedette (par ex. actants obligatoires ou non, actant humain ou non humain, préposition à employer), ce qui correspond aux attentes de chaque lexie. La combinatoire lexicale nous ramène également au principe d'attentes et concerne les phénomènes collocationnels.

Nous ne rentrons pas davantage dans les détails de la Théorie et du Modèle Sens-Texte ici[5]. Nous les compléterons lors de la présentation plus précise du DEC dans la section *Ressources* p. 192. Bien qu'orientée du sens vers le texte, c'est-à-dire vers la synthèse et non l'analyse du langage, la Théorie Sens-Texte contient des informations pertinentes pour vérifier la grammaticalité des textes. Ce que nous en retenons notamment, c'est l'extension aux noms, adjectifs ou encore conjonctions de la notion d'actant, initialement consacrée au verbes chez Tesnière, de même que dans le dictionnaire DICOVALENCE. Le DEC, en proposant une description des actants possibles pour une lexie, ainsi que les façons dont ces actants se combinent avec la lexie, constitue une base d'informations qui pourrait s'avérer très intéressante pour l'application de vérification grammaticale que nous visons.

8.2.3 Des attentes de différents niveaux

Nous avons vu que les relations de dépendance et de rection traduisent en des termes linguistiques les phénomènes d'attentes qu'une unité d'une phrase peut avoir pour une autre unité de cette phrase. Elles traduisent également, sur un plan cognitif, les attentes que l'humain a pour certains éléments (lexicaux, morphosyntaxiques, sémantiques) à la lecture d'autres éléments. En conséquence de ces attentes, réalisées sous forme d'actants selon la terminologie des grammaires de dépendance, des inattendus peuvent apparaître dans une phrase. Il s'agira alors vraisemblablement d'une phrase agrammaticale, dans laquelle se trouvera un défaut d'utilisation d'un actant (actant manquant, mal construit, en trop).

4. « une ***lexie*** ou ***unité lexicale***, est soit un mot pris dans une acception bien spécifique (= ***lexème***), soit encore une locution, elle aussi prise dans une acception bien spécifique (= ***phrasème***) » [Mel'čuk *et al.*, 1995, p. 16].

5. Le lecteur intéressé pourra neanmoins se référer à Mel'čuk [1984, 1988b, 1992, 1997, 1999] et Mel'čuk *et al.* [1995].

Précédemment, en regard des erreurs de notre corpus, nous avons émis l'hypothèse que les attentes du scripteur sur son texte sont de différents types : attentes de flexions dans les syntagmes nominaux et verbaux, attentes syntaxiques, lexicales, sémantiques et phonologiques (*cf.* § 7.2.2 p. 163).

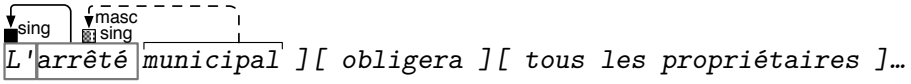
Les attentes sémantiques ne peuvent pour le moment pas être prises en compte par notre modèle, mais cela pourra être envisagé par la suite avec des ressources adaptées. Les autres attentes, quant à elles, peuvent être traduites en attentes adaptées à la modélisation. Elles doivent intervenir sur des niveaux et informations connues par le modèle, à savoir les différents niveaux de segmentation possibles (token et chunk) et les caractéristiques morphosyntaxiques des éléments de ces niveaux.

Telles que nous les envisageons, les attentes sont donc de plusieurs types. Nous effectuons une première distinction au niveau de leur portée. Nous distinguons les attentes intra-chunk et les attentes extra-chunk. Les premières englobent toutes les attentes créées par un élément pour un autre élément au niveau syntagmatique, c'est-à-dire un sein du même chunk. Les secondes correspondent aux attentes qui vont au-delà de la frontière du chunk, au niveau syntaxique.

La seconde distinction concerne le niveau de granularité sur lequel porte l'attente. Nous distinguons les attentes pour un chunk, les attentes pour un mot (ou token), les attentes pour une catégorie grammaticale et les attentes pour des traits de sous-catégorisation. Elles correspondent en grande partie aux attentes induites par les structures actancielles des mots.

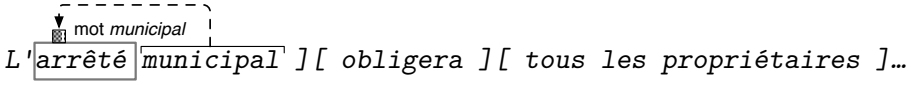
Il y a ainsi six types d'attentes différentes :

- Traits morphosyntaxiques attendus dans le même chunk :

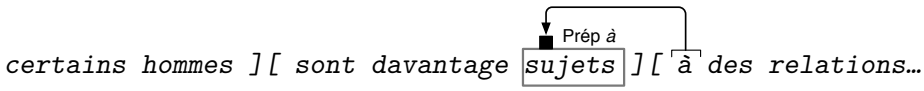
  Correspond à des attentes de flexions dans les syntagmes nominaux et verbaux.

- Lemme attendu dans le même chunk :

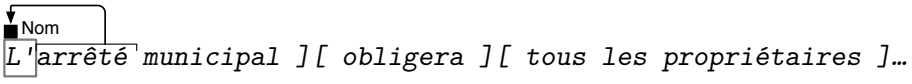
  Correspond en partie à des attentes lexicales, syntaxiques et phonologiques.

- Lemme attendu dans un autre chunk :

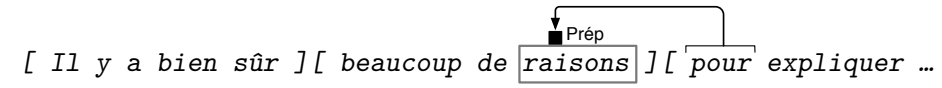
  Correspond en partie à des attentes lexicales, syntaxiques et phonologiques.

- Catégorie grammaticale attendue dans le même chunk :

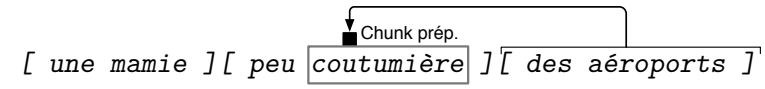
  Correspond en partie aux attentes de flexions dans les syntagmes verbaux, aux attentes syntaxiques et phonologiques.

- Catégorie grammaticale attendue dans un autre chunk :



Correspond en partie aux attentes syntaxiques et phonologiques.

- Autre chunk attendu :



Correspond en partie aux attentes de flexions dans les syntagmes verbaux, aux attentes syntaxiques et phonologiques.

De manière générale, les attentes de lemme et de catégorie peuvent permettre de détecter les erreurs de substitution de lemmes, en particulier lorsqu'elles impliquent une altération de la catégorie ce qui est généralement le cas avec les fréquentes confusions de mots grammaticaux. Elles permettent également de détecter les erreurs de syntaxe et les erreurs de mode verbal, elles aussi nombreuses, en particulier les confusions entre les finales en /E/. Les attentes de sous-catégories, quant à elles, permettent la détection des erreurs d'accord par le mécanisme d'unification.

8.2.4 Un traitement par piles

Qu'elles soient intra- ou extra-chunk, les attentes impliquent de se souvenir, lors de la lecture d'un token ou d'un chunk, de la ou des attente(s) créées auparavant. Vergne [1998, p. 23] décrit ce processus comme un « processus de mise en relation en réception » et le précise comme « un travail du récepteur (lecteur - auditeur) sur sa mémoire, un calcul présent sur la représentation présente d'évènements passés ». L'auteur donne en exemple la mise en relation entre le sujet et le verbe :

« en lisant - entendant un verbe :
a) le récepteur se souvient du sujet,
b) il relie le verbe au sujet,
c) puis il oublie ce sujet qui a trouvé son verbe, et n'en attend plus. »

Vergne [1998] appelle cette relation orientée du verbe vers le sujet « souvenance », pour la distinguer de la relation de dépendance, orientée du sujet vers le verbe.

En réalisant la vérification grammaticale de gauche à droite au fur et à mesure de la lecture des tokens, notre modèle, tout comme l'humain, doit se souvenir des attentes créées au fil du traitement pour vérifier ensuite si un élément y répond ou non. Afin de rendre cette « souvenance » possible, nous dotons le modèle de piles dans lesquelles seront empilées les attentes. Nous prévoyons une pile pour chacun des six types d'attentes présentés (voir figure 8.1 p. 177).

Précisons que les attentes ne sont pas limitées aux relations de dépendance ou aux relations de souvenance. Elles sont constituées des unes et des autres. Un sujet attend un verbe dont il dépend, mais un verbe attend un complément d'objet qu'il régit.

Chaque nouvelle attente vient se positionner au-dessus des autres attentes du même type dans leur pile. Lorsque l'attente du sommet de la pile (la dernière ajoutée) est comblée, elle n'est pas retirée de la pile mais étiquetée comme « comblée », afin de ne plus être considérée comme « en attente ». Nous dirons pour simplifier qu'une attente est « active » ou « activée » tant qu'elle

n'est pas comblée, et « inactive » ou « désactivée » dans le cas contraire.

8.2.5 Contenu des piles

En même temps que l'attente, l'élément qui en est l'origine ainsi que tout ce qui le concerne (le token ou chunk, ses catégorie et sous-catégorie et éventuellement sa fonction syntaxique) sont enregistrés. Nous appellerons cet élément « attendant », par opposition à l'élément « attendu ». Nous emploierons à présent le terme « attente » pour désigner le couple attendant-attendu.

Le fait d'enregistrer dans les piles l'attendant avec l'attendu permet au processus d'analyse de réaliser trois types d'actions. Tout d'abord, il dispose ainsi de toutes les informations nécessaires pour une rétroaction contextuelle si un défaut d'attente est détecté. Ensuite, lorsque plusieurs attendus possibles existent pour un même attendant, le fait de combler une des attentes conduit à désactiver les autres, qui sont facilement identifiables grâce à l'élément attendant. Enfin, si le scripteur effectue une modification du texte déjà analysé, en cas de correction suite au signalement d'une incohérence par exemple, les éléments des piles sont dépilés jusqu'aux attentes du dernier attendant non modifié. Par exemple, à l'analyse de *`*[...] utiliser tout un panoplie` :

- `utiliser` attend un chunk nominal objet ;
- `tout` attend à la fois un nom et du masculin singulier ;
- `un` comble l'attente de masculin singulier, et attend à la fois un nom et du masculin singulier ;
- `panoplie` comble l'attente de nom, mais pas de masculin singulier ;
- une incohérence est signalée au scripteur, qui décide de corriger `tout un` en `toute une` ;
- toutes les piles sont vidées jusqu'au niveau des attendus de `utiliser` ;
- l'analyse reprend avec `toute`.

Il se peut que plusieurs attentes soient ajoutées simultanément à une même pile comme des alternatives les unes des autres. Par exemple, la préposition `pour` peut attendre aussi bien un nom qu'un verbe infinitif. Dans les situations de la sorte, l'une des différentes alternatives devra être comblée, les autres étant alors désactivées. Nous les représenterons séparées d'un « | » (par ex. « `N|Vinf` »).

Il se peut également qu'une attente soit facultative, c'est-à-dire qu'elle ne doive pas être impérativement comblée pour ne pas déclencher de signalement d'incohérence. Il peut s'agir par exemple d'un auxiliaire qui pourra attendre un participe passé, mais pas obligatoirement. Dans ce cas, l'attente est ajoutée à la pile concernée comme l'alternative d'une attente de « rien », que nous représenterons par « ∅ ». Un auxiliaire, pour reprendre notre exemple, attendra ainsi soit un participe passé, soit rien : « `Ppé|∅` ».

Pour finir sur le contenu des piles, nous revenons sur le fait que la fonction syntaxique soit enregistrée en même temps que l'attendant. Cette information est utile pour le traitement des attentes des chunks verbaux, en particulier dans les phrases dont la structure ne suit pas l'ordre canonique (Sujet Verbe Objet en français). Si par exemple aucun sujet potentiel ne peut créer une attente pour un chunk verbal avant que celui-ci n'arrive, le chunk verbal doit créer une attente pour un sujet qui le suivra. Inversement, un chunk nominal peut constituer l'objet antéposé d'un chunk verbal, et doit être reconnu comme tel, en créant une attente de chunk verbal, pour que le processus d'analyse ne signale pas son absence. Nous retrouvons ces deux situations dans l'extrait suivant : `vu les décisions que prend notre fabuleux gouvernement.`

- Le pronom relatif `que`, qui attend un chunk verbal, remplit potentiellement le rôle de

complément d'objet, ce qui est indiqué dans la pile ;
- *prend* comble l'attente de chunk verbal initiée par un complément d'objet et n'a du coup pas besoin d'attendre de complément d'objet. En revanche, il ne constitue pas l'attendu d'un sujet, et doit donc en attendre un.
- *notre fabuleux gouvernement* constitue un chunk nominal potentiellement sujet et comble ainsi l'attente initiée par le verbe.

8.2.6 Portée des attentes

Une attente non comblée en temps voulu est le signe d'une probable incohérence et entraîne la génération d'une rétroaction contextuelle. Par « en temps voulu », nous faisons allusion au niveau dans lequel une attente doit être comblée. Par exemple, une attente de catégorie intra-chunk doit impérativement être comblée avant l'ouverture d'un nouveau chunk. De même, une attente extra-chunk doit être comblée avant la fin de la phrase.

Des différences existent cependant dans la gestion des piles selon leur type. Si les attentes de token et de catégorie doivent être comblées avant la fermeture du chunk dans les piles intra-chunk, ce n'est pas le cas pour les attentes de traits de sous-catégorie. Ces dernières servent à contrôler que le token lu possède bien les mêmes traits que ceux lus précédemment dans le même chunk et stockés dans la pile correspondante. Dans ce cas précis, nous avons recours au principe d'unification qui simplifie la vérification des accords. Le principe d'unification est issu des grammaires d'unification qui sont nées à la fin des années 70, en réaction aux grammaires génératives et transformationnelles chomskiennes. Tout en restant dans la lignée du courant générativiste de Chomsky, les grammaires d'unification se présentent comme des théories alternatives (voir Abeillé [1998, 2007] pour un historique plus complet). Elles sont le fruit « d'une collaboration entre syntacticiens et psycholinguistes insatisfaits du modèle transformationnel d'une part, logiciens et informaticiens à la recherche de formalismes pour le traitement automatique des langues d'autre part » [Abeillé, 1998, p. 24]. Elles ont d'ailleurs l'avantage d'être facilement implémentables.

Les formalismes d'unification présentent une approche de surface et considèrent la structure de surface de la phrase et non sa structure profonde. Ils sont également dits déclaratifs, et non procéduraux, en ce sens qu'ils décrivent les structures grammaticales et non les procédures qui permettent de les obtenir [Bouillon, 1998]. Les descriptions des unités de la phrase répondent toutes à un format unique de représentation, qu'il s'agisse de décrire le lexique, la syntaxe ou le sens : sous la forme de structures de traits. L'unification des ces structures permet d'enrichir les descriptions et de vérifier la compatibilité des traits communs aux deux structures objets de l'unification. Selon la théorie de Smedt [1990], l'unification serait d'ailleurs utilisée au niveau cognitif lors de la réalisation des accords (*cf.* § 7.3 p. 172).

a) Les structures de traits

Une Structure de Traits (ST) décrit un élément d'une phrase en énumérant ses caractéristiques linguistiques (lexicales, syntaxiques et/ou sémantiques) sous la forme de liste de couples attribut-valeur (ou trait-valeur). L'exemple de la figure 8.4 illustre une ST pour n'importe quel substantif masculin singulier. Un attribut *Cat* prend pour valeur la catégorie de l'unité à laquelle est associée la ST, ici un nom. Les traits *Genre* et *Nombre* et leurs valeurs respectives forment une sous-ST qui constitue elle-même la valeur de l'attribut *Accord*. Nous voyons ici un exemple de la possibilité d'enchâssement et de récursivité des traits. Un trait peut ainsi avoir comme valeur une ST.

$$\begin{bmatrix} \texttt{Cat} : & \text{N} \\ \texttt{Accord} : & \begin{bmatrix} \texttt{Genre} : & \text{masc} \\ \texttt{Nombre} : & \text{sing} \end{bmatrix} \end{bmatrix}$$

FIGURE 8.4 : Exemple de structure de traits pour un nom masculin singulier

Une autre caractéristique des ST est de présenter les informations sous forme d'une liste structurée mais non ordonnée. L'ordre dans lequel apparaissent les paires trait-valeur est libre. Enfin, les ST peuvent présenter des informations linguistiques partielles. Pour un déterminant comme *des*, qui peut être masculin ou féminin, le trait de genre ayant une valeur indéterminée n'a pas besoin d'être spécifié. La figure 8.5 donne deux exemples de ST possibles :

$$\begin{bmatrix} \texttt{Cat} : & \text{D} \\ \texttt{Accord} : & \begin{bmatrix} \texttt{Nombre} : & \text{pluriel} \end{bmatrix} \end{bmatrix} \qquad \begin{bmatrix} \texttt{Lex} : & \text{des} \\ \texttt{Cat} : & \text{D} \\ \texttt{Nombre} : & \text{pluriel} \end{bmatrix}$$

FIGURE 8.5 : Exemples de structure de traits pour le déterminant *des*

b) L'unification

« L'unification intervient lorsqu'on veut enrichir la description d'un objet en combinant les informations de deux ST qui le décrivent partiellement » [Bouillon, 1998, p. 94]. Il s'agit de la principale opération de comparaison et de combinaison des ST dans les formalismes d'unification.

Cette opération consiste à comparer les valeurs que possèdent les traits communs à deux ST différentes et à combiner ses traits en une nouvelle ST lorsque les valeurs ne sont pas contradictoires. Si les traits sont compatibles, une nouvelle ST est créée, formée de l'union des traits des deux ST initiales. Dans le cas contraire, on dit que l'unification échoue. L'union de deux ST A et B est notée A ∪ B, tandis que l'échec de l'unification est noté ⊥ [Abeillé, 2007, p. 40].

La figure 8.6 illustre une unification entre les ST décrivant les genre et nombre d'un déterminant et d'un nom, qui résulte en la ST de l'accord du syntagme nominal formé de ces constituants. La figure 8.7 illustre au contraire l'échec de l'unification des ST du même type que précédemment, mais où les traits de nombre ne sont pas compatibles et révèlent une agrammaticalité du syntagme nominal correspondant.

$$\begin{matrix} \textit{Les :} \\ \begin{bmatrix} \texttt{Nombre} : & \text{pluriel} \end{bmatrix} \end{matrix} \cup \begin{matrix} \textit{effluves :} \\ \begin{bmatrix} \texttt{Genre} : & \text{masculin} \\ \texttt{Nombre} : & \text{pluriel} \end{bmatrix} \end{matrix} = \begin{matrix} \textit{Les effluves :} \\ \begin{bmatrix} \texttt{Genre} : & \text{masculin} \\ \texttt{Nombre} : & \text{pluriel} \end{bmatrix} \end{matrix}$$

FIGURE 8.6 : Exemple d'unification qui réussit

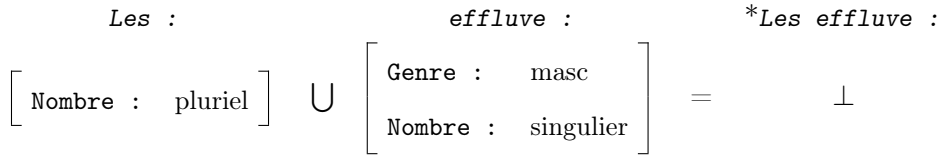


FIGURE 8.7 : Exemple d'unification qui échoue

Les ST représentent donc un moyen d'énumérer les caractéristiques des unités syntaxiques sous la forme de listes non ordonnées de couples attributs-valeurs. Elles peuvent être notamment le résultat d'un étiquetage morphosyntaxique, préalable incontournable à la vérification grammaticale. Ces ST présentent alors l'avantage de pouvoir contenir des descriptions partielles, qui pourront éventuellement être complétées par la suite, mais qui permettent surtout d'alléger les étiquettes associées aux unités traitées en cas d'ambiguïtés ou de données indéfinies.

Les ST sont également intéressantes en TAL par leur facilité d'implantation. Il en va de même pour le mécanisme d'unification, grâce auquel les ST de deux unités devant s'accorder peuvent être comparées, en permettant ainsi de détecter des anomalies d'accord. Le principal intérêt de l'unification se trouve donc dans la vérification des accords.

Dans le cadre de notre modèle, à chaque nouveau token d'un chunk, ses traits sont comparés à ceux du sommet de la pile, par un mécanisme d'unification. Les traits d'un constituant d'un chunk attendent en effet du suivant qu'il possède des traits concordants.

La vérification que l'attente est comblée, c'est-à-dire la vérification de la pile, consiste donc à réaliser ce calcul d'unification. Il peut avoir deux issues :

- si les traits du token et les traits de la pile présentent des discordances, alors le processus de rétroaction est activé ;
- si les traits du token et les traits de la pile ne présentent pas de discordance, alors les traits de la pile sont désactivés et les traits résultant de l'unification sont ajoutés.

Lorsque le chunk est fermé, tous les attendus de traits de sous-catégorie sont désactivés.

Dans les piles extra-chunk, le traitement des traits morphosyntaxiques diffère des piles intra-chunk. Ils n'ont pas de pile propre. Lorsqu'une catégorie ou un chunk sont attendus, les traits sont indiqués en même temps dans la pile. Ainsi, nous pouvons avoir une attente pour un chunk verbal à la troisième personne du singulier, ou pour un nom au pluriel. Les attentes de token ou de catégorie sont désactivées dès que l'élément attendu est lu. En revanche, les attentes de chunks nécessitent que le chunk attendu soit fermé pour être considérées comme comblées, et donc désactivées.

8.3 Ressources

Le processus d'analyse nécessite différentes ressources, qui sont principalement des bases de règles du type condition(s)-action(s) intervenant pour déterminer les attentes ou opérer une segmentation en chunks (voir figure 8.1 p. 177). Le token en cours de lecture, son chunk et/ou le contenu des piles d'attentes constituent les conditions permettant l'application d'une règle et la réalisation de l'action qu'elle prescrit.

8.3.1 Règles de segmentation en chunks

Un ensemble de règles est dédié à la segmentation en chunks. En s'appuyant sur des faits (token lu et chunk en cours de construction), les règles permettent au moteur d'inférence de prendre une décision parmi trois possibles :

- ouvrir un nouveau chunk avec le token lu,
- fermer le chunk en cours avant le token lu,
- inclure le token au chunk en cours.

Si nous prenons l'exemple d'un token déterminant, il ouvre un chunk nominal, sauf s'il est précédé d'une préposition. En termes de règles, ceci pourrait être énoncé par une première règle restrictive du type :
1 : « *Si le token est un déterminant, si un chunk prépositionnel est ouvert, alors inclure le token au chunk* ».
complétée par une seconde règle générale s'appliquant dans tous les autres cas :
2 : « *Si le token est un déterminant, alors ouvrir un chunk nominal.* »

Ces règles peuvent être formalisées de la façon suivante :
1 : `« $Det$ [SP → - »`
2 : `« $Det$ → [SN »`

La partie gauche contient la ou les conditions. Elle est au minimum constituée de la catégorie du token en train d'être lu, présentée entre deux $. Dans la règle 1, une seconde condition est qu'un syntagme prépositionnel soit ouvert (symbolisé par un crochet ouvrant). La partie droite contient l'action. Le tiret signifie l'inclusion du token au chunk ouvert, et donc aucune segmentation. Les crochets ouvrants ou fermants symbolisent une ouverture ou une fermeture de chunk (ici l'ouverture d'un chunk nominal).

Les règles sont présentées de la plus restrictive à la plus générique, et dès que l'une d'entre elles est appliquée, le processus s'interrompt jusqu'au token suivant.

8.3.2 Ressources pour les attentes

Outre les règles de segmentation en chunks, le processus d'analyse utilise des ressources de deux types pour gérer les attentes : des piles, que nous avons déjà présentées, et des bases de règles. Chaque base de règles est spécialisée dans un niveau grammatical donné. Lorsqu'un token est lu, sa catégorie, son lemme et son chunk peuvent générer des attentes déterminées grâce aux règles respectivement des niveaux morphosyntaxique, lexical et syntagmatique.

Nous n'avons pas mentionné les traits de sous-catégorie car ils ne requièrent pas de base de règles pour leurs attentes. Celles-ci sont traitées par un mécanisme d'unification effectué par le processus d'analyse (*cf.* § 8.2.6 p. 190).

a) Attentes de la catégorie

Les attentes créées par la catégorie grammaticale sont déterminées par le biais de règles, et gérées à l'aide des différentes piles. Le processus d'analyse procède à la vérification du contenu des piles à chaque nouveau token lu, puis s'appuie sur la base de règles pour mettre à jour les piles avec d'éventuelles nouveaux attendus.

Une règle dont les conditions portent sur la catégorie peut être du type :
1 : « *Si la catégorie est Préposition alors ajouter une attente de Nom ou de Verbe infinitif dans le même chunk.* »
ou
2 : « *Si la catégorie est Conjonction de subordination alors ajouter une attente de chunk verbal.* »
dont une formalisation possible serait :
1 : `« $Prep$ → (Cat-I) N|Vinf »`
2 : `« $Sub$ → (Chk) SV »`

Dans la partie droite de la règle, la pile dans laquelle doit être ajoutée l'attente est indiquée entre parenthèses. Le `I` (et le `E`) indiquent si l'attente doit être ajoutée dans la pile Catégorie Intra- ou Extra-chunk. Les catégories `N` et `Vinf` doivent être ajoutées à la pile en une seule et même attente car c'est l'un des deux qui est attendu.

b) Attentes du lemme

Les attentes créées par la forme lexicale du token ne sont pas génériques comme les précédentes. C'est le lemme et non sa catégorie ou ses traits qui a des attendus, ce qui nécessite de disposer de règles pour les nombreux lemmes ayant des actants. Des ressources lexicographiques comme le DEC [Mel'čuk, 1984, 1988b, 1992, 1999] seraient utiles à cette fin, en décrivant les types d'actants et leur régime de construction.

Par exemple, dans le DEC, la lexie « outil » possède deux actants (X et Y, voir tableau 8.1), qui constituent des compléments du nom. La zone de combinatoire syntaxique décrit comment ils se réalisent. L'actant X peut être réalisé sous la forme de la préposition *de* suivie d'un nom (« *les outils* ***de Gégé*** ») ou sous la forme d'un adjectif possessif (« ***ses*** *outils* »). L'actant Y peut être réalisé sous la forme de la préposition *de* ou *pour* suivie d'un nom (« *un outil* ***de jardinage*** », « *un outil* ***pour la gravure*** ») ou sous la forme de la préposition *à* suivie d'un verbe à l'infinitif (« *un outil* ***à couper le bois*** ») [Mel'čuk, 1999].

Ces constructions actancielles (ou schémas de régime) décrites dans les articles du DEC représentent les attentes créées par chaque lexie. En lisant le mot « outil », on s'attendra ainsi à trouver à sa suite les prépositions *de*, *pour* ou *à* pour introduire un complément du nom, alors qu'une préposition comme *par* ou *avec* sera un indice d'une possible erreur.

I. *Outil de X pour Y* = Artefact α destiné à ce qu'une personne X l'utilise, dans le cadre de l'activité Y qui est la fabrication/construction/ réparation de quelque chose [...]

X = 1	Y = 2
1. *de* N	1. *de* N
2. A_{poss}	2. *pour* N
	3. *à* V_{inf}

Tableau 8.1 : Extrait de définition et tableau de régime de l'acception I de la lexie « Outil » [Mel'čuk, 1999]

La combinatoire lexicale également ramène au principe d'attentes en s'intéressant aux phénomènes collocationnels. Par le biais de « fonctions lexicales », la zone de combinatoire lexicale

décrit les collocations dans lesquelles intervient la lexie dont il est question. La fonction lexicale `Magn` est fréquemment donnée en exemple dans la littérature sur les modèles Sens-Texte et les fonctions lexicales. Cette fonction permet de décrire les intensificateurs correspondant à chaque lexie à laquelle elle est associée, intensificateurs qui sont souvent spécifiques à telle ou telle lexie. Ainsi, on dira par exemple « *un gros chagrin* » [6] mais pas « **un gros amour* » [7] , ou encore « *gravement/griévement blessé* » [8] et non pas « **très blessé* ». Avec l'existence de ces collocations, nous avons des attentes pour certains mots en fonction d'autres que nous rencontrons.

Les informations fournies par le DEC seraient pertinentes pour une utilisation par notre modèle. Ce dictionnaire a d'ailleurs été conçu dans une optique d'informatisation à l'usage du TAL. Cependant, cette informatisation n'est à ce jour pas encore pleinement disponible. L'utilisation d'une telle ressource langagière n'est donc pas encore tout à fait d'actualité, mais plusieurs projets ont néanmoins vu le jour autour du DEC et de la théorie afférente [Mel'čuk *et al.*, 1995]. Le plus récent est le projet RELIEF qui développe le Réseau Lexical de Français [Lux-Pogodalla & Polguère, 2011 ; Polguère, 2012], mais la ressource n'est pas encore distribuée. Un précédent projet, DiCo, a développé « une sorte de version simplifiée et informatisée du [DEC] » [Jousse & Polguère, 2005, p. 6]. Le résultat est consultable en ligne [9]. Cette ressource a été conçue en vue d'un traitement informatique, ce qui pourrait en faire une bonne candidate à adapter pour la gestion des attentes de lemmes.

Le dictionnaire DICOVALENCE [van den Eynde & Mertens, 2010], ou sa version simplifiée Dicovalence-Easy [Gardent & Vandenberghes, 2009] constituent également des ressources potentiellement utilisables. DICOVALENCE possède la particularité de se fonder sur l'« Approche Pronominale » en syntaxe pour décrire les cadres valenciels de plus de 3700 verbes les plus fréquents en français. Sans rentrer dans les détails de cette approche (voir [van den Eynde & Blanche-Benvéniste, 1978 ; van den Eynde & Mertens, 2003]), nous reprenons la présentation succincte proposée en introduction de [van den Eynde & Mertens, 2010] : « d'abord pour chaque place de valence (appelée « paradigme ») le dictionnaire précise le paradigme de pronoms qui y est associé et qui subsume les lexicalisations possibles ; ensuite, la délimitation d'un cadre de valence, appelé « formulation », repose non seulement sur la configuration (nombre, nature, caractère facultatif, composition) de ces paradigmes pronominaux, mais aussi sur les autres propriétés de construction associées à cette configuration, comme les « reformulations » passives. »

En d'autres termes, les actants ne sont pas énumérés sous forme lexicalisée, mais par les pronoms correspondants (ayant les mêmes traits morphosyntaxiques). van den Eynde & Mertens [2003, p. 3] expliquent que « le nombre restreint de pronoms permet de vérifier de façon systématique et exhaustive leurs combinaisons avec les prédicateurs [unités lexicales qui régissent un ou plusieurs actants] ».

La figure 8.8 présente une entrée du verbe `bricoler`. Le verbe possède deux actants, ou deux « paradigmes » - P0 et P1 - qui peuvent se réaliser sous une des formes pronominales énumérées respectivement aux lignes `P0$` et `P1$`. La ligne `FRAME$` quant à elle indique de manière concise pour chaque actant : sa fonction syntaxique (sujet, objet, etc.), son caractère obligatoire ou facultatif, ses réalisations syntagmatiques possibles (pronom, groupe nominal, infinitive, etc.) et des restrictions de sélection (traits sémantiques de type humain/non-humain, abstrait, etc.).

6. `Magn`(chagrin) = grand, gros < énorme, immense
7. `Magn`(amour) = grand < immense
8. `Magn`(blessé) = gravement, grièvement
9. Dicouèbe (accès par requêtes SQL) : http ://olst.ling.umontreal.ca/dicouebe/
Dicopop (accès grand public) : http ://olst.ling.umontreal.ca/dicopop/

```
VAL$     bricoler: P0 P1
VTYPE$   predicator simple
VERB$    BRICOLER/bricoler
NUM$     12440
EG$      les enfants ont bricolé une cage à lapin
TR_DU$   in elk zetten, provisorisch herstellen, opknappen
TR_EN$   knock up, tinker
FRAME$   subj:pron|n:[hum], obj:pron|n:[nhum,?abs]
P0$      qui, je, nous, elle, il, ils, on, celui-ci, ceux-ci
P1$      que, la, le, les, en Q, ça, ceci, celui-ci, ceux-ci
RP$      passif être, se passif
AUX$     avoir
```

FIGURE 8.8 : Une entrée de DICOVALENCE (*bricoler*)

Des ressources du type de celles présentées ici pourront venir compléter une base de règles dans laquelle les règles d'attentes de certains mots, en particulier les mots grammaticaux, pourraient être consignées. Nous pensons à des règles comme :
1 : « *Si le token est* à *alors ajouter une attente de Nom ou de Verbe infinitif dans le même chunk.* »
ou encore
2 : « *Si le token est* qui *alors ajouter une attente de chunk verbal.* »
que nous pouvons formaliser comme suit :
1 : « $à$ → (Cat-I) N|Vinf »
2 : « qui → (Cat-E) SV »

c) Attentes de chunk

Les attentes initiées par un chunk sont nécessairement des attentes extra-chunk. L'attente principale à laquelle nous pensons est l'attente d'un chunk nominal sujet pour un chunk verbal. Nous pouvons énoncer la règle suivante :
« *Si le chunk est le premier chunk nominal alors il attend un chunk verbal.* »
Dans la règle formalisée ci dessous
« SN /^[SN]* → (Chk) SV »
la condition « /^[SN] » signifie qu'il n'y a pas eu de chunk nominal (« ^[SN] ») depuis le début de la phrase (« / ») jusqu'au chunk en cours de traitement (« * »).

d) Mémoire des chunks

En plus des piles et des bases de règles, la gestion des attentes nécessite de connaître le type du chunk auquel appartient le token en cours de traitement, car les attentes ne sont pas les mêmes selon qu'il s'agit d'un chunk nominal ou d'un chunk verbal par exemple. L'état du chunk, c'est-à-dire notamment s'il est en cours de construction ou s'il vient d'être fermé, est également une information utile, tant pour la vérification des piles que pour leur actualisation. Par exemple, une attente intra-chunk non comblée alors que le chunk vient d'être fermé constitue une incohérence qui doit être signalée lors de la vérification. Par ailleurs, la fermeture d'un chunk permet d'actualiser les piles avec ses propres attentes. Enfin, les informations sur les chunks peuvent servir lors de la génération de certaines rétroactions.

La prise en compte de ces informations passe par l'utilisation de la ressource *Mémoire des chunks* que crée le processus d'analyse. À chaque nouveau chunk, tout ce qui le concerne (type, contenu, traits, fonction) est enregistré dans cette mémoire, en venant s'empiler au-dessus des chunks précédents. Outre l'accès aux informations du chunk en cours de construction, cette ressource donne également accès aux chunks antérieurs afin de les mettre en relation avec le chunk en cours, sachant qu'un chunk dépend généralement de son prédécesseur. Nous pensons par exemple au cas d'un chunk adjectival qui pourrait de la sorte être mis en relation avec un chunk précédent pour vérifier qu'ils s'accordent.

8.4 Fonctionnement attendu

Afin de vérifier que notre modèle est capable d'analyser des énoncés et surtout d'y détecter des incohérences, nous avons tenté de simuler son fonctionnement sur quelques extraits de notre corpus contenant des erreurs de différents types. Notre corpus revêt ainsi le rôle de support que nous lui attribuions dans le chapitre 4 (*cf.* § *Linguistique(s) de/sur corpus* p. 64) après avoir rempli celui d'apport pour l'analyse des erreurs.

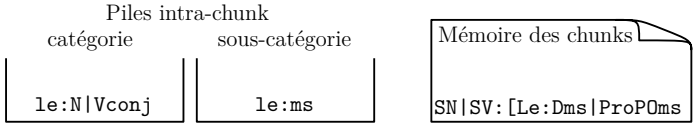
Dans les exemples que nous présentons ci-après, nous tentons d'indiquer pour chaque token les différents traitements effectués, à savoir : l'étiquetage, la segmentation en chunks, la vérification des attentes et leur détermination. Le premier exemple simule la détection d'une attente non comblée. Le second exemple concerne une détection d'erreur d'accord par échec de l'unification.

8.4.1 Exemple de détection par une attente non comblée

Prenons l'énoncé suivant, qui présente une erreur de confusion des mots grammaticaux `a`/`à`, avec en cascade une erreur de mode sur *`affirmé`* :
**Le tribunal de Paris a rendu un jugement qui tendrait a affirmé que non...*

1. {*Le*} est étiqueté Déterminant masculin singulier (`Dms`) et Pronom personnel objet masculin singulier (`ProPOms`).
   - Il est en début de phrase donc il ouvre un chunk, de type encore indéterminé, nominal ou verbal (`SN|SV`), car les 2 étiquettes déclenchent deux règles : `« $Det$ / → [SN »` et `« $ProPO$ / → [SV »`
   - Les piles sont vides donc aucune vérification n'a besoin d'être effectuée.
   - La catégorie Nom (`N`) est attendue dans le même chunk par la catégorie Déterminant (`D`)

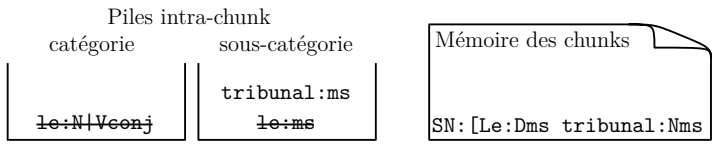
et est ajoutée à la pile (« D → (Cat-I) N »).
La catégorie Verbe conjugué (Vconj) est attendue dans le même chunk par la catégorie Pronom personnel objet (ProPO) et est ajoutée à la pile (« $ProPO$ → (Cat-I) Vconj »).
- La sous-catégorie masculin singulier (ms) est ajoutée à la pile.
- Les informations du chunk sont ajoutées à la mémoire des chunks.
- L'état des piles et de la mémoire des chunks à l'issue du traitement du token est le suivant :



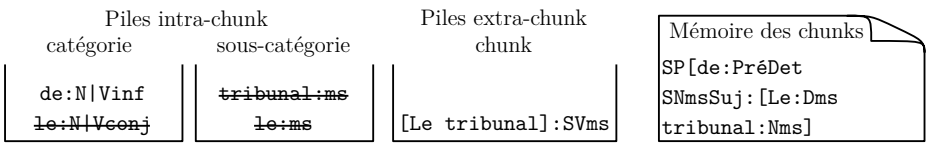
2. {*tribunal*} est étiqueté Nom masculin singulier (Nms).
   - Il est inclus au chunk ouvert dont le type peut être spécifié : chunk nominal (SN) (« N [SN →- »)
   - La catégorie N comble l'attente de la pile catégorie. L'attente est désactivée.
   - Les traits ms s'unifient avec succès avec les traits de la pile :

$$\underset{\textit{tribunal}:}{\begin{bmatrix}\text{genre:m}\\ \text{nombre:s}\end{bmatrix}} \cup \underset{\text{Pile ss-cat. :}}{\begin{bmatrix}\text{genre:m}\\ \text{nombre:s}\end{bmatrix}} = \begin{bmatrix}\text{genre:m}\\ \text{nombre:s}\end{bmatrix}$$

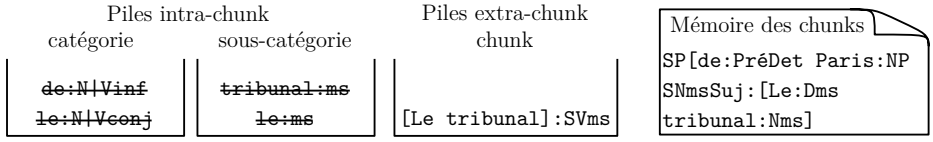
   - Le résultat de l'unification est ajouté à la pile.
   - Les traits de la pile sous-catégorie sont associés au chunk.
   - Les informations du chunk sont mises à jour dans la mémoire des chunks. (le chunk et *Le* sont désambiguïsés).
   - L'état des piles et de la mémoire des chunks à l'issue du traitement du token est le suivant :



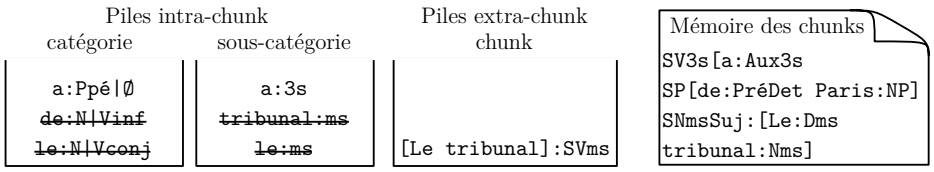
3. {*de*} est étiqueté Préposition déterminant (PréDet).
   - Il clôt le chunk nominal en ouvrant un chunk prépositionnel (SP).
   - Il est mis en attente le temps de traiter le chunk fermé.
     - Le chunk nominal clos ne comble pas d'attente de la pile chunk qui est vide.
     - Aucune attente n'est active dans les piles catégorie et mot intra-chunk donc pas d'incohérence.
     - Le chunk nominal clos est étiqueté comme sujet (Suj).
     - Il attend un chunk verbal masculin singulier (SVms) qui est ajouté à la pile.
     - Il est ajouté à la mémoire des chunks.
     - Les traits de sous catégorie de la pile intra-chunk sont désactivés.
   - {*de*} a une attente de catégorie N ou Verbe infinitif (Vinf) dans le même chunk.
   - Le chunk prépositionnel est ajouté à la mémoire des chunks.
   - L'état des piles et de la mémoire des chunks à l'issue du traitement du token est le suivant :



4. {*Paris*} est étiqueté Nom propre (NP).
   - Il est inclus au chunk prépositionnel en cours.
   - Il comble l'attente de N de la pile catégorie. L'attente est désactivée.
   - Les informations du chunk sont mises à jour dans la mémoire des chunks
   - L'état des piles et de la mémoire des chunks à l'issue du traitement du token est le suivant :



5. {a} est étiqueté Auxiliaire avoir indicatif présent 3e personne du singulier.
   - Il clôt le chunk prépositionnel en ouvrant un chunk verbal (SV).
   - Il est mis en attente le temps de traiter le chunk fermé.
     - Le chunk prépositionnel clos ne comble pas d'attente de la pile chunk.
     - Aucune attente n'est active dans les piles catégorie et mot intra-chunk donc pas d'incohérence.
     - Le chunk prépositionnel clos est ajouté à la mémoire des chunks.
     - Les traits de sous catégorie de la pile intra-chunk sont désactivés.
   - {a} ne comble pas d'attentes.
   - Il a une attente facultative pour un participe passé (Ppé) dans le même chunk, ajoutée à la pile.
   - Il a une attente de traits 3e personne singulier (3s), ajoutés à la pile.
   - Il donne au chunk verbal les mêmes traits que lui (3s).
   - Les informations du chunk sont mises à jour dans la mémoire des chunks
   - L'état des piles et de la mémoire des chunks à l'issue du traitement du token est le suivant :

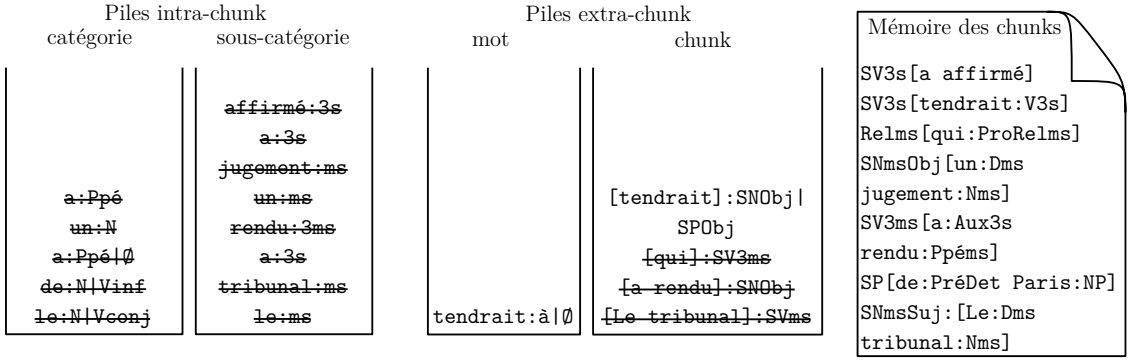


6. ...

   Le traitement est poursuivi de la sorte jusqu'à la conjonction *que* :

7. {*que*} est étiqueté Conjonction de subordination (Sub).
   - Il clôt le chunk verbal en ouvrant un chunk subordination (Ssub).
   - Il est mis en attente le temps de traiter le chunk fermé.
     - Le chunk verbal clos ne comble pas l'attente de chunk nominal ou prépositionnel objet créée par le chunk *[tendrait]*. Une incohérence doit être signalée au scripteur :
     - La génération de la rétroaction reprend les informations des attentes non comblées : « Après *tendrait* il faudrait un complément d'objet, qui peut-être introduit par la préposition *à*. Proposition : *tendrait à affirmer* »

– L'état des piles et de la mémoire des chunks à ce moment est le suivant :



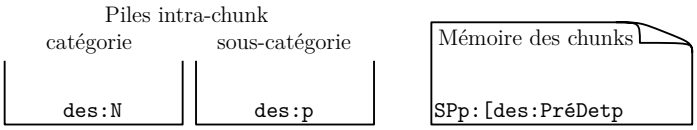
8.4.2 Exemples de détection par un échec d'unification

Les erreurs d'accord intra-chunk sont détectées grâce à un mécanisme d'unification, qui vérifie que les traits stockés dans la pile de sous-catégorie concordent avec les traits du token lu. Nous proposons un exemple de traitement aboutissant à la détection d'un erreur suite à un échec d'unification.

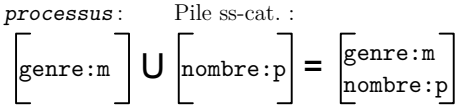
L'exemple concerne un accord au sein d'un syntagme nominal :
**il s'agit du plus bas niveau, des processus basique de tokenisation*

Nous ne présentons le traitement qu'à partir de *des processus* :

1. {*des*} est étiqueté Préposition déterminant pluriel (PréDetp).
   - Il ouvre un chunk prépositionnel.
   - Il ne comble pas d'attente.
   - Il crée une attente de N dans le même chunk.
   - La sous-catégorie p est ajoutée à la pile.
   - Il donne au chunk prépositionnel les mêmes traits que lui (p).
   - Les informations du chunk sont ajoutée dans la mémoire des chunks.
   - L'état des piles et de la mémoire des chunks à l'issue du traitement du token est le suivant :

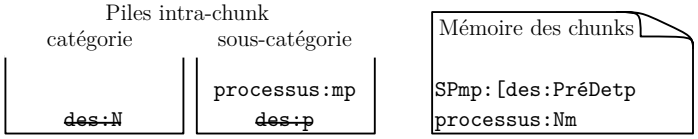


2. {*processus*} est étiqueté Nom masculin (Nm).
   - Il est inclus un chunk prépositionnel.
   - Il comble l'attente de N de la pile catégorie.
   - Ses traits et de ceux de la pile s'unifient avec succès :

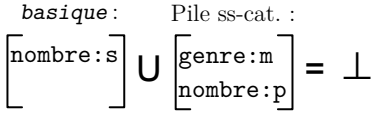


   - Le résultat de l'unification est ajouté à la pile.
   - Les traits de la pile sous-catégorie sont associés au chunk.
   - Les informations du chunk sont mises à jour dans la mémoire des chunks.

- L'état des piles et de la mémoire des chunks à l'issue du traitement du token est le suivant :



3. {*basique*} est étiqueté Adjectif singulier (As).
   - Il est inclus un chunk prépositionnel.
   - Il ne comble pas d'attente.
   - L'unification de ses traits et de ceux de la pile échoue car les valeurs des traits nombre sont différentes :



   - La génération de la rétroaction reprend les informations de l'unification qui a échoué et du chunk en cours :
     « *des processus* est au pluriel mais *basique* est au singulier. Proposition : *des processus basiques*. »

Les exemples traités dans cette section montrent que la détection d'erreurs est possible en se fondant sur le principe des attentes non validées. Ils montrent également des exemples de rétroactions qui peuvent être générées. Nous en présentons le fonctionnement dans la suite de cette section.

8.4.3 Rétroactions

Les rétroactions sont déclenchées si une attente n'est pas comblée. Elles ont pour but de prévenir le scripteur de la détection d'une incohérence, de lui indiquer sa localisation et de lui expliquer sa nature, afin qu'il puisse décider d'effectuer une modification de son texte ou non.

Il est primordial de rappeler qu'un vérificateur de texte tel que nous le proposons ne détecte pas une erreur mais une incohérence. Or toutes les remédiations de l'incohérence ne se valent pas. Il faudra présenter à l'utilisateur des informations pertinentes pour prendre sa décision. Les résultats de nos analyses peuvent alimenter une réflexion dans cette voie.

Par exemple, la prégnance de l'homophonie dans les erreurs (*cf.* § *L'homophonie dans les erreurs* p. 128) suggère qu'en cas d'erreur d'accord en nombre entre un déterminant et un nom, ce sera dans la plupart des cas le nom qui est mal orthographié. En effet, en français, la marque du pluriel est beaucoup plus souvent marquée phonétiquement sur les déterminants (un/des, le/les, etc.) que sur les noms.

Le processus de rétroaction consiste en une génération automatique de texte qui reprend diverses informations sur les élément impliqués dans l'incohérence détectée. Dans le cas d'une incohérence au niveau du token, sont repris :

- la donnée du token en cours de traitement qui n'a pas validé une attente : sa sous-catégorie pour une attente de la pile sous-catégorie, sa catégorie pour une attente des piles catégorie intra- ou extra-chunks, ou son lemme pour une attente des piles lemme intra- ou extra-chunks) ;

- l'élément attendu : une sous-catégorie, une catégorie, un lemme ou un chunk ;
- l'élément attendant : une sous-catégorie, une catégorie, un lemme ou un chunk.

Dans le cas d'une incohérence au niveau du chunk, sont repris :

- la donnée du chunk qui vient d'être fermé qui n'a pas validé une attente : sa sous-catégorie ou son type ;
- l'élément attendu : un type de chunk ou une sous-catégorie de chunk ;
- l'élément attendant : une catégorie, un lemme ou un chunk.

À partir de ces informations, un message peut être généré automatiquement. Pour une attente de catégorie non validée, il peut être du type :

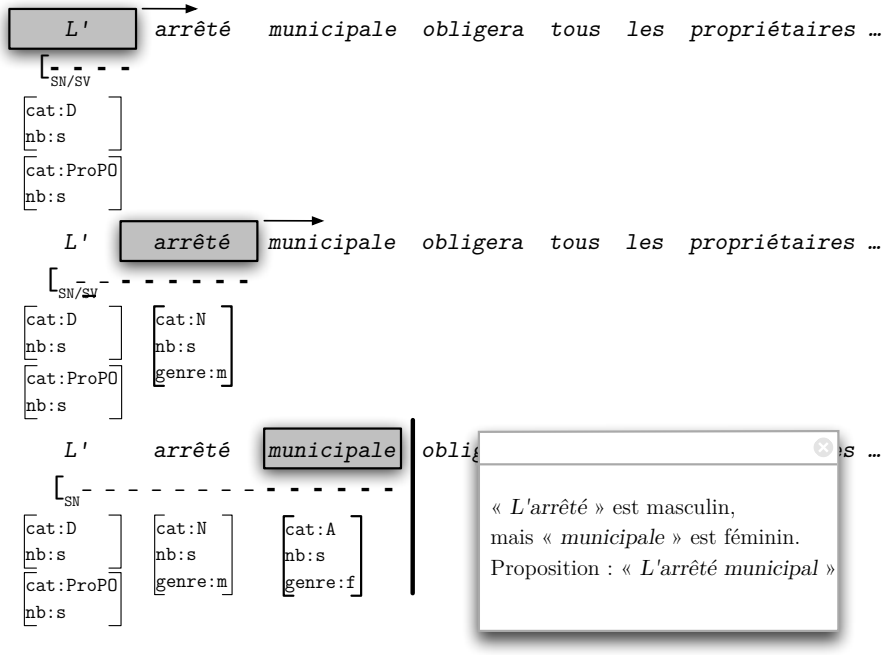
« `tokERR` est `Cat(tokERR)`. Après `Cat(tokATT)` `tokATT`, il faudrait `ATT`. »

où `tokERR` est le token ayant déclenché la rétroaction, `tokATT` est le token attendant, et `ATT` est l'élément attendu.

Par exemple, dans *`elle a atterrit`, l'auxiliaire attend un participe passé, et pas la forme conjuguée `atterrit`. Nous avons vu que ce type d'erreurs, qui consiste à conjuguer le participe passé des verbes des 2e et 3e groupes (*cf.* § Causes liées au scripteur p. 155), est assez fréquent et résulte sans doute d'une procédure cognitive d'automatisation. Dans ce cas, une rétroaction s'appuyant sur le token lu (`atterrit`) et sur l'attente (attendant : `a` ; attendu : *participe passé*) pourrait être :

« *atterrit* est un verbe conjugué. Après l'auxiliaire *a*, il faudrait un participe passé. »

Dans le cas d'une rétroaction suite à l'échec d'une unification, la génération de la rétroaction s'appuie sur les éléments ayant fait l'objet de la tentative d'unification pour contextualiser l'incohérence détectée. Pour cela, dans un cadre intra-chunk, il récupère le token lu ainsi que le chunk auquel il appartient, stocké dans la *mémoire des chunks*, afin de replacer le token dans son contexte immédiat. Il utilise ensuite les traits du token et ceux des autres constituants du chunks pour expliquer l'incohérence rencontrée. L'exemple ci-dessous illustre l'interruption du traitement et la rétroaction pour signaler une erreur d'accord sur *`municipale`* :



Suite à la rétroaction, si le scripteur décide d'effectuer une modification, l'analyse est reprise à partir de l'endroit de la modification.

Les exemples présentés dans cette section nous donnent une idée de ce que nous pouvons attendre de notre modèle, mais il ne s'agit que d'une simulation, et le traitement d'énoncés par un outil développé à partir de ce modèle sera susceptible de révéler des difficultés de gestion des attentes. L'efficacité d'un tel outil reposera en grande partie sur les ressources langagières, qui pourront la limiter. Nous abordons cette question dans les perspectives en conclusion de cette thèse. Nous reviendrons également de manière plus précise sur la question des rétroactions.

8.5 Conclusion

Dans ce chapitre nous présentons le modèle que nous avons élaboré pour la vérification grammaticale. Ce modèle présente l'originalité d'effectuer une analyse gauche-droite au fur et à mesure de l'écriture/la lecture. Il prend appui sur des concepts que nous avons fait émerger des erreurs de notre corpus, et dont nous trouvons des manifestations dans les mécanismes cognitifs de la production et de la révision d'écrit. Il s'agit des principes de dépendance, de segmentation en chunk et d'unification. Ils présentent distinctement divers intérêts à être utilisés pour la vérification grammaticale et sont d'ailleurs fréquemment employés dans le domaine du TAL. Leur utilisation conjointe ouvre encore d'autres possibilités.

Le principe des attentes, issu des grammaires de dépendance, est au centre du modèle. Il permet la détection d'incohérences grammaticales lorsqu'une attente n'est pas comblée, mais pourrait être étendu aux autres niveaux de traitements : attentes de ruptures pour la segmentation en tokens et en chunks et attentes d'étiquettes pour l'étiquetage morphosyntaxique. Ceci confère au modèle une certaine généricité.

Le découpage en chunks simplifie le traitement des attentes en délimitant la portée de ces attentes. Certaines attentes sont ainsi bornées au chunk et doivent y être validées. C'est le cas notamment des attentes d'accords, qui ont la particularité de pouvoir être vérifiées de façon très efficace en appliquant le principe d'unification. Combiner le découpage en chunks et l'unification de structures de traits simplifie la vérification des accords en restreignant la zone de calcul où tous les éléments doivent s'unifier au niveau de leurs traits morphosyntaxiques. En s'abstrayant des catégories grammaticales, il n'est plus nécessaire, comme dans les vérificateurs grammaticaux libres traditionnels (*cf.* § a) p. 43), d'avoir des règles de détection d'erreurs prévoyant toutes les combinaisons de mots pouvant constituer un syntagme. Par exemple, An Gramadóir [Scannell, 2003] contient environ 450 règles pour les accords dans les syntagmes nominaux.

La délimitation des syntagmes facilite également le traitement des attentes relatives aux actants. Il est plus facile de vérifier qu'un verbe possède bien tous ses dépendants, c'est-à-dire ses actants, si ceux-ci sont délimités dans des chunks (nominaux ou prépositionnels). Il est aussi plus facile de vérifier que ces actants sont bien construits, et plus généralement qu'il n'y a pas de mot de catégorie inattendue au sein d'un syntagme, puisqu'un chunk d'un certain type ne peut contenir que des mots d'un certain type (un chunk nominal ne contient pas de verbe par exemple). Ce sont ainsi, dans notre typologie, les erreurs de substitution de lemme, de mode verbal ou encore de syntaxe qui peuvent être détectées grâce aux attentes.

La détection des erreurs de grammaire passe par une bonne mise en relation des éléments dans la phrase. Des dépendances sont établies au sein des syntagmes, mais les syntagmes sont soumis aussi à des dépendances entre eux. Un chunk dépend ainsi généralement de son prédécesseur, sauf le chunk verbal principal. Celui-ci ne dépend d'aucun autre syntagme, mais régit au contraire des dépendances avec son sujet et ses compléments. Identifier les deux syntagmes sujet et verbal conduit à pouvoir tenter de les unifier pour vérifier qu'ils s'accordent correctement, comme pour les relations intra-chunk, et ce même s'ils sont très éloignés. Le découpage en syntagmes peut donc aider à résoudre le problème de détection de certaines fautes impliquées dans des relations distantes, auquel se heurte une grande partie des correcteurs.

Le modèle que nous avons proposé tire les bénéfices de l'utilisation conjointe des principes d'attentes, de chunks et d'unification. Il détecte les erreurs d'accords dans les syntagmes nominaux, qui sont les plus fréquemment commises (*cf.* § d) p. 122). Il détecte également les erreurs fréquentes d'accords sujet-verbe grâce à la délimitation des syntagmes et leur mise en relation. Les erreurs de confusion de lemmes, elles aussi fréquentes, sont détectables grâce aux attentes lorsqu'elles impliquent un changement de catégorie grammaticale, ce qui est souvent le cas dans les confusions de mots grammaticaux, mais moins avec les mots pleins. Ces derniers ne peuvent être détectés sans la sémantique. Par contre, les confusions courantes de mode verbal, et en particulier des finales en /E/ des verbes du premier groupe, sont identifiables car elles entraînent,elles aussi, un changement de catégorie grammaticale.

Conclusion et perspectives

Perspectives d'implantation du modèle

Sommaire

8.1 Structure du modèle 176
8.1.1 Mécanisme de lecture gauche-droite 176
8.1.2 Étiquetage morphosyntaxique 178
8.1.3 Segmentation en *chunks* 179
8.2 Des attentes aux piles 183
8.2.1 Les valences de Tesnière 183
8.2.2 Les actants de Mel'čuk 186
8.2.3 Des attentes de différents niveaux 186
8.2.4 Un traitement par piles 188
8.2.5 Contenu des piles 189
8.2.6 Portée des attentes 190
8.3 Ressources 192
8.3.1 Règles de segmentation en chunks 193
8.3.2 Ressources pour les attentes 193
8.4 Fonctionnement attendu 197
8.4.1 Exemple de détection par une attente non comblée 197
8.4.2 Exemples de détection par un échec d'unification 200
8.4.3 Rétroactions 201
8.5 Conclusion 203

Le modèle que nous avons présenté dans le chapitre 8 est une première étape pour la conception d'un outil original de vérification grammaticale fondé sur un principe d'attentes et une analyse gauche-droite. Il nous a permis de décrire dans les grandes lignes le fonctionnement d'une telle vérification.

Un processus d'analyse procède parallèlement à l'étiquetage morphosyntaxique des tokens, à la segmentation en tokens et à la gestion des attentes pour détecter des incohérences et générer des rétroactions. À tout moment du traitement d'un token, des attentes peuvent être créées, validées ou invalidées, à différents niveaux. En cas d'incohérence, l'analyse est interrompue pour générer une rétroaction, sans attendre que le traitement du token soit terminé. Il est même nécessaire que la rétroaction soit générée sitôt l'incohérence détectée afin de fournir à l'utilisateur les dernières informations traitées, c'est-à-dire celles qui sont impliquées dans la détection.

Pour prendre en considération les différents niveaux d'interactions linguistiques, nous proposons dans ce chapitre conclusif, une architecture sur le modèle des systèmes multi-agents. Nous présentons cette architecture avant de discuter certaines limites inhérentes au modèle. Nous approfondissons enfin la question des rétroactions.

1 Un système multi-agents

Les systèmes multi-agents sont des systèmes constitués d'agents qui évoluent au sein d'un environnement dans le quel ils interagissent. Wooldridge [2002, p. 15] définit un agent comme « *a computer system that is situated in some environment, and that is capable of autonomous action in this environment in order to meet its design objectives*[10] ».

L'environnement est un espace commun à tous les agents dans lequel se trouve l'ensemble des données utiles aux agents et sur lesquelles ils agissent. Il est constitué, dans notre cas, du token en cours de traitement, de ses étiquettes morphosyntaxiques, de la mémoire des chunks et des piles d'attentes.

Les agents observent l'environnement et décident, en fonction des données qu'ils perçoivent, d'effectuer des actions ou non. Ils agissent de façon autonome pour atteindre les objectifs qui leur ont été fixés. Pour la vérification grammaticale, un agent sera par exemple affecté à l'étiquetage morphosyntaxique, un autre à l'unification, etc. L'architecture que nous proposons comporte 13 agents. Nous les illustrons dans la figure 1 p. 209, en nous inspirant du mode de représentation adopté par Lebarbé & Girault [2000].

Les agents se décomposent en trois modules :
- Le module de perception, représenté par un œil, observe l'environnement en permanence et en surveille les modifications. Un filtre de perception lui permet de ne voir que les données le concernant.
- Le module de délibération, représenté par un « K » pour *Knowledge* (Connaissances), dispose de ressources propres et essaye de les appliquer aux données perçues de l'environnement pour inférer une action répondant à ses objectifs. Cette action est transmise au module d'action.
- Le module d'action reçoit les actions du module de délibération et les applique à l'environnement.

Dans la figure 1 p. 209, nous distinguons deux environnement : l'un interne et l'autre externe. L'environnement externe est extérieur à l'analyse. Il est composé du flux de tokens saisis et des messages à l'utilisateur dans le cadre des rétroactions. Il ne peut pas à proprement parler être considéré comme un environnement dans la mesure où seuls deux agents, (a) : perception et (m) : rétroactions, y sont présents mais n'y interagissent pas. L'on peut considérer l'utilisateur comme un troisième agent de cet environnement externe, lui aussi doté de la capacité de perception, d'interprétation et d'action.

L'environnement interne est constitué de tous les éléments résultant de l'analyse ainsi que du token en cours de traitement transmis par l'agent (a) qui l'a perçu dans l'environnement externe. Il est le lieu d'interaction de l'ensemble des agents (b) à (l), chacun détenteur d'une compétence et d'un ensemble de connaissances pour le traitement linguistique.

L'agent (a) a pour objectif de percevoir chaque nouveau token arrivant dans l'environnement externe (qui vient d'être saisi par exemple) et de le déposer dans l'environnement interne. Le token sera alors perçu par l'ensemble des autres agents. Par exemple, l'agent (b), en l'absence d'étiquetage morphosyntaxique, pourra initier une tentative d'identification des catégorie/sous-

10. Un agent est un programme informatique situé dans un environnement et capable d'effectuer des actions de manière autonome au sein de cet environnement afin d'atteindre les objectifs pour lesquels il a été conçu. (traduit dans [Lebarbé, 2002]).

catégorie. L'agent (c), s'il ne dispose pas d'un chunk en cours de constitution, pourra déclarer l'ouverture d'un chunk.

L'agent (b) est chargé de l'étiquetage morphosyntaxique. Il perçoit l'environnement interne et déclenche un raisonnement si l'environnement interne contient un token dont l'étiquette morphosyntaxique doit être identifiée ou clarifiée. Son raisonnement s'appuie d'une part sur la pile d'attentes de catégories intra-chunk (inscrite dans l'environnement interne), d'autre part sur le chunk en cours. Il dispose pour cette démarche d'un lexique étiqueté (*cf.* § *Étiquetage morphosyntaxique* p. 178).

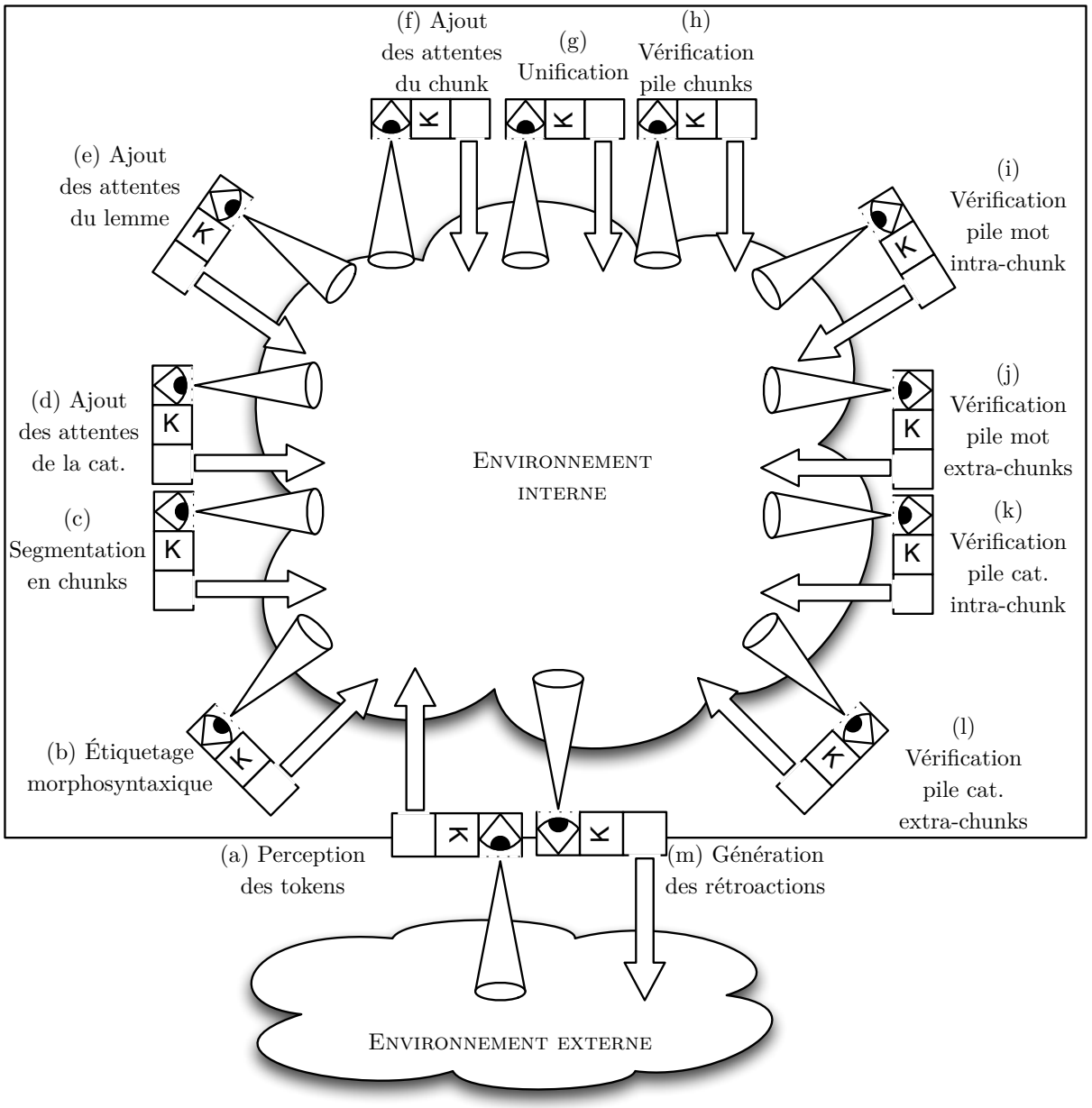


Figure 1 : Système multi-agents pour la vérification grammaticale

L'agent (c) tente de délimiter les chunks. Il perçoit le token et son étiquette, ainsi que l'état du dernier chunk et s'appuie à la fois sur des règles de segmentation et sur des attentes sur ses constituants pour modifier l'état du dernier chunk.

Les agents (d), (e) et (f) ont pour rôle d'ajouter des attentes aux piles en se fondant sur des bases de règles. L'agent (d) se fonde sur la catégorie du token pour générer des attentes d'ordre catégoriel, lexical ou syntagmatique dans les piles correspondantes. L'agent (e) s'appuie sur le lemme du token et génère de la même manière des attentes d'ordre catégoriel, lexical ou syntagmatique dans les piles correspondantes. L'agent (f) perçoit le chunk qui vient d'être fermé et tente d'appliquer une règle d'ajout d'attente dans une des piles à partir de ce chunk.

L'agent (g), observe la sous-catégorie du token et tente de l'unifier avec la sous-catégorie de la pile correspondante, qu'il perçoit également. Il est l'un des trois agents capable de détecter des incohérences. S'il parvient à effectuer l'unification, il ajoute le résultat à la pile. Dans le cas contraire, il laisse un message dans l'environnement interne. Celui-ci sera perçu par l'agent (m) qui génèrera une rétroaction.

Les agents (h) à (l) sont chargés de vérifier les différentes piles d'attentes. Chaque agent est responsable d'une pile. Il doit vérifier si les données qu'il perçoit de l'environnement valident une attente de la pile dont il s'occupe, et si la non validation est un signe d'incohérence, d'après des règles dont il dispose. Dans le cas d'une incohérence, il laisse un message dans l'environnement pour l'agent (m). L'agent (h) perçoit le chunk qui vient d'être fermé, vérifie s'il répond à une attente de la pile chunks et décide s'il y a une incohérence ou non. Les agents (i) et (j) perçoivent le lemme, vérifient s'il répond à une attente de la pile lemme intra-chunk pour l'un et extra-chunks pour l'autre, et décident s'il y a une incohérence ou non. Les agents (k) et (l) perçoivent la catégorie, vérifient si elle répond à une attente de la pile catégorie intra-chunk pour l'un et extra-chunks pour l'autre, et décident s'il y a une incohérence ou non.

Enfin l'agent (m) observe l'environnement interne pour percevoir un message d'incohérence laissé par un des agents (g) à (l). Il génère alors une rétroaction en fonction du message et des données qu'il a perçus et la transmet à l'environnement externe d'où elle sera rendue accessible à l'utilisateur.

D'autres agents pourraient venir se greffer à ce système multi-agents, comme un agent pour la vérification orthographique par exemple, qui vérifierait l'existence du token et laisserait un message dans l'environnement à destination de l'agent dédié à la génération de rétroactions pour l'utilisateur. Des agents pour prendre en charge la dimension phonologique peuvent également être envisagés, d'autant plus que l'homophonie touche la majorité des erreurs. Ces agents permettraient, par exemple, de proposer des corrections homophones des erreurs notamment dans le cadre des confusions de lemmes.

D'autres agents encore pourraient s'occuper des attentes sémantiques. Nous avons vu que les erreurs nécessitant un recours au sens pour être détectées par l'humain sont nombreuses et concernent notamment plus d'un tiers des erreurs d'accord (*cf.* § c) p. 170). Nous en avons plusieurs exemples dans les extraits (44) et (45) ci-dessous. Dans l'extrait (44), *des dizaines* et *des gammes* représentent un ensemble de plusieurs éléments et leurs compléments du nom devraient donc être au pluriel. Dans l'extrait (45), *vente* est pris au sens général, au singulier, mais nous inférons qu'il s'agit de la vente de plusieurs vêtements, et donc que *vêtement* devrait alors être au pluriel, comme *accessoirs*. De même, *chien* devrait logiquement être au pluriel,

puisque la *`vente de vêtement et accessoirs` n'est pas destinée à un unique chien.

(44) *`des dizaines de stand proposant des gammes de produit de beauté`

(45) *`un site internet de vente de vêtement et accessoirs pour chien`

Dans certains cas, ces erreurs de nombre pourraient être détectées grâce à la notion d'actants et à un dictionnaire les décrivant, en s'affranchissant du sens. Par exemple, dans l'extrait (44), il n'est syntaxiquement pas possible de déterminer si `stand` et `produit` doivent être au singulier ou au pluriel. En revanche, un dictionnaire d'actants indiquerait que `dizaines` peut avoir un complément du nom qui doit être au pluriel. Il en va de même pour `gamme`. Dans l'extrait (45), le recours aux actants ne permettraient pas de résoudre le problème du nombre de `vente` et de `chien` pour lesquels il est nécessaire de faire appel au sens.

Cette structure étant fondée sur un environnement interne d'accumulation de connaissances linguistiques et de mémoire du processus, d'autres compétences linguistiques peuvent être intégrées de manière à traiter différents aspects de la production écrite (orthographe, sémantique, stylistique,etc.). Des évolutions du système sont donc possibles, pour affiner la détection d'incohérences et la génération des rétroactions, et seraient facilitées par la modularité d'un système multi-agents.

2 Limitations du modèle

Le modèle, tel que nous l'avons présenté dans le chapitre 8, semble efficace pour détecter certaines erreurs, en particulier celles d'accord au sein d'un même chunk. Mais il fait également l'objet de quelques limitations que nous discutons dans cette section.

2.1 Complexité de la détection de certaines erreurs

a) Cas des accords

Si la vérification des accords intra-chunk ne pose *a priori* pas de problème grâce à l'unification des traits des constituants des chunks au fur et à mesure de leur lecture, certaines configurations peuvent « piéger » le modèle et lui faire manquer des erreurs. Le problème que nous soulevons ici est celui du rattachement d'un token au bon antécédent pour réaliser son accord.

Deux situations peuvent se présenter. La première est celle du rattachement d'un adjectif non pas au nom qui le précède directement, mais à un nom antérieur. Prenons les exemples suivants :

(46) *`[en laissant] [le bouton] [de la souris enfoncée]`

(47) *`[le langage] [des signes américains]`

Dans l'exemple (46), la segmentation en chunks mène à trois chunks, dont le chunk prépositionnel `[de la souris enfoncée]`. L'adjectif `enfoncée` étant inclus au chunk, le modèle réalise l'unification de ses traits `féminin singulier` avec ceux des constituants précédents, qui sont les mêmes, et ne conclut donc à aucune incohérence. Or, sémantiquement, l'adjectif devrait être rattaché à `bouton`. Il y a donc ici une erreur d'accord, indétectable par notre modèle car elle implique la dimension sémantique qu'il ne prend pas en compte pour le moment.

L'exemple (47) illustre une situation semblable. Par unification au sein du chunk préposition-

nel *des signes américains*, *américains* n'est pas détecté comme une erreur d'accord. Il devrait pourtant être accordé avec *le langage*. Ici encore intervient la dimension sémantique qui rend impossible la détection de l'erreur. Une première solution dans ce cas pourrait être la reconnaissance par le modèle, et plus précisément par l'étiqueteur morphosyntaxique, d'un token unique constitué de *langage des signes*, c'est-à-dire la collocation. La présence d'un adjectif masculin pluriel postposé à la collocation masculin singulier ferait échouer l'unification de traits et serait détecté comme incohérence. Une seconde solution serait d'intégrer la connaissance que dans un contexte « Nom de Nom Adjectif », l'adjectif peut qualifier l'un des deux noms.

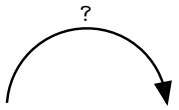
La première solution fait appel à plus de ressources externes mais est potentiellement plus précise que la seconde solution, qui nécessite moins de ressources externes mais engendre un risque de silence.

Le problème de la recherche de l'antécédent se pose également dans le cas de proposition participiale. Il est impossible, sans faire usage de la sémantique, de rattacher *situés* à *l'institut* dans l'exemple suivant :

(48) **Les fondateurs de l'institut de recherche sur les chimpanzés, situés sur le campus [...]*

Le même problème se pose pour rattacher un pronom relatif sujet à son antécédent, et lui assigner ainsi les bons traits, qui serviront à vérifier l'accord du verbe de la relative.

Une solution serait, en cas de doute sur un rattachement, de demander confirmation à l'utilisateur que tel mot dépend bien de tel autre, sous une forme graphique facilement interprétable comme ci-dessous :



Les fondateurs de l'intitut de recherche sur les **chimpanzés,** *situés* sur le campus

b) Cas des attentes inter-chunks

La mobilité des chunks au sein de la phrase, c'est-à-dire le fait que leur place ne soit pas figée au sein de la phrase, constitue une autre difficulté pour le modèle. Il faut généralement attendre la fin de la phrase pour être sûr qu'une attente de chunk, ou de catégorie/mot dans un autre chunk, n'est pas comblée, ce qui ne permet pas de détection au moment, ou proche du moment où l'erreur a été commise.

L'incohérence peut également se déplacer au fil de l'analyse de la phrase, c'est-à-dire ne pas être détectée au niveau de l'erreur, mais faire l'objet d'une détection plus loin, concernant un élément pas forcément erroné. Par exemple, dans la longue phrase (49), *à provoqué* contient une double erreur : d'une part l'auxiliaire et la préposition ont été confondus, d'autre part l'auxiliaire aurait dû être au pluriel puisque son sujet est *Les effluves*. Ici le modèle va détecter que le participe passé *provoqué* ne comble pas l'attente d'infinitif créée par la préposition *à*, et va suggérer au scripteur de remplacer *provoqué* par *provoquer*. Le premier verbe que le modèle trouvera ensuite est *a rapporté*, dont il pensera qu'il est le verbe attendu par le sujet *Les effluves* et dont il dira qu'il ne comble pas l'attente d'un verbe à la 3e personne du pluriel. Il signalera donc un problème d'accord, alors qu'il n'y en a pas. En revanche, la substitution de *a*

par `à`, seule erreur de la phrase, n'est pas détectée comme telle.

(49) *Les effluves d'une décoction de piments, que le cuisinier d'un restaurant de Londres préparait, à provoqué une alerte terroriste en plein centre de la capitale britanique, a rapporté mercredi le quotidien Times.*

2.2 Des ressources complexes à élaborer

a) Les attentes du lemme

Les bases de règles d'attentes sont complexes à construire, et en particulier celle concernant les attentes créées par les lemmes. Cette base contient les informations sur les actants des mots et sur leur réalisation syntaxique.

Concernant les verbes, nous pouvons envisager d'adapter le dictionnaire DICOVALENCE [van den Eynde & Mertens, 2010] ou sa version simplifiée Dicovalence-Easy [Gardent & Vandenberghes, 2009], tous deux assez conséquents et permettant de déterminer les attentes de milliers de verbes.

En revanche, pour les mots des autres catégories, il n'y a pas de ressource aussi complète. La plupart des mots autres que les verbes ne peuvent alors pas créer d'autres attentes que celles qu'ils créent par leur catégorie grammaticale ou leurs traits morphosyntaxiques. Il est bien sûr envisageable de compléter les ressources produites par des projets comme le projet DiCo [Jousse & Polguère, 2005], dont les acteurs ont développé une version simplifiée et informatisée du DEC de Mel'čuk [1984, 1988b, 1992, 1999], mais il s'agit d'un travail très coûteux qu'il faudrait confier à des lexicographes.

Une autre piste serait d'inférer les actants des mots à partir d'un corpus, comme c'est le cas pour extraire les collocations par exemple (voir [Nerima *et al.*, 2006] pour différentes méthodes d'extraction sur corpus). Une validation des actants extraits serait alors nécessaire et elle aussi coûteuse, à moins de l'envisager au travers des rétroactions de l'outil de vérification grammaticale. En fonction des actants extraits sur corpus, des attentes seraient créées au niveau du lemme et lorsqu'elles conduiraient à la détection d'une incohérence, la décision de l'utilisateur (corriger ou laisser en l'état) servirait à la validation ou l'invalidation de l'attente, et donc de l'actant.

Il est à noter que la prise en compte des attentes du lemme induit une autre difficulté qui est de disposer de la forme lemmatisée du token traité. Il est nécessaire pour cela d'avoir recours à un lexique contenant à la fois les formes graphiques (toutes les formes fléchies des mots) et les lemmes correspondants. Il pourrait s'agir du même lexique que celui utilisé pour l'étiquetage morphosyntaxique (par exemple le lexique de dicollecte [Ronez, 2014]), pour ne pas multiplier les ressources. Le choix de ce lexique est alors primordial car les informations qui en sont extraites conditionnent la plupart des traitements menant à la détection d'incohérences.

b) Les bases de règles d'attentes

Les bases de règles d'attentes de notre modèle sont au nombre de deux, si nous excluons les dictionnaires d'actants que nous venons d'aborder : les attentes créées par la catégorie et les attentes créées par le chunk.

La base de règles la plus délicate à élaborer, selon nous, est celle des attentes créées par la

catégorie. Elle doit tenir compte des différents contextes syntaxiques possibles dans lesquels une certaine catégorie attend un certain élément, qui peut être une catégorie un lemme ou un chunk. Le risque est alors de retomber dans le problème des règles que nous dénoncions au sujet des vérificateurs grammaticaux libres se fondant sur le principe du *pattern-matching* (*cf.* § a) p. 43). Trop peu de règles génère du silence dans la détection, et des règles trop nombreuses peuvent entraîner des redondances ou mener à des règles contradictoires.

Ce risque de redondance et de contradiction est d'autant plus grand qu'un même token génère des attentes à la fois par sa forme lemmatisée et par sa catégorie, et que ces attentes peuvent être pour des lemmes, pour des catégories et pour des chunks. Par exemple le lemme *tendre* a des attentes multiples : attente d'un chunk nominal ou d'un chunk prépositionnel, attente d'une catégorie préposition ou encore attente des lemmes *à* ou *vers*. Ces trois types d'attentes sont redondants, mais faut-il privilégier l'attente la plus générique (chunk) ou la plus spécifique (*à* et *vers*) ? En privilégiant la plus spécifique on évite le risque d'accepter une construction avec une mauvaise préposition. Mais il faut alors préciser que l'attente est facultative car *tendre* ne se construit pas obligatoirement avec un syntagme prépositionnel. La combinatoire de toutes les attentes créées par un même token et la gestion des priorités dans ces attentes constitue donc un enjeu pour le TAL.

3 La question des rétroactions explicites contextuelles

Nous avons dressé, dans le chapitre 3 (*cf.* § 3.1.3 p. 45), un état des lieux des rétroactions proposées dans différents outils de vérification grammaticale, commerciaux et libres. Nous avons constaté que ces rétroactions sont souvent avares en explications sur l'erreur détectée, et également prescriptives. L'utilisateur dispose de peu d'indices pour décider s'il y a effectivement une erreur, s'il doit la corriger et comment la corriger. À moins d'avoir une pratique et des connaissances avancées de l'écrit, et une compréhension du fonctionnement de son logiciel, il n'a d'autres choix que d'avoir une confiance aveugle en ce que lui dit le système.

Nous nous interrogeons alors sur le contenu que doivent avoir les rétroactions pour être utiles à l'utilisateur et sur la façon de présenter ce contenu pour être intelligible pour tous les publics.

3.1 Quel contenu ?

a) Le contexte de l'incohérence

Notre modèle offre la possibilité d'indiquer le token en cours de traitement et l'attente non validée, elle-même contenant l'attendant et l'attendu (*cf.* § *Rétroactions* p. 201). C'est la combinaison du token et de l'attente en attente de validation qui déclenche la rétroaction. Celle-ci est alors explicite car elle indique tous les éléments du contexte de l'incohérence détectée et leur implication.

Dans de nombreuses situations cependant, la génération d'une rétroaction qui soit explicite et qui replace l'incohérence détectée dans son contexte est difficile à mettre en œuvre. Dans l'extrait (50), il y a une confusion entre *peu* et *peut*. *peut* devient le verbe de *une mamie*, et *une mamie* le sujet de *peut*. Le verbe attend par ailleurs un complément sous forme de pronom ou de verbe infinitif. Quand arrive *emprunte*, il n'y a plus de sujet attendant un verbe. Un sujet pouvant être postposé, il faut continuer l'analyse de la phrase. Le dernier syntagme nominal pourrait

être considéré comme un sujet postposé. Dans ce cas il manquerait un complément d'objet. À l'inverse, si le dernier syntagme nominal est considéré comme un objet, alors il manque un sujet. Il faut attendre la fin de la phrase ici pour signaler une incohérence avec *emprunte*, qui soit n'a pas de sujet, soit pas de complément. C'est à ce moment-là aussi que sera signalée l'absence de complément du verbe *peut*.

La phrase conduit ainsi à deux rétroactions suite à une seule erreur car elle a induit deux incohérences. La rétroaction concernant *emprunte* qui n'a pas de sujet ou pas de complément est indésirable. Quant à la rétroaction sur l'absence de complément de *peut*, elle présente l'intérêt d'attirer l'attention sur le mot erroné, mais n'est pas en mesure d'indiquer qu'une confusion a sans doute été faite entre deux mots grammaticaux, et encore moins de proposer une correction.

(50) *une mamie peut coutumière des aéroports emprunte le tapis-bagages* .

De manière plus générale, les informations stockées dans les piles et la mémoire des chunks permettent normalement de contextualiser les incohérences relevées. Le problème se situe davantage au niveau des incohérences elles-mêmes, qui peuvent se situer sur des éléments sans rapport direct avec l'erreur effective, comme dans l'exemple précédent.

Dans un cas comme celui-ci, mais aussi de manière plus générale, il faudrait que la décision du scripteur de refuser la rétroaction soit prise en compte par le système pour revenir sur son analyse et en proposer une nouvelle.

b) Des propositions de correction

Proposer un moyen de résoudre l'incohérence détectée est difficilement réalisable dans les cas que nous venons de citer où l'incohérence est sans lien direct avec la véritable erreur. En revanche, dans les cas d'erreurs d'accord détectées par l'échec de l'unification, par exemple, il est envisageable de recourir à un lexique étiqueté contenant les lemmes de chaque forme fléchie. Il serait possible de proposer dans la rétroaction la forme issue du lexique qui correspond au lemme et à la bonne valeur de trait. En effet, le système connaît les deux informations nécessaires, à savoir le trait du token en cours de traitement qui a fait échouer l'unification, et la valeur il aurait dû avoir.

Dans les autres cas d'erreurs, cette procédure ne fonctionne que si un mot particulier est attendu, ou si le token déclencheur de la rétroaction est le bon lemme mais sous une forme incorrecte, comme dans l'exemple que nous donnions dans le précédent chapitre (*cf.* § *Rétroactions* p. 201) : **elle a atterrit*. Ici, une forme conjuguée est utilisée à la place du participe passé, mais avec le même lemme. Il est possible alors, avec une attente de participe passé non validée, de proposer en correction le participe passé correspondant au lemme de *atterit*. Indiquer simplement à l'utilisateur qu'il fallait un participe passé après l'auxiliaire n'est peut-être pas suffisant. Rien n'indique que l'utilisateur sache comment orthographier le participe passé, ni qu'il comprenne la rétroaction s'il pense que le participe passé de *atterrir* prend un *t*.

Une autre piste de proposition de correction se trouve dans la phonologie. Nous avons fait ressortir de notre corpus 78% d'erreurs homophones de la forme correcte (*cf.* § *L'homophonie dans les erreurs* p. 128). Il est alors légitime de penser que l'homophonie pourrait être un critère dans la proposition de correction. Cela nécessiterait néanmoins une annotation des tokens au niveau phonémique, ce qui implique alors de disposer d'un lexique contenant une transcription phonémique des formes fléchies, comme le propose le dictionnaire DELAP [Laporte, 1990].

3.2 Quelle représentation ?

Après avoir abordé le « quoi dire » dans les rétroactions, se pose la question de « comment le dire ». Si nous restons dans les rétroactions classiques textuelles, il nous semble pertinent de tenir compte du profil de l'utilisateur.

L'outil idéal devrait prendre en compte notamment la langue maternelle du scripteur. En effet, l'analyse des erreurs de notre corpus a fait ressortir la particularité des scripteurs non francophones natifs, qui commettent des erreurs plus nombreuses, notamment sur les accords des déterminants et les substitutions de lemmes (*cf.* section *Résumé des principaux résultats* p. 136). Quand un un francophone natif aura par exemple simplement besoin qu'une erreur d'accord en genre soit mise en évidence pour la corriger, sans nécessiter davantage d'explications, l'apprenant en FLE aura, quant à lui, besoin d'une explication sur le genre des mots impliqués dans l'erreur.

Ainsi les rétroactions pourraient varier en fonction de la langue du scripteur. Il nous semble qu'elles pourraient varier également en fonction du niveau de connaissances linguistiques du scripteur. Nous nous posons ici la question des formulations à employer pour décrire et expliquer les incohérences détectées. Développer un outil destiné à tout type d'utilisateur pose la question de leur connaissances grammaticales et de leur capacité à comprendre des termes comme « complément d'objet », « participe passé » ou encore « conjonction ». De plus la formulation même des rétroactions doit être la plus explicite possible.

Une solution se trouve peut-être dans des rétroactions à plusieurs niveaux de détail, du type de celles que propose le logiciel Antidote (*cf.* § *Exemples de logiciels commerciaux* p. 46). Un premier niveau indique le mot erroné et un mot-clé sur le type d'erreur, un second niveau explicite et contextualise l'erreur en quelques mots, et enfin un troisième niveau propose la règle de grammaire convoquée. Le premier niveau est généralement suffisant pour corriger l'erreur, mais si ce n'est pas le cas, ou en cas de doute, l'utilisateur peut accéder au second puis au troisième niveau. Pour mettre en œuvre des rétroactions à plusieurs niveaux, le modèle/outil de vérification grammaticale doit être capable de faire le lien entre une incohérence détectée et une règle de grammaire.

Les rétroactions pourraient également être envisagées de manière graphique, dans une approche inductive, pour être comprises plus facilement [Wade-Stein & Kintsch, 2004], en faisant l'économie d'un texte. Par exemple, l'erreur d'accord pourrait être présentée comme en figure 2, par les deux éléments devant s'accorder, reliés par une flèche indiquant quel trait (genre ou nombre) appliquer à quel mot.

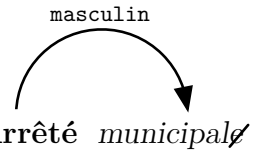


FIGURE 2 : Exemple de rétroaction graphique

Les rétroactions pourraient également s'inspirer de la méthode *Silent Way* de Gattegno [2010] pour s'abstraire du métalangage grammatical afin d'expliquer les incohérences grammaticales détectées. Il s'agit d'une méthode d'apprentissage inductif qui se fonde sur des codes couleur et

des réglettes de couleur pour appréhender la prononciation, le vocabulaire et la grammaire d'une langue étrangère. Cette méthode a été adaptée dans l'enseignement supérieur pour l'apprentissage de la grammaire en langue maternelle [Boch *et al.*, 2012].

3.3 Prise en compte de la décision de l'utilisateur

Lorsqu'une rétroaction est générée, c'est-à-dire lorsqu'une incohérence a été détectée, l'utilisateur a le choix entre modifier son texte ou ignorer la rétroaction. Dans le premier cas, l'analyse de la phrase reprend à l'endroit de la modification, avec un dépilage de toutes les attentes jusqu'à celles générées par le token juste avant le texte modifié.

Si l'utilisateur choisit d'ignorer une rétroaction, probablement parce qu'il juge qu'il s'agit d'une fausse alarme, il faudrait que le système en tienne compte. En effet, s'il a détecté une incohérence mais qu'elle ne se situe pas au niveau du token qu'il a signalé, c'est qu'il y a certainement une autre incohérence à côté de laquelle il est passé. L'exemple (50), dont nous décrivons le traitement p. 215, illustre une situation de ce type, avec une erreur sur *peut* qui ne provoque une incohérence qu'au niveau de *emprunte*.

(50) *une mamie peut coutumière des aéroports emprunte le tapis-bagages* .

Dans un cas comme celui-ci, il faudrait que le système reprenne l'analyse de la phrase, en se focalisant sur une possible confusion de lemmes. Les confusions de mots grammaticaux notamment sont fréquentes (50% des erreurs de substitution de lemmes dans notre corpus). Une liste des confusions les plus fréquentes, comme *est*/*et*, *à*/*a*, *sait*/*c'est*, etc., pourrait venir compléter les ressources utilisées par l'outil et lui permettre de réaliser une analyse alternative en considérant non plus le mot saisi mais sa correction probable. Cette stratégie constitue un complément à celle proposée p. 210. Dans l'exemple (50), la nouvelle analyse considèrerait que *peut* pourrait se substituer à *peu*, ce qui serait cohérent au niveau des chunks et des attentes. Une incohérence serait donc signalée sur *peut*. Cette stratégie (agentifiée) pourrait compléter une stratégie plus générique par un agent phonologique tel que préconisé p. 210.

Si l'utilisateur ignore la rétroaction, il se peut également que l'incohérence détectée trouve son origine dans un mot dont les traits de catégorisation et/ou sous-catégorisation seraient inconnus du système, c'est-à-dire inconnus du lexique étiqueté utilisé (qui ne peut être exhaustif), sans toutefois être mal orthographié. Dans ce cas, le système étiquettera le mot avec les traits qu'il connaît mais qui correspondent à une autre acception du token. Ces traits créeront, valideront ou invalideront alors des attentes que n'auraient pas créées, validées ou invalidées les véritables traits du token, d'où l'incohérence détectée. Le système pourrait alors inférer les traits du token d'après les attentes qu'il doit valider, et demander à l'utilisateur de confirmer qu'il s'agit des bons traits. La nouvelle acception du token pourrait alors venir compléter le lexique étiqueté.

4 Conclusion

Dans la présente thèse, nous sommes partie de la distinction qui peut être faite entre les points de vue linguistique et informatique sur « l'erreur de grammaire » pour délimiter les contours de notre recherche. À la lumière de cette définition, nous avons passé en revue les vérificateurs libres existants pour constater un réel besoin d'amélioration. Les systèmes de vérification linguistique étant destinés à une analyse de textes numériques, nous avons remarqué un manque dans la

littérature quant à l'analyse des erreurs spécifiquement tapuscrites.

Afin d'appuyer nos réflexions sur des données authentiques, nous avons constitué un corpus d'erreurs tapuscrites. L'annotation de celui-ci a demandé un travail préalable sur notre typologie qui avait vocation à être descriptive. Ceci nous a permis de nous positionner par rapports aux typologies existantes. Enfin, l'annotation du corpus nous a permis de mettre à jour nos descripteurs.

Après avoir constaté certaines propriétés génériques de notre corpus (influence de la situation de scription, du scripteur, du texte), nous nous sommes intéressée aux processus cognitifs de rédaction et de révision chez l'humain. Le croisement de ces deux dimensions (corpus d'erreurs et processus cognitifs) nous a permis de proposer un modèle fondé sur les mécanismes suivants :

- Les attentes : élément central des processus cognitifs de révision, les attentes correspondent également à un principe fondateur des grammaires de dépendances et occupent une place centrale dans notre modèle. Elles permettent de détecter des erreurs de lexique, de syntaxe et de mode verbal ;
- La segmentation en chunks : le chunk est une unité de calcul au sein des processus cognitifs. De plus nos données montrent que les deux tiers des erreurs de grammaire produites touchaient un mot dont le référent est dans le même chunk ;
- L'unification : il s'agit d'un principe linguistique, ainsi qu'une partie du processus cognitif de rédaction qui permet au scripteur de réaliser les accords. Ces erreurs constituent à elles seules la moitié des erreurs de grammaire.

Le modèle présenté nécessite à présent d'être mis en œuvre et testé sur des données. Le corpus que nous avons constitué fournirait une première étape à ces tests. L'intégralité des erreurs qu'il contient a été annoté, avec pour chaque erreur une indication de son type et de la correction suggérée. Il est donc possible d'automatiser toute l'analyse du corpus par un prototype, ce qui permettrait d'en mesurer la précision et le rappel, y compris type d'erreurs par type d'erreurs.

Toutefois notre corpus ne permettra pas d'analyser les contextes qui suscitent le plus de faux positifs. Il faudra alors s'appuyer sur nos résultats méthodologiques pour compléter notre corpus. Pour approfondir l'analyse, des informations supplémentaires pourraient être annotées, comme les caractéristiques morphosyntaxiques de tous les mots, et plus seulement des erreurs. La prise en compte des informations contextuelles permettrait de décrire les situations où la détection échoue. Il peut en effet y avoir de fausses incohérences détectées dans certaines circonstances, ou au contraire de vraies erreurs passées sous silence, voire une combinaison des deux lorsqu'une incohérence est détectée de manière décalée sur un mot qui se situe plus loin que l'erreur (*cf.* § *Le contexte de l'incohérence* p. 214).

L'amélioration de la détection passe ainsi par un corpus annoté plus en détail, mais aussi par un corpus éventuellement plus vaste. En effet, notre corpus ne couvre qu'une petite variété de types d'écrits, et nous avons vu que les erreurs commises sont sensibles à cette variété (*cf.* section *Résumé des principaux résultats* p. 136). Élargir notamment les situations de scription permettra peut-être de faire émerger des situations où la vérifications grammaticale orientée vers le français standard ne correspond pas aux attentes des utilisateurs. En effet, on peut se demander s'il est pertinent de corriger en français standard des écrits produits dans un style oralisé, par exemple dans le cadre de mails informels (*cf.* section *Résumé des principaux résultats* p. 136).

Dès lors que nous nous intéressons à l'adaptation de la variété de langue au type d'écrits, se pose celle de la détection de la situation de scription, des besoins de l'utilisateur. Nous avons

évoqué la pertinence de tenir compte du profil du scripteur, de sa langue maternelle et des connaissances linguistiques en particulier (*cf.* section *7.3* p. 171 et § *8.3.2* p. 216). Un enjeu majeur consiste à prendre en compte les actions de l'utilisateur vis-à-vis des rétroactions et la façon dont le modèle peut réagir à ces actions. Nous avons avancé quelques pistes dans la dernière section (*cf.* § *Prise en compte de la décision de l'utilisateur* p. 217), mais ces considérations sont susceptibles d'amener à des modifications du modèle en profondeur. Celui-ci n'est en effet pas conçu, à l'heure actuelle, pour être en mesure d'analyser une seconde fois un énoncé pour lequel l'utilisateur aurait ignoré une rétroaction.

La réflexion sur les questions posées par les rétroactions et les interactions entre le système et l'utilisateur constitue un large axe de recherche qui viendra compléter et faire évoluer les premiers jalons posés en vue d'une implantation. En effet, l'évaluation d'un système de vérification grammaticale, ne peut — à l'heure actuelle — s'arrêter à une simple analyse des sorties des systèmes, qui ne sont pas en mesure d'être utilisés en situation réelle sans opérateur. Or, nos réflexions montrent que le système peut répercuter les rétroactions de l'utilisateur dans sa propre analyse (*cf.* § *Prise en compte de la décision de l'utilisateur* p. 217) et ainsi améliorer la fiabilité du système.

Acronymes

ALAO	Apprentissage des Langues Assisté par Ordinateur
CES	*Corpus Encoding Standard*
COD	Complément d'Objet Direct
COI	Complément d'Objet Indirect
CSS	*Cascading Style Sheet*
DTD	*Document Type Definition*
EIAL	Environnements Informatiques pour l'Apprentissage des Langues
FLE	Français Langue Étrangère
NLP	*Natural Language Processing*
OCR	*Optical Character Recognition*
ST	Structure de Traits
TAL	Traitement Automatique des Langues
TEI	*Text Encoding Initiative*
XML	*eXtensive Mark-up Language*
XSL	*eXtensible Stylesheet Language*

Bibliographie

[Abeillé, 1993] Anne Abeillé (1993). *Les nouvelles syntaxes - grammaires d'unification et analyse du français*. Armand Colin.

[Abeillé, 1998] Anne Abeillé (1998). Grammaires génératives et grammaires d'unification. *Langages*, 32(129):24–36.

[Abeillé, 2007] Anne Abeillé (2007). *Les grammaires d'unification*. Hermès, Lavoisier. ISBN : 978-2-7462-1251-0.

[Abney, 1991] Steven P. Abney (1991). Principle-based parsing : Computation and psycholinguistics. *In* R.C. Berwick, S. Abney & C. Tenny, éditeurs, *Principle-Based Parsing : Computation and Psycholinguistics*, chapitre Parsing by chunks, pages 257–278. Kluwer Academic Publishers.

[Académie Française, 1932] Académie Française (1932). *Grammaire de l'académie française*. Firmin-Didot.

[Alamargot & Chanquoy, 2001] Denis Alamargot & Lucile Chanquoy (2001). *Through the models of writing*. Kluwer Academic Publishers.

[Allal *et al.*, 2004] Linda Allal, Lucile Chanquoy & Pierre Largy, éditeurs (2004). *Revision : cognitive and instructional processes*. Kluwer Academic Publishers. ISBN : 1-4020-7729-7.

[Antoniadis *et al.*, 2010] Georges Antoniadis, Claude Ponton & Virginie Zampa (2010). Exxelant et mirto : Deux exemples d'environnement d'alao intégrant des outils tal. *Multilinguisme et traitement des langues naturelles*, pages 151–165. ISBN : 978-2-7605-2569-6.

[Arrivé *et al.*, 1986] Michel Arrivé, Françoise Gadet & Michel Galmiche (1986). *La grammaire d'aujourd'hui*. Flammarion. ISBN : 2-08-112003-8.

[Astolfi, 1997] Jean-Pierre Astolfi (1997). *L'erreur, un outil pour enseigner*. esf. ISBN : 2-7101-1203-5, ISSN : 1275-0212.

[Atkins *et al.*, 1992] Beryl T. Sue Atkins, Jeremy Clear & Nicholas Ostler (1992). Corpus design criteria. *Literary and Linguist Computing*, 7(1):1–16.

[Atkins & Rundell, 2008] Beryl T. Sue Atkins & Michael Rundell (2008). *The oxford guide to practical lexicography*. Oxford University Press.

[Babu, 2007] Jean-Philippe Babu (2007). L'influence de la tradition grammaticale gréco-latine sur la grammaire du thaï. *Journal of Humanities Naresuan University*, pages 21–41.

[Baddeley, 1986] Alan Baddeley (1986). *Working memory.* Oxford University Press.

[Baddeley, 1992] Alan Baddeley (1992). *La mémoire humaine : théorie et pratique.* Presses Universitaires de Grenoble, édition française de [**?**] (traduction : Solange Hollard). ISBN : 2-7061-0471-6.

[Baddeley, 2000] Alan Baddeley (2000). The episodic buffer : A new component of working memory. *Trends in Cognitive Sciences*, 4(11):417–423.

[Besse & Porquier, 1991] Henri Besse & Rémy Porquier (1991). *Grammaire et didactique des langues.* Langues et apprentissage des langues. Hatier/Didier. ISBN : 9782278069330.

[Biais, 2005] Maxime Biais (2005). Grac.
url : `http://grac.sourceforge.net/grac_architecture.pdf`.

[Biber *et al.*, 1998] Douglas Biber, Susan Conrad & Randi Reppen (1998). *Corpus linguistics : Investigating language structure and use.* Cambridge Approaches to Linguistics. Cambridge University Press. ISBN : 978-0521499576.

[Billières, 1988] Michel Billières (1988). Crible phonique, crible psychologique et intégration phonétique en langue seconde. *Travaux de didactique du Français langue étrangère*, 19:5–30. ISSN : 0765-1643.

[Boch *et al.*, 2012] Françoise Boch, Laurence Buson & Carole Blondel (2012). Orthographe & grammaire à l'université. quels besoins ? quelles démarches pédagogiques ? *Scripta*, 16(30):31–51.
url : `http://periodicos.pucminas.br/index.php/scripta/article/view/4238`.

[Bock & Cuting, 1992] Kathryn Bock & J. Cooper Cuting (1992). Regulating mental energy : performance units in language production. *Journal of memory and language*, 31:99–127.

[Boissière *et al.*, 2007] Philippe Boissière, Jean-Leon Bouraoui, Frédéric Vella, Aurélie Lagarrigue, Mustapha Mojahid, Nadine Vigouroux & Jean-Luc Nespoulous (2007). Méthodologie d'annotation des erreurs en production écrite. principes et résultats préliminaires. Actes de *TALN (Traitement Automatique des Langues Naturelles)* (Toulouse, 5-8 juin).
url : `http://www.irit.fr/~Philippe.Boissiere/VITIPI/PUB/WNC_07.pdf`.

[Boudreau & Kittredge, 2005] Sylvie Boudreau & Richard Kittredge (2005). Résolution des anaphores et détermintation des chaînes de coréférences. modèles et algorithmes pour la résolution d'annaphores. *Traitement automatique des langues*, 46(1/2005):41–69.

[Bouillon, 1998] Pierrette Bouillon (1998). *Traitement automatique des langues naturelles.* Duculot.

[Bouraoui *et al.*, 2009] Jean-Leon Bouraoui, Philippe Boissière, Mustapha Mojahid, Nadine Vigouroux, Aurélie Lagarrigue, Frédéric Vella & Jean-Luc Nespoulous (2009). Problématique d'analyse et de modélisation des erreurs en production écrite. approche interdisciplinaire. Actes de *TALN (Traitement Automatique des Langues Naturelles)* (Senlis, France, 24-26 juin).
url : `http://www.atala.org/taln_archives/TALN/TALN-2009/taln-2009-court-029.pdf`.

[Braffort, 1981] Paul Braffort (1981). Formalismes pour l'analyse et la synthèse de textes littéraires. *In* OuLiPo, éditeur, *Atlas de Littérature Potentielle*, pages 108–137. Gallimard. ISBN : 978-2070325009.

[Brill, 1997] Eric Brill (1997). Unsupervised learning of disambiguation rules for part of speech tagging. *Natural Language Processing Using Very Large Corpora*, 11:27–42.

[Brissaud *et al.*, 2006] Catherine Brissaud, Jean-Pierre Chevrot & Pascale François (2006). Les formes verbales homophones en /e/ entre 8 et 15 ans : contraintes et conflits dans la construction des savoirs sur une difficulté orthographique majeure du français. *Langue française*, 141:74–93.
url : `http://www.cairn.info/revue-langue-francaise-2006-3-page-74.htm`.

[Brunelle, 2004] Éric Brunelle (2004). Antidote : Correcteur, dictionnaire et plus. *BULAG*, Correction automatique : bilan et perspectives(29):25–32. ISBN : 2-84867-080-0.

[Bureau, 1985] Conrad Bureau (1985). *Le français écrit au secondaire : une enquête et ses implications pédagogiques*. Conseil de la langue française. ISBN : 9782551090327.

[Burnard, 1991] Lou Burnard (1991). An introduction to the text encoding initiative. *In* Daniel I. Greenstein, éditeur, *SGML : TEI Introduction*, pages 81–91. Max-Planck-Institut für Geschichte.
url : `http://xml.coverpages.org/edw26.html`.

[Burney, 1970] Pierre Burney (1970). *L'orthographe*, volume 685 de *Que sais-je ?* Presses universitaires de France.

[Butterfield *et al.*, 1996] Earl C. Butterfield, Douglas J. Hacker & Luann R. Albertson (1996). Environmental, cognitive, and metacognitive influences on text revision : Assessing the evidence. *Educational Psychological Review*, 8:239–297.

[Catach, 1991] Nina Catach (1991). *L'orthographe en débat. dossiers pour un changement*. Nathan.

[Catach *et al.*, 1980] Nina Catach, Daniel Duprez & Michel Legris (1980). *L'enseignement de l'orthographe : l'alphabet phonétique international, la typologie des fautes, la typologie des exercices*. Dossiers didactiques. Nathan. ISBN : 978-2-09-120900-5.

[Chanquoy & Alamargot, 2002] Lucile Chanquoy & Denis Alamargot (2002). Mémoire de travail et rédaction de textes : évolution des modèles et bilan des premiers travaux. *L'année psychologique*, 102(2):363–398.
url : `http://www.persee.fr/web/revues/home/prescript/article/psy_0003-5033_2002_num_102_2_29596`.

[Chanter, 1975] Dennis O. Chanter (1975). Modifications of the angular transformation. *Journal of the Royal Statistical Society. Series B (Applied Statistics)*, 24(3):354–359.
url : `http://www.jstor.org/stable/2347101`.

[Chiss, 2009] Jean-Louis Chiss (2009). Sciences du langage et didactique des langues. *Synergies Roumanie*, 4:127–137.

[Chomsky, 1965] Noam Chomsky (1965). *Aspects of the theory of syntax*. MIT Press. ISBN : 0-262-53007-4.

[Chomsky, 1969] Noam Chomsky (1969). *Le langage et la pensée*. Payot.

[Chomsky, 1971] Noam Chomsky (1971). *Aspects de la théorie syntaxique*. Éditions du Seuil, édition française de [Chomsky, 1965] (traduction : Jean-Claude Milner).

[Chomsky, 1979] Noam Chomsky (1979). *Structures syntaxiques*. Éditions Seuil.

[Clear, 1992] Jeremy Clear (1992). Corpus sampling. *In* Gerhard Leitner, éditeur, *New directions in English language corpora : methodology, results, software developments*, pages 21–31. Mouton de Gruyter. ISBN : 3-11-013201-X.

[Clément *et al.*, 2009] Lionel Clément, Kim Gerdes & Renaud Marlet (2009). Grammaire d'erreur - correction grammaticale avec analyse profonde et proposition de correction minimale. Actes de *TALN (Traitement Automatique des Langues Naturelles)* (Senlis, France, 24-26 juin 2009).
url : `https://hal.archives-ouvertes.fr/file/index/docid/396229/filename/susabda-taln2009.pdf`.

[Cordary, 2010] Noëlle Cordary (2010). L'orthographe du participe passé : les entretiens métagraphiques pour évaluer et comprendre les difficultés des élèves en classe de seconde. *Synergies France*, 6:77–84.

[Corder, 1980] Stephen Pit Corder (1980). Que signifient les erreurs des apprenants ? *Langages*, 57:9–15.
url : `http://www.persee.fr/web/revues/home/prescript/article/lgge_0458-726x_1980_num_14_57_1833`.

[Cori & David, 2008] Marcel Cori & Sophie David (2008). Les corpus fondent-ils une nouvelle linguistique ? *Langages*, 3(171):111–129. ISBN : 978-2-200-92467-6.

[COVAREC, 1994] COVAREC (1994). Corpus de variations orthographiques. Laboratoire LIDILEM, Université Stendhal Grenoble 3.

[Cuq, 2010] Jean-Pierre Cuq (2010). Sciences du langage et didactique des langues. *Synergies Brésil*, spécial 1:85–88.

[Cuq & Gruca, 2005] Jean-Pierre Cuq & Isabelle Gruca (2005). *Cours de didactique du français langue étrangère et seconde*. Presses Universitaires de Grenoble. ISBN : 2-7061-1301-4.

[Dabène, 1987] Michel Dabène (1987). *L'adulte et l'écriture. contribution à une didactique de l'écrit en langue maternelle*. De Boeck-Wesmael. BU-sdl : EL 86 A. ISBN : 2-8041-0996-8.

[de Saussure, 1916] Ferdinand de Saussure (1916). *Cours de linguistique générale*. Payot.

[Debyser, 1970] Francis Debyser (1970). La linguistique contrastive et les interférences. *Langue française*, 8(8):31–61.

[Debyser *et al.*, 1967] Francis Debyser, Maurice Houis & Carlo Rojas (1967). *Grille de classement typologique des fautes*. BELC.

[Dédéyan & Largy, 2003] Alexandra Dédéyan & Pierre Largy (2003). Réviser la morphologie flexionnelle verbale : étude chez l'enfant et l'adulte. *Rééducation Orthophonique*, 213:97–113.

[Dédéyan *et al.*, 2006] Alexandra Dédéyan, Pierre Largy & Isabelle Negro (2006). Mémoire de travail et détection d'erreurs d'accord verbal : étude chez le novice et l'expert. *Langages*, 4(164):57–70.
url : URL:www.cairn.info/revue-langages-2006-4-page-57.htm.

[Déjean, 1998] Hervé Déjean (1998). *Concepts et algorithmes pour la découverte des structures formelles des langues*. Thèse de doctorat, Université de Caen.

[Demirtaş & Gümüş, 2009] Lokman Demirtaş & Hüseyin Gümüş (2009). De la faute à l'erreur : une pédagogie alternative pour améliorer la production écrite en fle. *Synergies Turquie*, 2:125–138.

[Dessus *et al.*, 2009] Philippe Dessus, Stefan Trausan-Matu, Virginie Zampa, Traian Rebedea, Sonia Mandin & Mihai Dascalu (2009). Vers un environnement-tuteur d'apprentissage dialogique.
url : https://hal.archives-ouvertes.fr/hal-00404842.

[Di Benedetto, 1958] Vincenzo Di Benedetto (1958). Dionisio trace e la techne a lui attribuita. *Annali della Scuola Normale di Pisa*, XXVII:169–210. ISSN : 0392-095X.

[Dister, 1997] Anne Dister (1997). Problématique des fins de phrase en traitement automatique du français. Actes de *Colloque international et interdisciplinaire de Liège* (Liège, Belgique, 13-15 mars 1997). ISBN : 2-8011-1200-3.

[Doll & Coulombe, 2004] Frédéric Doll & Claude Coulombe (2004). L'avenir des correcteurs grammaticaux : un point de vue industriel. *BULAG*, Correction automatique : bilan et perspectives(29):33–50. ISBN : 2-84867-080-0.

[Dortier, 2004] Jean-François Dortier (2004). *Dictionnaire des sciences humaines*. Éditions Sciences Humaines.

[Dubois *et al.*, 1994] Jean Dubois, Mathée Giacomo, Louis Guespin, Christiane Marcellesi, Jean-Baptiste Marcellesi & Jean-Pierre Mével (1994). *Dictionnaire de linguistique*. Larousse, 2002.

[Ducard *et al.*, 1995] Dominique Ducard, Renée Honvault & Jean-Pierre Jaffré (1995). *L'orthographe en trois dimensions*. Théories & pratiques. Nathan pédagogie. ISBN : 2-09-120417-X.

[Ducrot & Schaeffer, 1972] Oswald Ducrot & Jean-Marie Schaeffer (1972). *Nouveau dictionnaire encyclopédique des sciences du langage*. Éditions Seuil, 1995.

[EAGLES, 2000] EAGLES (2000). Corpus encoding standard.
url : http://www.cs.vassar.edu/CES/.

[Fayol & Got, 1991] Michel Fayol & Constance Got (1991). Automatisme et contrôle dans la production écrite : les erreurs d'accord sujet verbe chez l'enfant et l'adulte. *L'année psychologique*, 91(2):187–205.
url : /web/revues/home/prescript/article/psy_0003-5033_1991_num_91_2_29453.

[Fayol & Largy, 1992] Michel Fayol & Pierre Largy (1992). Une approche cognitive fonctionnelle de l'orthographe grammaticale. *Langue Française*, 95:80–98. ISSN : 0023-8368.

url : http://www.persee.fr/web/revues/home/prescript/article/lfr_0023-8368_1992_num_95_1_5773.

[Fayol *et al.*, 1994] Michel Fayol, Pierre Largy & Patrick Lemaire (1994). When cognitive overload enhances subject–verb agreement errors. a study in french written language. *Quarterly Journal of Experimental Psychology*, 47:437–464.

[Fayol & Pacton, 2006] Michel Fayol & Sébastien Pacton (2006). L'accord du participe passé : entre compétition de procédures et récupération en mémoire. *Langue française*, 151:59–73.
url : http://www.persee.fr/web/revues/home/prescript/article/lfr_0023-8368_2006_num_151_3_6774.

[Fitzgerald & Markham, 1987] Jill Fitzgerald & Lynda R. Markham (1987). Teaching children about revision in writing. *Cognition and instruction*, 4(1):3–24.

[Flower *et al.*, 1986] Linda Flower, John R. Hayes, Linda Carey, Karen Schriver & James Stratman (1986). Detection, diagnosis and the strategies of revision. *College composition and communication*, 37(1):16–55.

[Fontenelle, 2004] Thierry Fontenelle (2004). When syntax meets semantics in electronic dictionaries. Actes de *Symposium Internacional de Lexicografia* (Barcelone, Espagne, 16-18 mai 2002), pages 81–88. ISBN : 84-96367-06-1.

[Fontenelle, 2005a] Thierry Fontenelle (2005a). Dictionnaires et outils de correction linguistique. *Revue française de linguistique appliquée*, X-2:119–128.

[Fontenelle, 2005b] Thierry Fontenelle (2005b). Identifying tokens : Is word-breaking so easy ? *In* Philippe Hiligsmannn, Guy Janssens & Joseph Vromans, éditeurs, *Woord voor woord, zin voor zin : Liber Amicorum voor Siegfried Theissen*, pages 109–115. Koninklijke Academie voor Nederlandse Taal- en Letterkunde. ISBN : 9072474627.

[Fontenelle, 2006] Thierry Fontenelle (2006). Les nouveaux outils de correction linguistique de microsoft. Actes de *TALN (Traitement Automatique des Langues Naturelles)* (Louvain, Belgique, 10-13 avril 2006), volume 1 de *Cahiers du Cental*, pages 3–19. Presses Universitaires de Louvain. ISBN : 2-87463-023-3.

[Franck & Hupet, 2002] Julie Franck & Michel Hupet (2002). L'autonomie de la syntaxe revisitée : flux d'information dans la réalisation de l'accord grammatical. *In* Agnès Florin ans José Morais, éditeur, *La maîtrise du langage*, pages 61–78. Presses Universitaires de Rennes. ISBN : 2-86847-692-9.

[François, 1974] Frédérique François (1974). *L'enseignement et la diversité des grammaires.* Hachette. ISBN : 2-01-000260-1.

[Freeman & Tukey, 1950] Murray F. Freeman & John W. Tukey (1950). Transformations related to the angular and the square root. *The Annals of Mathematical Statistics*, 21(4):607–611.
url : http://projecteuclid.org/DPubS/Repository/1.0/Disseminate?view=body&id=pdf_1&handle=euclid.aoms/1177729756.

[Frei, 1928] Henri Frei (1928). *La grammaire des fautes, introduction à la linguistique fonctionnelle. assimilation et différenciation, brièveté et invariabilité, expressivité.* Thèse de doctorat, Genève.

[Friburger *et al.*, 2000] Nathalie Friburger, Anne Dister & Denis Maurel (2000). Améliorer le découpage en phrases sous intex. Actes de *Troisièmes journées Intex* (Liège, Belgique, 13-14 juin 2000), numéro 36, pages 181–200.

[Fuchs, 1996] Catherine Fuchs (1996). *Les ambiguïtés du français*. OPHRYS.

[Galisson & Coste, 1976] Robert Galisson & Daniel Coste (1976). Dictionnaire de didactique des langues.

[Gaonac'h & Larigauderie, 2000] Daniel Gaonac'h & Pacsale Larigauderie (2000). *Mémoire et fonctionnement cognitif : La mémoire de travail*. Armand Colin. BU-Droit-Lettres - 153.1GAON - 3e étage. ISBN : 2-200-01826-0.

[Gardent & Vandenberghes, 2009] Claire Gardent & Nathanaël Vandenberghes (2009). Dicovalence-easy.
url : `http://www.loria.fr/~gardent/resources/dicovalence-easy-240709.txt`.

[Gary-Prieur, 1985] Marie-Noelle Gary-Prieur (1985). *De la grammaire à la linguistique : L'etude de la phrase*. Armand Colin. ISBN : 2-200-31213-X.

[Gattegno, 2010] Caleb Gattegno (2010). *Teaching foreign languages in schools : the silent way*. Educational Solutions Worldwide Inc.

[Giguet, 1998] Emmanuel Giguet (1998). *Méthode pour l'analyse automatique de structures formelles sur documents multilingues*. Thèse de doctorat, Université de Caen.

[Gillam, 2003] Richard Gillam (2003). *Unicode demystified : a practical programmer's guide to the encoding standard*. Addison-Wesley. ISBN : 0-201-70052-2.

[Granger, 2007] Sylviane Granger (2007). Corpus d'apprenants, annotation d'erreurs et alao : une synergie prometteuse. *Cahier de lexicologie*, 91(2007-2):117–132.

[Granger *et al.*, 2001] Sylviane Granger, Anne Vandeventer & Marie-Josée Hamel (2001). Analyse de corpus d'apprenants pour l'elao basé sur le tal. *In* Béatrice Daille & Laurent Romary, éditeurs, *Linguistique de corpus*, volume 42 de *Traitement Automatique des Langues*, pages 609–621. Hermès, Lavoisier. ISBN : 2-7462-0411-8.

[Grevisse, 1993] Maurice Grevisse (1993). *Le bon usage*. De Boeck - Duculot, treizième édition par andré goosse.

[Grondeux, 2000] Anne Grondeux (2000). La grammatica positiva dans le bas moyen-âge. *In* Sylvain Auroux, Ernst Frideryk Konrad Koerner, Hans-Josef Niederehe & Kees Versteegh, éditeurs, *History of the Language Sciences : An International Handbook on the Evolution of the Study of Language from the Beginnings to the Present*, volume 1, chapitre 81, pages 598–610. Walter de Gruyter. ISBN : 978-3-11-011103-3.

[Grunig, 1993] Blanche-Noëlle Grunig (1993). Charges mémorielles et prédictions syntaxiques. *Cahiers de Grammaire*, 18:13–29.

[Habert, 1998] Benoît Habert (1998). *De l'écrit au numérique : constituter, normaliser et exploiter les corpus électroniques*. InterEditions. ISBN : 2-225-82953-5.

[Habert *et al.*, 1997] Benoît Habert, Adeline Nazarenko & André Salem (1997). *Les linguistiques de corpus*, volume Collection U. Série Linguistique. Armand Colin.

[Hayes, 1996] John R. Hayes (1996). The science of writing : theories, methods, individual differences, and applications. *In* Levy & Ransdell [1996], chapitre A new framework for understanding cognition and affect in writing, pages 1–28.

[Hayes, 2004] John R. Hayes (2004). What triggers revision ? *In* Allal *et al.* [2004], pages 9–20. ISBN : 1-4020-7729-7.

[Hayes & Flower, 1980] John R. Hayes & Linda Flower (1980). Identifying the organization of writing processes. *In* Lee Gregg & Erwin Steinberg, éditeurs, *Cognitive Processes in Writing*, pages 3–30. Lawrence Erlbaum Associates.

[Hayes *et al.*, 1987] John R. Hayes, Linda Flower, Karen A. Schriver, James F. Stratman & Linda Carey (1987). Cognitive processes in revision. *In* Sheldon Rosenberg, éditeur, *Advances in applied psycholinguistics*, volume 2 - Reading, writing, and language learning, pages 176–240. Cambridge University Press.

[Heurley, 2006] Laurent Heurley (2006). La révision de texte : l'approche de la psychologie cognitive. *Langages*, 4(164):10–25.
url : `http://www.persee.fr/web/revues/home/prescript/article/lgge_0458-726x_2006_num_40_164_2669`.

[Heurley & Garnier, 2002] Laurent Heurley & Franck Garnier (2002). La production des textes techniques écrits. *In* Michel Fayol, éditeur, *Traité de sciences cognitives : la production du langage*, pages 227–247. Hermès.

[Howell, 1998] David C. Howell (1998). *Méthodes statistiques en sciences humaines.* De Boeck Université. ISBN : 978-2744500084.

[Hugo, 1827] Victor Hugo (1827). *Cromwell (préface).* Eugène Renduel.

[Ildefonse, 1997] Frédérique Ildefonse (1997). *La naissance de la grammaire dans l'antiquité grecque.* Librairie Philosophique J. Vrin. ISBN : 978-2-7116-1311-3.

[Imprimerie Nationale, 2002] Imprimerie Nationale (2002). *Lexique des règles typographiques en usage à l'imprimerie nationale.* Imprimerie Nationale. ISBN : 978-2-7433-0482-9.

[INSEE, 2009] INSEE (2009). Statistiques sur les ressources et les conditions de vie (srcv).
url : `http://www.insee.fr/fr/themes/tableau.asp?reg_id=0&ref_id=NATnon05140`.

[Jacques, 2002] Marie Paule Jacques (2002). Pourquoi une linguistique de corpus ? *In* Williams [2005], pages 21–30.

[Jacquet-Pfau, 2001] Christine Jacquet-Pfau (2001). Correcteurs orthographiques et grammaticaux. quel(s) outil(s) pour quel rédacteur ? *Revue française de linguistique appliquée*, IV:81–94.

[Jaffré, 2003] Jean-Pierre Jaffré (2003). Les commentaires métagraphiques. *Faits de langues*, 22:67–76.

[Jespersen, 1924] Otto Jespersen (1924). *The philosophy of grammar.* George Allen and Unwin Ltd.

[Johnson & Kotz, 1969] Norman Lloyd Johnson & Samuel Kotz (1969). *Distributions in statistics : Discrete distributions.* Houghton Mifflin.

[Joshi, 1987] Arvind Joshi (1987). Introduction to tree adjoining grammar. *In* Alexis Manaster-Ramer, éditeur, *Mathematics of Language : Proceedings of a conference held at the University of Michigan, Ann Arbor, October 1984*, pages 87–114. John Benjamins. ISBN : 978-90-272-2049-3.

[Jousse & Polguère, 2005] Anne-Laure Jousse & Alain Polguère (2005). Le dico et sa version dicouèbe. document descriptif et manuel d'utilisation. Document technique, Département de linguistique et de traduction. Université de Montréal.

[Judd *et al.*, 1995] Charles M. Judd, Gary H. McClelland & Sara E. Culhane (1995). Data analysis : continuing issues in the everyday analysis of psychological data. *Annual Review of Psychology*, 46:433–465.

[Kahane, 2000] Sylvain Kahane (2000). Présentation. *In* Sylvain Kahane, éditeur, *Grammaires de dépendance*, volume 41 de *Traitement Automatique des Langues*, pages 3–8. Hermès Science. ISSN : 1248-9433.

[Kahane, 2001] Sylvain Kahane (2001). Grammaires de dépendance formelles et théorie sens-texte. Actes de *TALN (Traitement Automatique des Langues Naturelles)* (Tours, France, 2-5 juillet 2001).
url : `http://olst.ling.umontreal.ca/pdf/Kahane2001.pdf`.

[Kaplan & Bresnan, 1995] Ronald Kaplan & Joan Bresnan (1995). Lexical-functional grammar : a formal system for grammatical representation. *In* Mary Dalrymphe, Ronald M. Kaplan, John T. Maxwell III & Annie Zaeenen, éditeurs, *Formal Issues in Lexical Functional Grammar*, pages 29–130. MIT Press.

[Kellogg, 1996] Ronald T. Kellogg (1996). A model of working memory in writting. *In* Levy & Ransdell [1996], pages 57–72.

[Lacroix, 2009] Chantal Lacroix (2009). Les dépenses de consommation des ménages en biens et services culturels et télécommunications. *Culture Chiffres*, 2009-2.
url : `http://www2.culture.gouv.fr/culture/deps/2008/pdf/Cchiffres09_2.pdf`.

[Lallot, 1998] Jean Lallot (1998). *La grammaire de denys le thrace,*. CNRS-Éditions.

[Laporte, 1990] Éric Laporte (1990). Le dictionnaire phonémique delap. *Langue Française*, 87:59–70.

[Largy, 2003] Pierre Largy (2003). Du contrôle de l'orthographe grammaticale. 1re partie : du contrôle pré-graphique. *Le Langage et l'Homme - Logopédie, psychologie, audiologie*, XXX-VIII(2):139–152. ISBN : 2-930342-17-X.

[Largy *et al.*, 2004a] Pierre Largy, Lucile Chanquoy & Alexandra Dédéyan (2004a). Orthographic revision : the case of subject-verb agreement in french. *In* Linda Allal, Lucile Chanquoy & Pierre Largy, éditeurs, *Revision : Cognitive and Instructional Processes*, pages 39–62. Kluwer Academic Publishers. ISBN : 1-4020-7729-7.

[Largy *et al.*, 2005] Pierre Largy, Marie-Paule Cousin & Alexandra Dédéyan (2005). Produire et réviser la morphologie flexionnelle du nombre : de l'accès à une expertise. *In* Société Française de Psychologie [2005], pages 339–350. ISSN : 0033-2984.
url : `http://www.sciencedirect.com/science/article/pii/S0033298405000300`.

[Largy & Dédéyan, 2002] Pierre Largy & Alexandra Dédéyan (2002). Automatisme en détection d'erreurs d'accord sujet-verbe : étude chez l'enfant et l'adulte. *L'année psychologique*, 102(2):201–234.

[Largy *et al.*, 2004b] Pierre Largy, Alexandra Dédéyan & Michel Hupet (2004b). Orthographic revision : A developmental study of how revisers check verbal agreements in written texts. *British Journal of Developmental Psychology*, 74:533–550.
url : `http://w3.coll-tble-lang.univ-tlse2.fr/pierrelargy/BJEP2004.pdf`.

[Larousse, 2006] Pierre Larousse (2006). Le petit larousse illustré.

[Larruy, 2003] Martine Marquillo Larruy (2003). *L'interprétation de l'erreur*. CLE International/VUEF.

[László, 2009] Németh László (2009). Lightproof grammar checker.
url : `http://extensions.services.openoffice.org/project/lightproof`.

[Lebarbé, 2002] Thomas Lebarbé (2002). *Hiérarchie inclusive des unités linguistiques en analyse syntaxique coopérative*. Thèse de doctorat, Université de Caen - GREYC.

[Lebarbé & Girault, 2000] Thomas Lebarbé & François Girault (2000). Tapas : Traitement et analyse par perception augmentée en syntaxe. *Revue française de linguistique appliquée*, V-2:71–83.

[Levy & Ransdell, 1996] C. Michael Levy & Sarah Ellen Ransdell (1996). *The science of writing : theories, methods, individual differences, and applications*. Lawrence Erlbaum Associates.

[Lucci & Millet, 1994] Vincent Lucci & Agnès Millet (1994). *L'orthographe de tous les jours ; enquête sur les pratiques orthographiques des français*. Champion.

[Luhtala, 2005] Anneli Luhtala (2005). *Grammar and philosophy in late antiquity : a study of priscian's sources*. Studies in the History of the Language Science. John Benjamins Publishing Company. ISBN : 90-272-4598-3.

[Lusson, 2013] Charlotte Lusson (2013). *Influence des facteurs non syntaxiques sur l'accord en nombre : approche développementale*. Thèse de doctorat, Université Nice Sophia Antipolis.

[Luste-Chaa, 2009] Olha Luste-Chaa (2009). *Les acquisitions lexicales en français langue seconde : conceptions et applications*. Thèse de doctorat, Université Paul Verlaine, Metz.

[Lux-Pogodalla & Polguère, 2011] Veronika Lux-Pogodalla & Alain Polguère (2011). Construction of a french lexical network : Methodological issues. Actes de *First international Workshop on Lexical Resources, WoLeR 2011. An ESSLI Workshop* (Ljubljana, Slovénie), pages 54–61.

[Madec, 2004] Henir Madec (2004). Une approche cognitive de la correction automatique des fautes de syntaxe. *BULAG*, Correction automatique : bilan et perspectives(29):85–103. ISBN : 2-84867-080-0.

[Marshman, 2003] Elizabeth Marshman (2003). Construction et gestion des corpus : Résumé et essai d'uniformisation du processus pour la terminologie. *Définition des normes pour le groupe ÉCLECTIK*.
url : `http://olst.ling.umontreal.ca/pdf/terminotique/corpusnormes.pdf`.

[Martinet, 1985] André Martinet (1985). *Syntaxe générale*. Armand Colin. ISBN : 2-200-31211-3.

[Max & Wisniewski, 2010] Aurélien Max & Guillaume Wisniewski (2010). Mining naturally-occurring corrections and paraphrases from wikipedia's revision history. Actes de *Language Resources and Evaluation (LREC'10)* (Valletta, Malte, 19-21 mai 2010).

[Mayaffre, 2005] Damon Mayaffre (2005). Rôle et place des corpus en linguistique : réflexions introductives. *Texto !*, X(4).
url : `http://www.revue-texto.net/Reperes/Themes/Mayaffre_Corpus.html`.

[McEnery & Xiao, 2005] Anthony McEnery & Richard Xiao (2005). Character encoding in corpus construction. *In* Martin Wynne, éditeur, *Developing Linguistic Corpora : a Guide to Good Practice*, chapitre 4, pages 47–58. Oxbow Books. ISSN : 1463 5194.
url : `http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm`.

[Meigret, 1550] Louis Meigret (1550). *Tretté de la grammere françoeze*.

[Mel'čuk, 1984] Igor Mel'čuk, éditeur (1984). *Dictionnaire explicatif et combinatoire du français contemporain : recherches lexico-sémantiques i*, volume 1. Presses de l'Université de Montréal. ISBN : 2-7606-0659-7.

[Mel'čuk, 1988a] Igor Mel'čuk (1988a). *Dependency syntax : Theory and practice*. The SUNY Press. ISBN : 0-88706-450-7.

[Mel'čuk, 1988b] Igor Mel'čuk, éditeur (1988b). *Dictionnaire explicatif et combinatoire du français contemporain : recherches lexico-sémantiques ii*, volume 2. Presses de l'Université de Montréal. ISBN : 2-7606-0804-2.

[Mel'čuk, 1992] Igor Mel'čuk, éditeur (1992). *Dictionnaire explicatif et combinatoire du français contemporain : recherches lexico-sémantiques iii*, volume 3. Presses de l'Université de Montréal. ISBN : 2-7606-1559-6.

[Mel'čuk, 1997] Igor Mel'čuk (1997). *Vers une linguistique sens-texte. leçon inaugurale*. Collège de France.

[Mel'čuk, 1999] Igor Mel'čuk, éditeur (1999). *Dictionnaire explicatif et combinatoire du français contemporain : recherches lexico-sémantiques iv*, volume 4. Presses de l'Université de Montréal. ISBN : 2-7606-1738-6.

[Mel'čuk *et al.*, 1995] Igor Mel'čuk, André Clas & Alain Polguère (1995). *Introduction à la lexicologie explicative et combinatoire*. Duculot.

[Mout, 2013] Tiphaine Mout (2013). *L'orthographe du français : usages et représentations d'adultes socio-differencies. approche pluridisciplinaire*. Thèse de doctorat, Université de Grenoble.

[Naber, 2003a] Daniel Naber (2003a). Languagetool - open source language checker.
url : `http://www.languagetool.org`.

[Naber, 2003b] Daniel Naber (2003b). *A rule-based and grammar checker*. Thèse de doctorat, Technishe Fakultät, Universität Beilefeld.

[Nadasdi & Sinclair, 2001] Terry Nadasdi & Stéfan Sinclair (2001). Bonpatron. Nadaclair Language Technologies.
url : `http://bonpatron.com`.

[Negro, 2003] Isabelle Negro (2003). Le rôle des informations conceptuelles sur le traitement du nombre chez des rédacteurs experts. *Le Langage et l'Homme - Logopédie, psychologie, audiologie*, XXXVIII(2):123–137. ISBN : 2-930342-17-X.

[Nerima *et al.*, 2006] Luka Nerima, Violeta Seretan & Eric Wehrli (2006). Le problème des collocations en tal. *Nouveaux cahiers de linguistique française*, 27:95–115.

[New, 2006] Boris New (2006). Lexique 3 : Une nouvelle base de données lexicales. Actes de *TALN (Traitement Automatique des Langues Naturelles)* (Louvain, 10-13 avril 2006).

[Nicot, 1606] Jean Nicot (1606). *Le thresor de la langue françoyse tant ancienne que moderne.*

[Nouveau, 2007] Dominique Nouveau (2007). Homophonie et morphologie verbale en fle : Les groupes verbaux en -ir. *Radboud University Nijmegen.*
url : `http://www.ru.nl/publish/pages/530439/homophonieenfleversionlongue.pdf`.

[Olive, 2004] Thierry Olive (2004). Working memory in writing : empirical evidence from the dual-task technique. *European Psychologist*, 9(1):32–42.

[Olive & Piolat, 2005] Thierry Olive & Annie Piolat (2005). Le rôle de la mémoire de travail dans la production écrite de textes. *In* Société Française de Psychologie [2005], pages 373–390.

[Palsgrave, 1530] John Palsgrave (1530). *L'esclarcissement de la langue françoise.*

[Péry-Woodley, 1995] Marie-Paule Péry-Woodley (1995). Quels corpus pour quels traitements automatiques ? *Traitement Automatique des Langues*, 36(1-2):213–232. ISSN : 1248-9433.

[Picoche, 2010] Jacqueline Picoche (2010). *Précis de lexicologie française.* Éditions Vigdor. ISBN : 2-84771-033-7.

[Piolat, 2004] Anne Piolat (2004). Approche cognitive de l'activité rédactionnelle et de son acquisition. le rôle de la mémoire de travail. *Linx*, 51:55–74.
url : `http://linx.revues.org/174`.

[Piolat, 2007] Anne Piolat (2007). Les avantages et les inconvénients d'un traitement de texte pour réviser. *In* Jocelyne Bisaillon, éditeur, *La révision professionnelle : processus, stratégies et pratiques*, pages 189–211. Éditions Nota bene. ISBN : 9782895182696.

[Piolat *et al.*, 2004] Annie Piolat, Jean-Yves Roussey, Thierry Olive & Murielle Amada (2004). Processing time and cognitive effort in revision : effects of error type and of working memory capacity. *In* Allal *et al.* [2004], pages 21–38. ISBN : 1-4020-7729-7.

[Polguère, 1998] Alain Polguère (1998). La théorie sens-texte. *Dialangue*, 8-9:9–30.

[Polguère, 2012] Alain Polguère (2012). Lexicographie des dictionnaires virtuels. *In* Juri Apresjan, Igor Boguslavsky, Marie-Claude L'Homme, Leonid Iomdin, Jasmina Milićević, Alain Polguère & Leo Wanner, éditeurs, *Meanings, Texts, and Other Exciting Things. A Festschrift to Commemorate the 80th Anniversary of Professor Igor Alexandrovič Mel'čuk*, pages 509–523. Jazyki slavjanskoj kultury Publishers.

[Pollard & Sag, 1994] Carl Jesse Pollard & Ivan Sag (1994). *Head-driven phrase structure grammar*. University of Chicago Press. ISBN : 9780226674476.

[Pollard, 1979] J. H. Pollard (1979). *A handbook of numerical and statistical techniques : with examples mainly from the life sciences*. Cambridge University Press. ISBN : 9780521297509.

[Porquier, 1977] Rémy Porquier (1977). L'analyse des erreurs. problèmes et perspectives. *Études de linguistique appliquée*, Apprentissage et enseignement de la grammaire d'une langue non-maternelle(25):23–43.

[Queneau, 1959] Raymond Queneau (1959). *Zazie dans le métro*. Gallimard.

[Rabadi & Odeh, 2010] Najib Rabadi & Akram Odeh (2010). L'analyse des erreurs en fle chez des apprenants jordaniens et bahreïniens. *Jordan Journal of Modern Languages and Literature*, 2(2):163–177.

[Riegel *et al.*, 1994] Martin Riegel, Jean-Christophe Pellat & René Rioul (1994). *Grammaire méthodique du français*. Presses Universitaires de France, 2003.

[Robert, 2006] Paul Robert (2006). Le petit robert de la langue française.

[Rock, 2001] Frances Rock (2001). Policy and practice in the anonymisation of linguistic data. *International journal of corpus linguistics*, 6(1):1–26. ISSN : 1384-6655.

[Ronez, 2011] Olivier Ronez (2011). Grammalecte.
url : `http://www.dicollecte.org/grammalecte/`.

[Ronez, 2014] Olivier Ronez. dicollecte. (2014).
url : `http://www.dicollecte.org/home.php?prj=fr`.

[Roussey & Piolat, 2005] Jean-Yves Roussey & Anne Piolat (2005). La révision du texte : une activité de contrôle et de réflexion. *In* Société Française de Psychologie [2005], pages 351–372.
url : `http://sites.univ-provence.fr/wpsycle/documentpdf/DocPiolat/Publications/RousseyPiolaPFt2005.pdf`.

[Sabah, 1989] Gérard Sabah (1989). *L'intelligence artificielle et le langage*, volume 2. Éditions Hermès.

[Scannell, 2003] Kevin Scannell (2003). An gramadóir.
url : `http://borel.slu.edu/gramadoir`.

[Scardamalia & Bereiter, 1983] Marlene Scardamalia & Carl Bereiter (1983). The development of evaluative, diagnostic and remedial capabilitiesin children's composing. *In* M. Martlew, éditeur, *The psychology of written language : A developmental approache*, pages 67–95. Wiley.

[Selinker, 1972] Larry Selinker (1972). Interlanguage. *IRAL (International Review of Applied Linguistics in Language Teaching)*, 10:209–231.

[Simard, 1997] Claude Simard (1997). *Éléments de didactique du français langue première*. Pratiques pédagogiques. De Boeck Université. ISBN : 2-8041-2614-5.

[Sinclair, 1991] John Sinclair (1991). *Corpus concordance collocation*. Oxford University Press.

[Sinclair, 1996] John Sinclair (1996). Preliminary recommendations on corpus typology. Document technique, EAGLES (Expert Advisory on Language Engineering standards).
url : `http://www.ilc.cnr.it/EAGLES96/pub/eagles/corpora/corpustyp.ps.gz`.

[Sinclair, 2005] John Sinclair (2005). Corpus and text - basic principles. *In* Martin Wynne, éditeur, *Developing Linguistic Corpora : a Guide to Good Practice*, chapitre 1, pages 1–16. Oxbow Books. ISSN : 1463 5194.
url : `http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm`.

[Smedt, 1990] Koenraad De Smedt (1990). Ipf : An incremental parallel formulator. *In* Robert Dale, Chris Mellish & Michael Zock, éditeurs, *Current research in natural language generation*, pages 167–192. Academic Press.

[Société Française de Psychologie, 2005] Société Française de Psychologie (2005). *Psychologie française*, volume 50. Elsevier Masson.

[Souque, 2006] Agnès Souque (2006). Générateur automatique d'affixes pour la compression des dictionnaires de correction orthographique.

[Souque, 2007] Agnès Souque (2007). Conception et développement d'un formalisme de correction grammaticale automatique - application au français -. Master de master 2 sciencecs du langage, Université Stendhal - Grenoble 3.

[Souque, 2008] Agnès Souque (2008). Vers une nouvelle approche de la correction grammaticale automatique. Actes de *RECITAL (Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues)* (Avignon, 9-13 juin).

[Swiggers & Wouters, 2001] Pierre Swiggers & Alfons Wouters (2001). Philosophie du langage et linguistique dans l'antiquité classique. *In* Martin Haspelmath, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible, éditeurs, *Language Typology and Language Universals : an international Handbook*, volume 1, chapitre 13, pages 181–192. Walter de Gruyter. ISBN : 3-11-011423-2.

[Synapse Développement, 2006] Synapse Développement (2006). À propos de la correction grammaticale française et des correcteurs informatiques.
url : `http://www.synapse-fr.com/descr_technique/A_propos_des_correcteurs.htm`.

[Tesnière, 1959] Lucien Tesnière (1959). *Éléments de syntaxe structurale*. Klincksieck.

[Thouet, 2004] Myriam Thouet (2004). Prise en compte des propriétés sémantiques des unités lexicales pour améliorer les correcteurs. *BULAG*, Correction automatique : bilan et perspectives(29):153–162. ISBN : 2-84867-080-0.

[Tisserand, 2004] Jean-Sébastien Tisserand (2004). Parsers, grammaires formalisées et fautes de grammaire du français. *BULAG*, Correction automatique : bilan et perspectives(29):163–181. ISBN : 2-84867-080-0.

[Totereau *et al.*, 1998] Corinne Totereau, Pierre Barrouillet & Michel Fayol (1998). Overgeneralizations of number inflections in the learning of written french : The case of noun and verb. *British Journal of Developmental Psychology*, 16:447–464.

[Tournier, 1980] Claude Tournier (1980). Histoire des idées sur la ponctuation, des débuts de l'imprimerie à nos jours. *Langue française*, 45:28–40.
url : http://www.persee.fr/web/revues/home/prescript/article/lfr_0023-8368_1980_num_45_1_5261.

[Trouilleux, 2009] François Trouilleux (2009). Un analyseur de surface non déterministe pour le français. Actes de *TALN (Traitement Automatique des Langues Naturelles)* (Senlis, France, 24-26 juin 2009).
url : http://www.atala.org/taln_archives/TALN/TALN-2009/taln-2009-long-024.pdf.

[van den Eynde & Blanche-Benvéniste, 1978] Karel van den Eynde & Claire Blanche-Benvéniste (1978). Syntaxe et mécanismes descriptifs : présentation de l'approche pronominale. *Cahier de lexicologie*, 32:3–27.

[van den Eynde & Mertens, 2003] Karel van den Eynde & Piet Mertens (2003). La valence : l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies*, 13(1):63–104.
url : http://bach.arts.kuleuven.be/pmertens/papers/proton2002.pdf.

[van den Eynde & Mertens, 2010] Karel van den Eynde & Piet Mertens (2010). Le dictionnaire de valence dicovalence : manuel d'utilisation.
url : http://bach.arts.kuleuven.be/dicovalence/manuel_061117.pdf.

[Vergne, 1998] Jacques Vergne (1998). Entre arbre de dépendance et ordre linéaire, les deux processus de transformation : linéarisation, puis reconstruction de l'arbre. *Cahiers de Grammaire*, (23):95–136.

[Vergne, 1999] Jacques Vergne (1999). *Études et modélisation de la syntaxe des langues à l'aide de l'ordinateur. analyse syntaxique automatique non combinatoire.* Hdr, Université de Caen - GREYC.

[Vergne & Giguet, 1998] Jacques Vergne & Emmanuel Giguet (1998). Regards théoriques sur le "tagging". Actes de *TALN (Traitement Automatique des Langues Naturelles)* (Paris, France, 10-12 juin 1998).
url : http://www.atala.org/taln_archives/TALN/TALN-1998/taln-1998-long-010.pdf.

[Véronis, 2005] Jean Véronis (2005). Blog du 05 juillet 2005.
url : http://blog.veronis.fr/2005/07/rcr-pourriss-vos-texte.html.

[Vienney, 2004] Séverine Vienney (2004). Présentation. *In* Séverine Vienney & Mounira Bioud, éditeurs, *Correction automatique : bilan et perspectives*, numéro 29, pages 5–8. Presses Universitaires de Franche-Comté. ISBN : 2-84867-080-0.

[von Humboldt, 1836] Whilhelm von Humboldt (1836). *Über die verschiedenheit des menschlichen sprachbaues und ihren einfluss auf die geistige entwicklung des menschengeschlechts.* F. Dümmler.

[Wade-Stein & Kintsch, 2004] David Wade-Stein & Eileen Kintsch (2004). Summary street : interactive computer support for writing. *Cognition and instruction*, 22(3):333–362.

[Wikipédia, 2014] Wikipédia. Openoffice.org. (2014).
url : http://fr.wikipedia.org/wiki/OpenOffice.org.

[Williams, 2005] Geoffrey Williams (2005). *La linguistique de corpus.* Presses Universitaires de Rennes.

[Wisniewski *et al.*, 2010] Guillaume Wisniewski, Aurélien Max & François Yvon (2010). Recueil et analyse d'un corpus écologique de corrections orthographiques extrait des révisions de wikipédia. Actes de *TALN (Traitement Automatique des Langues Naturelles)* (Montréal, Québec, 19-23 juillet 2010).
url : `wicopaco.limsi.fr/pub/taln10.pdf`.

[Wooldridge, 2002] Michael Wooldridge (2002). *An introduction to multiagent systems.* John Wiley and sons, LTD.

[Zampa & Dessus, 2010] Virginie Zampa & Philippe Dessus (2010). Écrire des résumés pour apprendre des cours : un système d'aide à l'apprentissage autorégulé. Actes de *Tice2010.*

[Žolkovskij & Mel'čuk, 1965] Aleksandr Žolkovskij & Igor Mel'čuk (1965). O vozmožnom metode i instrumentax semantičeskogo sinteza âÂ» [sur une méthode possible de synthèse sémantique et les outils correspondants]. *Naučno-texničeskaja informacija*, 5:23–28.

[Žolkovskij & Mel'čuk, 1967] Aleksandr Žolkovskij & Igor Mel'čuk (1967). O semanticeskom sinteze [sur la synthèse sémantique]. *Problemy kibernetiki*, 19:177–238.

Annexes

Annexe A

Tableaux de données

A.1 Effectifs d'erreurs

Catégories d'erreurs	Dictées	Résumés	Mails	Texte FLE	CORPUS
adjectif	76	16	4	7	103
déterminant	15	6	2	10	33
nom	104	16	7	5	132
Total dans les syntagmes nominaux	*195*	*38*	*13*	*22*	*268*
verbe	18	8	21	4	51
attribut	1	3	1	3	8
p.p. avec verbe d'état	46	2	7	2	57
sujet-p.p. avec *avoir*	23	0	0	0	23
objet-p.p. avec *avoir*	15	0	1	3	19
Total dans les syntagmes verbaux	*103*	*13*	*30*	*12*	*158*
Total ACCORD	**298**	**51**	**43**	**34**	**426**
coréférence	0	1	1	1	3
euphonie	0	2	1	1	4
lemme	112	16	7	28	163
union	45	5	7	2	59
Total LEXIQUE	**157**	**24**	**16**	**32**	**229**
ajout	6	3	0	7	16
ordre	0	0	0	1	1
oubli	19	2	9	14	44
Total SYNTAXE	**25**	**5**	**9**	**22**	**61**
mode	46	3	11	5	65
temps	0	0	0	1	1
Total VERBE	**46**	**3**	**11**	**6**	**66**
abréviation	8	5	4	0	17
diacritique	180	7	22	13	222
graphie	220	99	27	46	392
majuscule	177	15	21	30	243
morphologie	6	0	0	5	11
segmentation	11	14	3	5	33
Total ORTHOGRAPHE	**602**	**140**	**77**	**99**	**918**
espace	210	15	74	56	355
signe	22	3	1	15	41
Total PONCTUATION	**232**	**18**	**75**	**71**	**396**
TOTAL ERREURS	**1360**	**241**	**231**	**264**	**2096**

Tableau A.1 : Effectifs d'erreurs par (sous-)catégories et types de textes

Catégories d'erreurs	Dictées	Résumés	Mails	Textes FLE	CORPUS
ACCORD	266	45	41	20	372
LEXIQUE	141	12	13	17	183
SYNTAXE	0	0	0	0	0
VERBE	39	3	11	4	57
ORTHOGRAPHE	390	52	46	65	553
Total erreurs homophones	**836**	**112**	**111**	**106**	**1165**

Tableau A.2 : Effectifs d'erreurs à caractère homophone

Catégories d'erreurs	Humain	Texte	Situation	Humain + Texte	Humain + Situation	Texte + Situation	Humain + Texte + Situation	CORPUS
ACCORD	265 (62,5%)	57 (13,4%)	4 (0,9%)	93 (21,9%)	4 (0,9%)	1 (0,2%)	0	424 **(100%)**
LEXIQUE	98 (42,8%)	91 (39,7%)	5 (2,2%)	26 (11,4%)	2 (0,9%)	3 (1,3%)	4 (1,7%)	229 **(100%)**
SYNTAXE	33 (55%)	1 (1,7%)	22 (36,7%)	0	4 (6,7%)	0	0	60 **(100%)**
VERBE	7 (10,4%)	36 (53,7%)	0	24 (35,8%)	0	0	0	67 **(100%)**
Toutes erreurs grammaticales	**403 (51,7%)**	**185 (23,7%)**	**31 (4%)**	**143 (18,3%)**	**10 (1,3%)**	**4 (0,5%)**	**4 (0,5%)**	**780 (100%)**

Tableau A.3 : Répartition des catégories d'erreurs par type de causes

A.2 Proportions moyennes d'erreurs

Catégories d'erreur	Dictées		Résumés		Mails		Textes FLE	
	occ.	(%)	occ.	(%)	occ.	(%)	occ.	(%)
Accord	298	(21,9%)	51	(21,2%)	43	(18,6%)	34	(12,9%)
Lexique	157	(11,5%)	24	(10,0%)	16	(6,9%)	32	(12,1%)
Syntaxe	25	(1,8%)	5	(2,1%)	9	(3,9%)	22	(8,3%)
Verbe	46	(3,4%)	3	(1,2%)	11	(4,8%)	6	(2,3%)
Orthographe	602	(44,3%)	140	(58,1%)	77	(33,3%)	99	(37,5%)
Ponctuation	232	(17,1%)	18	(7,5%)	75	(32,5%)	71	(26,9%)
Total	**1360**	**(100%)**	**241**	**(100%)**	**231**	**(100%)**	**264**	**(100%)**

Tableau A.4 : Proportion moyenne (ET) des catégories d'erreur dans les différents types de textes. Moyennes calculées à partir des effectifs totaux

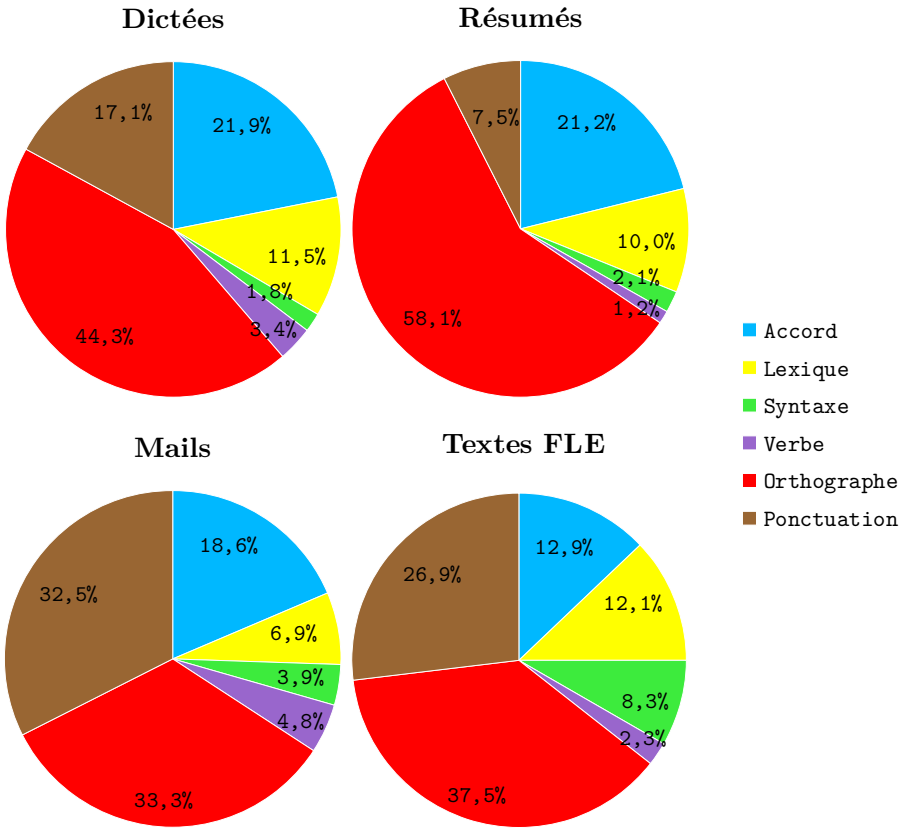


FIGURE A.1 : Proportion des catégories d'erreur dans les différents types de textes. Moyennes calculées à partir des effectifs totaux.

LEXIQUE	Dictées	Résumés	Mails	Textes FLE	CORPUS
coreférence	-	5,6% (23,6)	7,1% (26,7)	3,8% (13,9)	2% (13,2)
euphonie	-	8,3% (25,7)	7,1% (26,7)	3,8% (13,9)	2,3% (13,8)
lemme	65,2% (40,5)	72,2% (42,8)	46,4% (49,9)	86,5% (21,9)	66,3% (41,1)
union	34,8% (40,5)	13,9% (33,5)	39,3% (48,7)	5,8% (15)	29,4% (39,9)

Tableau A.5 : Proportion moyenne (ET) des erreurs de la catégorie LEXIQUE dans chaque type de textes

SYNTAXE	Dictées	Résumés	Mails	Textes FLE	CORPUS
ajout	27,1% (40,3)	62,5% (47,9)	-	40,3% (41,7)	29,2% (40,3)
ordre	-	-	-	4,2% (14,4)	1,3% (7,9)
oubli	72,9% (40,3)	37,5% (47,9)	100% (0)	55,6% (45,1)	69,6% (41,7)

Tableau A.6 : Proportion moyenne (ET) des erreurs de la catégorie SYNTAXE dans chaque type de textes

VERBE	Dictées	Résumés	Mails	Textes FLE	CORPUS
mode	100% (0)	100% (0)	100% (0)	90% (0,2)	99% (0,1)
temps	-	-	-	10% (0,2)	1% (0,1)

Tableau A.7 : Proportion moyenne (ET) des erreurs de la catégorie VERBE dans chaque type de textes

PONCTUATION	Dictées	Résumés	Mails	Textes FLE	CORPUS
espace	89% (29,9)	84,8% (34,5)	96,9% (12,5)	74,6% (37,4)	88% (30)
signe	11% (29,9)	15,2% (34,5)	3,1% (12,5)	25,4% (37,4)	12% (30)

Tableau A.8 : Proportion moyenne (ET) des erreurs de la catégorie PONCTUATION dans chaque type de textes

ORTHO-GRAPHE	Dictées	Résumés	Mails	Textes FLE	CORPUS
abreviation	0,3% (1,8)	5,9% (19,5)	5,6% (19,8)	-	2,2% (11,7)
diacritique	20,2% (26,4)	4,1% (13,6)	21,7% (37,6)	12,1% (17,1)	16,9% (26,9)
graphie	45,7% (35,4)	67,7% (29,1)	39% (46,2)	43,3% (26,7)	48,1% (36,7)
majuscule	30,1% (31,2)	11,1% (18,1)	28% (38,7)	29% (28,1)	26,3% (31,1)
morphologie	2,2% (12,2)	-	-	10,8% (27,4)	2,3% (12,6)
segmentation	1,4% (7,1)	11,3% (19,7)	5,7% (20,1)	4,9% (12,3)	4,2% (13,7)

Tableau A.9 : Proportion moyenne (ET) des erreurs de la catégorie ORTHOGRAPHE dans chaque type de textes

A.3 Densités d'erreurs

Catégories d'erreurs	Dictées	Résumés	Mails	Texte FLE	CORPUS
adjectif	0,003 (0,007)	0,004 (0,008)	0,001 (0,003)	0,005 (0,010)	0,003 (0,007)
déterminant	0,001 (0,002)	0,001 (0,003)	0,001 (0,004)	0,010 (0,018)	0,001 (0,006)
nom	0,005 (0,008)	0,003 (0,004)	0,002 (0,008)	0,002 (0,006)	0,004 (0,007)
verbe	0,001 (0,002)	0,002 (0,004)	0,007 (0,012)	0,004 (0,010)	0,002 (0,007)
attribut	0 (0,000)	0,001 (0,003)	0,000 (0,002)	0,004 (0,013)	0,001 (0,004)
p.p. avec verbe d'état	0,002 (0,004)	0,000 (0,002)	0,001 (0,004)	0,001 (0,004)	0,002 (0,004)
sujet-p.p. avec *avoir*	0,001 (0,003)	0	0	0	0,001 (0,002)
objet-p.p. avec *avoir*	0,001 (0,002)	0	0,001 (0,004)	0,001 (0,005)	0,001 (0,003)
Total ACCORD	**0,013 (0,016)**	**0,010 (0,011)**	**0,014 (0,016)**	**0,027 (0,034)**	**0,014 (0,018)**
coréférence	0	0,000 (0,002)	0,000 (0,001)	0,000 (0,001)	0,000 (0,001)
euphonie	0	0,000 (0,002)	0,000 (0,002)	0,002 (0,008)	0,000 (0,002)
lemme	0,005 (0,008)	0,004 (0,005)	0,003 (0,007)	0,024 (0,036)	0,006 (0,013)
union	0,002 (0,004)	0,001 (0,003)	0,002 (0,006)	0,001 (0,004)	0,002 (0,004)
Total LEXIQUE	**0,007 (0,009)**	**0,005 (0,006)**	**0,005 (0,011)**	**0,027 (0,037)**	**0,008 (0,014)**
ajout	0,000 (0,001)	0,001 (0,003)	0	0,008 (0,014)	0,001 (0,005)
ordre	0	0	0	0,001 (0,003)	0,000 (0,001)
oubli	0,001 (0,004)	0,001 (0,002)	0,003 (0,007)	0,010 (0,017)	0,002 (0,007)
Total SYNTAXE	**0,001 (0,005)**	**0,002 (0,005)**	**0,003 (0,007)**	**0,019 (0,019)**	**0,003 (0,009)**
mode	0,002 (0,005)	0,001 (0,002)	0,003 (0,009)	0,002 (0,005)	0,002 (0,006)
temps	0	0	0	0,001 (0,003)	0,000 (0,001)
Total VERBE	**0,002 (0,005)**	**0,001 (0,002)**	**0,003 (0,009)**	**0,003 (0,007)**	**0,002 (0,006)**
abréviation	0,000 (0,002)	0,001 (0,003)	0,001 (0,004)	0	0,001 (0,003)
diacritique	0,009 (0,028)	0,001 (0,005)	0,006 (0,018)	0,008 (0,015)	0,007 (0,023)
graphie	0,010 (0,016)	0,019 (0,016)	0,006 (0,012)	0,029 (0,026)	0,012 (0,017)
majuscule	0,008 (0,017)	0,003 (0,005)	0,006 (0,014)	0,017 (0,017)	0,008 (0,015)
morphologie	0,000 (0,002)	0	0	0,003 (0,008)	0,000 (0,003)
segmentation	0,001 (0,003)	0,003 (0,006)	0,001 (0,003)	0,002 (0,005)	0,001 (0,004)
Total ORTHOGRAPHE	**0,028 (0,057)**	**0,028 (0,018)**	**0,019 (0,031)**	**0,060 (0,041)**	**0,029 (0,048)**
espace	0,010 (0,014)	0,003 (0,008)	0,012 (0,027)	0,042 (0,049)	0,012 (0,023)
signe	0,001 (0,003)	0,000 (0,002)	0 (0,000)	0,008 (0,014)	0,001 (0,005)
Total PONCTUATION	**0,011 (0,014)**	**0,004 (0,008)**	**0,012 (0,027)**	**0,050 (0,056)**	**0,013 (0,025)**
TOTAL ERREURS	**0,063 (0,080)**	**0,049 (0,028)**	**0,055 (0,048)**	**0,185 (0,079)**	**0,069 (0,077)**

Tableau A.10 : Densité d'erreurs calculées ($\frac{nb\ d'erreurs}{nb\ de\ mots}$)

Résumé

Nous proposons un modèle de vérification grammaticale automatique gauche-droite issu de l'analyse d'un corpus d'erreurs tapuscrites. Les travaux menés en psychologie cognitive ont montré que le processus de révision procède au travers de la confrontation d'une attente à un résultat. Ainsi, la détection d'une erreur grammaticale reposerait, chez l'humain, sur une attente du réviseur non comblée. Ce principe est à la base du modèle que nous avons élaboré.

Pour faciliter la gestion des attentes du point de vue traitement numérique, nous convions deux concepts courants en TAL : le principe d'unification et la segmentation en chunks. Le premier est particulièrement adapté à la vérification des accords et le second constitue une unité de calcul intermédiaire permettant de définir des bornes simplifiant la recherche d'incohérences grammaticales. Enfin, l'originalité de ce modèle réside dans une analyse gauche-droite construite au fur et à mesure de la lecture/écriture.

Mots-clés: Correction grammaticale, TAL, Corpus, Unification, Chunk

Abstract

This thesis presents a model for automated left-right grammar checking based on analysis of a corpus of typescript errors. Studies in cognitive psychology have shown that the revision process works by confronting expectations with results. For humans, detecting a grammatical error therefore relies on an unfulfilled expectation on the part of the revisor. The model presented here is based on this principle.

In order to deal with expectations from the point of view of computational processing, two common concepts in NLP are called upon : the unification principle and chunk segmentation. The former is particularly adapted to checking agreements, while the latter provides an intermediate computational unit to delimit, and therefore simplify, detection of grammatical inconsistencies. Finally, the model's originality lies in the left-right analysis it provides, which is constructed as the text is produced/read.

Keywords: Grammar checking, NLP, Corpus, Unification, Chunk