Apport du TAL à l'exploitation linguistique d'un corpus scolaire longitudinal

Claire Wolfarth

▶ **To cite this version:**

Claire Wolfarth. Apport du TAL à l'exploitation linguistique d'un corpus scolaire longitudinal. Linguistique. Université Grenoble Alpes, 2019. Français. NNT : 2019GREAL025 . tel-02517320

HAL Id: tel-02517320
https://theses.hal.science/tel-02517320

Submitted on 24 Mar 2020

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

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

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEURE DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

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

Arrêté ministériel : 25 mai 2016

Présentée par

Claire WOLFARTH

Thèse dirigée par **Catherine Brissaud**, Professeure, **Université Grenoble Alpes**, et
co-encadrée par **Claude Ponton**, Maitre de conférences, **Université Grenoble Alpes**

préparée au sein du **Laboratoire LIDILEM**
dans l'**École Doctorale Langues, Littérature et Sciences Humaines**

Apport du TAL à l'exploitation linguistique d'un corpus scolaire longitudinal

Thèse soutenue publiquement le 09 décembre 2019, devant le jury composé de :

Mme Catherine BRISSAUD
Professeure, Université Grenoble Alpes, Directrice
Mme Claire DOQUET
Professeure, Université Sorbonne Nouvelle, Rapporteure
M. Cédrick FAIRON
Professeur, Université Catholique de Louvain, Rapporteur
Mme Karën FORT
Maitresse de conférences, Sorbonne Université, Examinatrice
M. Claude PONTON
Maitre de conférences, Université Grenoble Alpes, Co-encadrant
Mme Lilia TERUGGI
Professeure, Università degli Studi di Milano-Bicocca, Examinatrice

Remerciements

Je voudrais adresser ici un très grand merci à ma directrice et mon directeur de thèse, Catherine Brissaud et Claude Ponton, pour leur accompagnement tout au long de cette thèse, et tout particulièrement pour leur présence soutenue lors de ces derniers mois de rédaction. Malgré un cadre pas toujours favorable engendré par le milieu de l'enseignement supérieur et de la recherche, particulièrement pour les jeunes chercheurs et chercheuses, ils ont su m'accorder du temps et m'apporter un encadrement de qualité.

Je remercie Claire Doquet, Cédrick Fairon, Karën Fort et Lilia Teruggi d'avoir accepté de faire partie de mon jury et d'évaluer ce travail.

Un grand merci à Louise-Amélie Cougnon, Cédrick Fairon et toute l'équipe du *CENTAL* et du *MiiL* pour leur accueil lors de mon séjour de recherche. Ils ont su à la fois me pousser dans mon raisonnement scientifique et m'initier au mode de vie belge.

Je souhaiterais remercier chaleureusement tous les enseignants et enseignantes, tous les directeurs et directrices qui ont accepté de nous accueillir dans leurs écoles et dans leurs salles de classe et tous les parents d'élèves qui ont accepté que nous utilisions les productions de leurs enfants. Je souhaiterais remercier plus chaleureusement encore tous les élèves qui ont accepté de nous confier leurs productions.

Je profite de cet espace pour remercier tous les étudiants et étudiantes, stagiaires ou vacataires, qui ont travaillé sur le projet *Scoledit* et qui ont chacun et chacune successivement apporté une pierre à cet édifice.

Ce travail représente pour moi la fin d'un long parcours de formation scolaire et institutionnel, pour lequel je ne peux qu'être reconnaissante. Mais je voudrais remercier aussi tous les collectifs qui m'ont nourrie et formée, tout autant que les institutions républicaines ont pu le faire. Parce qu'ils représentent des espaces où il m'a été possible de prendre du recul sur les expériences, bonnes ou mauvaises, que je vivais à l'université, mais aussi de co-construire des solutions ou des alternatives aux difficultés que nous, puisque je n'étais pas la seule, y rencontrions, certains d'entre eux ont contribué plus ou moins directement à l'aboutissement de cette thèse.

Enfin, je voudrais remercier l'ensemble de mes amis et amies, colocataires, collègues et membres de ma famille qui m'ont accompagnée et soutenue tout au long de mon parcours et de mes projets.

J’adresse un merci tout particulier à celles et ceux qui ont constitué une « cellule de crise » dans les derniers jours de rédaction afin de m’aider à finaliser les aspects très matériels de cette tâche.

Table des matières

Remerciements v
Table des matières vii
Glossaire xvii
Introduction 1

1. Contexte 1
2. Le projet *Scoledit* 2
3. Inscription dans un dialogue multi-équipe 3
4. Problématique 4
5. Plan de la thèse 5

Partie 1 - Linguistique de corpus, TAL et corpus scolaires 7

Chapitre 1 - Corpus et linguistique de corpus 9

1. Qu'est-ce qu'un corpus ? 9

1.1. Critère linguistique 10
1.2. Critère d'authenticité 12
1.3. Critère de représentativité 12
1.4. Degré de numérisation 13

2. Linguistique de corpus, rappels historiques 15

2.1. Premières approches sur corpus 15
2.2. Linguistique introspective et apparition des premiers grands corpus 15
2.3. Expansion des corpus et enrichissement des corpus 16
2.4. Des approches sur corpus à la linguistique de corpus 17

3. La constitution de la linguistique de corpus en tant que discipline 18

3.1. Contre la linguistique introspective 19
3.2. Apports supposés de la linguistique de corpus 19
3.3. Critiques de la linguistique de corpus 20
3.4. Statut de discipline 21
3.5. Courants internes 21

4. Conclusion 22

Chapitre 2 - Traitement automatique des langues et apport à l'exploitation des corpus 25

1. Comprendre le traitement automatique des langues 25

1.1. Domaines d'application 26
1.2. Fonctionnement d'un système de traitement automatique des langues 27
1.3. Méthodes du TAL 28
1.4. Evolution du TAL 29
1.5. TAL et corpus 30

2. Apport du traitement automatique des langues à l'exploitation de corpus 30
2.1. Elaboration de corpus et traitement automatique des langues 31
2.1.1. Formalisme et codage des corpus 31
2.1.2. La segmentation 32
2.1.3. La lemmatisation et l'étiquetage morphosyntaxique 32
2.1.4. L'annotation syntaxique 33
2.1.5. L'annotation sémantique 33
2.1.6. L'alignement multilingue 33
2.2. Outils d'exploration des corpus 34
2.2.1. Outils génériques et outils spécifiques 34
2.2.2. Types d'exploration 35
3. Apports du traitement automatique des langues aux corpus d'écrits peu normés 35
3.1. Le traitement automatique des langues et les corpus de communication électronique écrite 36
3.2. Apports du TAL au traitement des corpus d'apprenants 38
3.2.1. Annotations linguistiques et outils TAL 40
3.2.2. Du traitement automatique des langues pour assister l'annotation d'erreurs 41
3.2.3. Exploitation automatique des productions d'apprenants 42
Chapitre 3 - Les corpus d'écrits scolaires : un domaine en pleine dynamique 45
1. Corpus scolaires : état des lieux 45
2. Nécessité de constituer des corpus d'écrits scolaires accessibles 53
3. Traitements automatiques de ces corpus 55
4. Conclusion 57
Partie 2 - Constitution du corpus *Scoledit*, ressource longitudinale d'écrits scolaires annotés .. 59
Chapitre 4 - Constitution et exploitation d'un corpus scolaire 61
1. Numérisation d'un corpus 62
1.1. Enjeux de la transcription 64
1.1.1. Comment transcrire ? 65
1.1.2. Que transcrire ? 69
1.2. Réflexion nationale autour des conventions de transcription 71
2. Enrichissement et annotation d'un corpus 73
2.1. Types d'enrichissements linguistiques 74
2.2. La transcription est-elle une annotation ? 75
2.3. Normalisation : qu'en est-il de la correction et de la standardisation ? 77
2.3.1. Statut de la correction 77
2.3.2. Quel format pour la correction ? 78
2.3.3. Correction automatique ou manuelle ? 79
2.3.4. Quelles variations annoter ? 79
2.4. Approche du corpus *Scoledit* 81

3. L'approche par comparaison : un préambule à l'exploitation 83
3.1. Genèse de l'approche 83
3.2. Exposition de la méthode 84
3.2.1. Étapes 84
3.2.2. Niveaux d'applications 85
4. Conclusion 85
Chapitre 5 - Conception et numérisation du corpus *Scoledit* 87
1. Enjeux et motivations 87
2. Recueil du corpus 89
2.1. Phases de recueil 89
2.1.1. Première phase de recueil : le recueil du projet « Lire - Écrire au CP » 89
2.1.2. Deuxième phase de recueil : le recueil du projet *Scoledit* 90
2.2. Consignes de recueil 93
2.2.1. Productions de texte 93
2.2.1.1. Consigne de production de texte en classe de CP 94
2.2.1.2. Consigne de production de texte pour les classes de CE1 à CM2 95
2.2.2. Dictées 97
2.2.2.1. Dictée du projet « Lire – Écrire au CP » (*DictéeIFÉ(1)* et *DictéeIFÉ(2)*) 97
2.2.2.2. Dictées de mots et de phrases © DEPP (*DictéeDEPP(1)* et *DictéeDEPP(2)*) 97
2.2.2.3. Texte *Le corbeau* (*DictéeCorbeau*) 98
2.3. Difficultés d'un recueil longitudinal en milieu scolaire 98
2.3.1. Complexité des démarches administratives 99
2.3.2. Complexité de l'organisation du recueil 100
2.3.3. Difficultés du suivi longitudinal 101
2.3.4. Diversité des contextes de recueil 101
3. Caractérisation et structure du corpus de textes 101
3.1. Structure du corpus 101
3.2. Caractérisation du corpus 102
4. Métadonnées 103
5. Numérisation du corpus 104
5.1. Choix de transcription pour le projet *Scoledit* 105
5.1.1. Eléments génétiques, visuels ou méta-textuels 106
5.1.2. Eléments textuels 109
5.1.3. Indications de l'incertitude du transcripteur 112
5.2. Outil de transcription 113
5.3. Processus de transcription 116
6. Diffusion du corpus 117
7. Conclusion 120

Chapitre 6 - La normalisation, une annotation déportée 121
1. Type de d'enrichissement choisi 121
2. Enjeux de la normalisation 122
3. Principes de normalisation 123
4. Conventions de normalisation 125
4.1. À l'échelle lexicale et micro-syntaxique 126
4.2. À l'échelle syntaxique 127
4.3. Marques de structuration 130
4.4. Marques d'incertitudes 131
5. Processus de normalisation 131
6. Conclusion 132
Partie 3 - Application de l'approche par comparaison : aligner transcriptions et normalisations 135
Chapitre 7 - Algorithmes d'alignement, revue des méthodes 137
1. Introduction 137
1.1. Définition des notions 137
1.2. Principe de l'approche par comparaison appliquée aux formes 139
1.3. Algorithme d'alignement, quelques explications 141
1.4. Domaines d'application des algorithmes d'alignement 142
2. État de l'art des méthodes d'alignement 143
2.1. Distance d'édition ou distance de Levenshtein 144
2.1.1. Principe général 144
2.1.2. Distance d'édition avec poids 146
2.1.3. Unités de mesure 147
2.1.4. Distance d'édition normalisée 147
2.2. Critères et indices d'alignement 147
2.2.1. Méthodes basées sur des indices graphiques 148
2.2.1.1. Mesures de distance et de similarité 148
2.2.1.2. N-grammes 148
2.2.1.3. Séquences communes 149
2.2.1.4. Règles orthographiques 149
2.2.1.5. Génération de candidats par règles 150
2.2.2. Méthodes basées sur des indices phonologiques 150
2.2.2.1. Distance d'édition avec variation de poids 150
2.2.2.2. Méthodes à partir de clés phonétiques 151
2.2.2.3. Méthodes par conversion phonétique 152
2.2.2.4. Méthodes par apprentissage automatique 153
2.2.3. Méthodes basées sur des indices stochastiques 153
2.3. Méthodes d'alignement 155
2.4. Unité d'alignement 155

2.5. Adaptabilité des méthodes 155
3. Conclusion 156
Chapitre 8 - *AliScol*, un système d'alignement pour les écrits scolaires 157
1. Composition du corpus de développement 158
2. Unité et critères d'alignement 159
2.1. Comparaison à l'aide d'indices graphiques 159
2.2. Comparaison à l'aide d'indices phonologiques 160
2.3. Comparaison à l'aide d'indices archiphonologiques 160
3. Prétraitements nécessaires 161
4. Algorithme d'alignement 166
4.1. Mesure de comparaison 166
4.2. Parcours algorithmique 169
4.2.1. Parcours séquentiel (A) 169
4.2.2. Parcours matriciel (B) 175
4.2.2.1. Calcul de la distance d'édition de Levenshtein 176
4.2.2.2. Calcul de la distance d'édition de Damereau-Levenshtein (opération d'inversion) 177
4.2.2.3. Prise en compte des phénomènes d'hyposegmentation et d'hypersegmentation 178
5. Conclusion 179
Chapitre 9 - Évaluation de l'aligneur *AliScol* 181
1. Élaboration des données de référence 181
1.1. Méthode manuelle 181
1.2. Méthode par correction 182
2. Mesures d'évaluation 182
2.1. Accord inter-annotateurs 183
2.2. Score d'évaluation 183
3. Evaluation d'*AliScol* 184
3.1. Composition du corpus de référence 185
3.2. Comparaisons des différentes versions de l'aligneur 187
3.2.1. Méthode séquentielle 187
3.2.2. Méthode matricielle 191
3.2.3. Evaluation de l'algorithme par niveau scolaire 194
3.3. Observations des erreurs de l'aligneur *AliScol* 195
4. Conclusion 196

Partie 4 - Application de l'approche par comparaison : autres exemples 197

Chapitre 10 - Comparer formes transcrites et formes normalisées, quelques enseignements 199

1. Erreurs de segmentation en mots 200

1.1. Recensement des phénomènes d'hypersegmentation et d'hyposegmentation 200
1.2. L'élision, un facteur d'hyposegmentation 202

2. Comparaison des formes au niveau graphique et phonologique 203

2.1. Définition des catégories 203
2.1.1. Les segments normés 203
2.1.2. Les segments phonologiquement normés 203
2.1.3. Les formes archiphonologiquement normée 204
2.1.4. Les segments non normés 205
2.2. Données chiffrées 206
2.2.1. Données chiffrées longitudinales 206
2.2.2. Données chiffrées par catégories syntaxiques 207
2.3. Évaluation de la fiabilité des résultats 208

3. Conclusion 210

Chapitre 11 - Appliquer la méthode par comparaison aux graphèmes 213

1. Définition opératoire du graphème 215

1.1. Définition du graphème dans la littérature 215
1.2. Définition du graphème adoptée pour l'algorithme *AliScol_Graph* 217
1.2.1. Les graphèmes 217
1.2.1.1. Application des règles de discrimination des graphèmes, le critère phonogrammique 218
1.2.1.2. Application des règles de discrimination des graphèmes, le critère morphogrammique 220
1.2.2. Les ponctuants 221
1.2.3. Les balises 221
1.2.4. Les nombres et les symboles 221

2. Algorithme développé 222

2.1. Découpage en graphèmes (prétraitements, étape2) 222
2.1.1. Appel de *LIA-PHON* 222
2.1.2. Ajustement du découpage en graphèmes 223
2.1.3. Calcul des représentations archiphonologiques 224
2.2. Alignement des graphèmes (étape 3) 224

3. Évaluation de l'algorithme 225

4. Conclusion 226

Chapitre 12 - Analyser la morphographie verbale grâce à l'approche par comparaison 227
1. Quelle théorie du verbe ? 227
2. Caractérisation des verbes 229
2.1. Répartition des temps verbaux 229
2.2. Répartition des erreurs par temps 231
3. Distinction des erreurs sur la base et sur la désinence 233
3.1. Le cas de l'imparfait 234
3.2. Le cas du passé simple 235
4. Conclusion 236
Conclusion et perspectives 239
1. Travaux réalisés au cours de la thèse 239
2. Apports des travaux réalisés 240
3. Perspectives d'amélioration des ressources développées 240
3.1. Le corpus 240
3.2. Les outils développés 241
3.3. Développement d'un outil de normalisation 242
4. Perspectives d'exploitation linguistique et didactique des ressources développées.. 242
Bibliographie 245
Liste des tableaux 246
Liste des figures 250
Annexes 254
Annexe 1 : Les différents corpus utilisés 256
Étude 1 : Alignement des formes (Chapitre 8, 9 et 10) 256
1. Corpus de développement 256
2. Corpus d'évaluation 257
3. Corpus de travail 257
Étude 2 : Alignement des graphèmes (Chapitre 11) 258
Étude 3 : Etude de la morphographie verbale (Chapitre 12) 258
Annexe 2 : Conventions nationales de constitution de corpus scolaires 260
Annexe 3 : Livrets de recueil 268
1. Livret de consigne en classe de CP 268
2. Support d'écriture, classe de CP et de CE1 270
3. Livret de consigne en classe de CE1 271

4. Support d'écriture, classe de CE2, CM1 et CM2 273
5. Consignes CE2-CM2 274
Annexe 4 : Ensemble des productions recueillies pour un élève (élève 96) 276
Annexe 5 : Guide de transcription 282
Transcription des éléments génétiques 282
Transcription des éléments extratextuels 283
Transcription des éléments textuels 284
Transcription de l'incertitude du transcripteur ou de la transcriptrice 286
Annexe 6 : Guide de normalisation 288
1. Segmentation en mots, orthographe et accords 288
1.1. Segmentation en mots 288
1.2. Orthographe 289
1.3. Les accords 289
1.4. La morphologie verbale 289
2. Cohésion et utilisation des temps 290
3. Le plan lexical 290
3.1. Omission d'un mot 290
3.2. Répétition ou ajout d'un mot 291
3.3. Choix lexicaux 291
4. Les constructions issues de l'oral 292
5. Syntaxe et ponctuation 292
5.1. Organisation des mots dans la proposition 292
5.2. La ponctuation 292
5.3. La segmentation du texte 293
5.4. 5.4. Le cas de la virgule 295
5.5. La balise <s/> 296
6. Traitement des dialogues 297
7. Notations spécifiques 298
Annexe 7 : French *TreeTagger* Part-of-Speech Tags 300
Annexe 8 : Table des correspondances des formats de transcription phonologique 302
Annexe 9 : Liste des correspondances graphophonémiques établie par J. Riou et R. Goigoux (2017) 304
Annexe 10 : Protocole d'alignement en graphèmes 306
Unités d'alignement 306
A. Les graphèmes 306
B. Les ponctuants 308

C. Les balises 309
D. Les espaces 311

Résumé 312

Mots clés 312

Abstract 314

Keywords 314

Glossaire

Corpus

Un **corpus** est un ensemble de données rassemblées ou collectées selon un critère linguistique défini, ou dans le cadre d'une situation de communication unifiée, stockées électroniquement et de manière exploitable informatiquement sous un format texte, pouvant être agrémentées de données supplémentaires (annotations, version normalisée, représentation phonologique, etc.) (cf. Chapitre 1).

Nous appelons **corpus longitudinal** le sous-corpus qui ne contient que les productions des élèves présents toutes les années du CP au CM1. Il est composé de 1 492 textes de niveaux différents (CP – CM1) produits par 373 élèves.

Nous appelons **corpus de développement** les sous-corpus utilisés pour développer les outils de traitement automatique.

Nous appelons **corpus d'évaluation** les sous-corpus utilisés pour évaluer les outils de traitement automatique.

Nous appelons **corpus de travail** les sous-corpus utilisés pour mener des études linguistiques.

Le descriptif des différents corpus utilisés au cours de ce travail de thèse est donné en annexe (Annexe 1).

Transcription et normalisation

La **transcription** d'une production est une version numérisée de cette production. Elle vise la reproduction de l'ensemble des caractères textuels de celle-ci ainsi que certaines caractéristiques visuelles (cf. Chapitre 5).

La **normalisation** d'une production est une version numérisée et normalisée de cette production. Elle vise à fournir une version plus proche de la norme linguistique pour laquelle les outils informatiques ont été conçus (cf. Chapitre 6).

Conventions d'écriture : Les « éléments transcrits » sont donnés entre guillemets français, tandis que les *éléments normalisés* sont donnés en italique.

Par exemple :

- Production entière ou production partielle :

« sa mamant il dormer le petit chat il a marcher » (70, CP)
[normalisation] Sa maman elle dormait <segmentation/> le petit chat il a marché.

- Forme ou groupe de formes : « itonb » (*il tombe*, 1138 - CP)
- Lettre, graphème ou groupe de lettres : « on » (*om*)

Mot, token, forme et segment

Ces termes renvoient à des notions proches mais non identiques.

Le terme **forme** est utilisé préférentiellement pour désigner les formes normalisées contenues dans la version normalisée des productions. Par exemple, *Il tombe* contient deux formes : *il* et *tombe*.

Le terme **segment** est utilisé pour désigner les segments produits par les scripteurs et reproduits dans la version transcrite des productions. Par exemple*, «* itonb » est un seul segment, bien qu'il renvoie aux deux formes *il* et *tombe*.

Le terme **token** est un terme utilisé dans le domaine du traitement automatique des langues, il est réservé à la conception des outils de traitement automatique. Il désigne toute séquence de caractères encadrée par des espacements ou par des signes de ponctuation. Il peut renvoyer indifféremment à une forme normalisée ou à un segment transcrit.

Le terme **mot** est un terme particulièrement ambigu que nous évitons d'utiliser, dans la mesure du possible, exception faite lorsque son utilisation permet une meilleure compréhension.

Niveau de scolarité

Les niveaux de scolarité sont désignés selon l'usage français. Les correspondances classe – âge moyen des élèves – année de scolarité sont données dans le tableau ci-dessous :

Cycle	*Ecole française*	*Classe française*	*Correspondance avec la plupart des systèmes éducatifs*	*Âge moyen des élèves en fin d'année*
Cycle 2	primaire	CP	1ère année primaire	7 ans
		CE1	2e année primaire	8 ans
		CE2	3e année primaire	9 ans
Cycle 3		CM1	4e année primaire	10 ans
		CM2	5e année primaire	11 ans
	secondaire	6e	6e année primaire	12 ans

Tableau 1 : Correspondances classe – âge moyen des élèves – année de scolarité

Introduction

1. Contexte 1
2. Le projet *Scoledit* 2
3. Inscription dans un dialogue multi-équipe 3
4. Problématique 4
5. Plan de la thèse 5

1. Contexte

Que ce soit lors de l'écriture d'un mail professionnel, de l'envoi d'un SMS de félicitation ou de la rédaction d'un formulaire de réclamation, l'écrit est de plus en plus présent dans nos sociétés. L'apprentissage de l'écriture et de la production d'écrit représente donc un enjeu majeur pour l'école d'aujourd'hui et un nombre grandissant de travaux de recherche y est consacré.

Parallèlement, alors qu'on assiste depuis plusieurs décennies à un développement des études linguistiques sur corpus, les écrits des apprenants de langue maternelle étaient jusqu'à récemment très peu représentés dans les corpus de grande taille constitués ou en cours de constitution, et ce malgré l'accroissement des recherches sur l'écriture et les processus d'écriture. Observer les textes d'apprenants permettrait pourtant de suivre le processus évolutif à l'œuvre dans l'apprentissage de l'écriture. Écrire un texte est en effet une activité complexe et apprendre à écrire des textes est un processus qui s'inscrit dans la durée et qui est loin d'être achevé en fin de scolarité primaire (Fayol, 2013). L'élaboration de grands corpus scolaires permettrait d'appuyer ces observations. En effet, dès 2011 M.-L. Elalouf (2011, p. 67) entrevoit qu'il est temps d'opérer un renouvellement des pratiques de recherche et un changement d'échelle comparable à celui qui s'est produit en sciences du langage suite à l'emploi de grands corpus.

Face à ce besoin émergent depuis une dizaine d'années divers projets poursuivant l'objectif de constituer des corpus d'écrits scolaires (Elalouf, 2005 ; Gunnarsson-Largy & Auriac-Slusarczyk, 2013 ; Garcia-Debanc & Bonnemaison, 2014 ; Boré & Elalouf, 2017 ; Doquet, Enoiu, Fleury, & Maziotti, 2017). Mais ces corpus, pour la plupart encore en cours d'élaboration, sont souvent encore trop restreints ou peu accessibles. De plus, peu s'intéressent à la notion de progressivité et particulièrement dès la première année de l'école primaire.

Par ailleurs, ce changement d'échelle ne pourra se faire sans des outils numériques adaptés. Il y a donc une nécessité à employer des méthodes issues du traitement automatique des langues (désormais TAL), alors même que l'absence de travaux sur les écrits scolaires dans ce domaine est criant.

2. Le projet *Scoledit*

De ces deux constats, à savoir l'absence de grands corpus longitudinaux accessibles et exploitables et le besoin de méthodes issues du TAL pour accompagner l'élaboration et l'exploitation de ces corpus, est né en 2014 le projet *Scoledit*, initié par Claude Ponton, Catherine Brissaud et Corinne Totereau au sein du laboratoire *Lidilem* de l'Université Grenoble Alpes.

Ce projet vise l'élaboration d'un corpus longitudinal de productions scolaires recueillies du CP au CM2 (6 - 12 ans), ainsi que son accessibilité en ligne, son exploitation et son usage à des fins linguistiques et didactiques. Pour ce faire, des dictées et des productions de textes sont recueillies à différents moments de la scolarité élémentaire de plus de 1 000 élèves. L'ensemble du corpus rassemble plusieurs milliers de productions permettant l'étude longitudinale des procédés d'écriture de différents niveaux (orthographe, syntaxe, ponctuation, cohérence textuelle).

L'enjeu d'un tel projet est multiple. Il s'agit, dans un premier temps, de permettre une description linguistique des écrits d'apprenants à l'école primaire à la fois en synchronie et en diachronie, du CP au CM2. Pour ce faire, il est important que le recueil soit véritablement longitudinal, c'est-à-dire réalisé à différents moments auprès des mêmes élèves et avec une consigne proche, et non uniquement transversal, c'est-à-dire rassemblant des textes produits à différents moments de la scolarité, sans qu'ils soient forcément produits par les mêmes élèves. Un tel corpus devrait apporter une connaissance plus fine des phénomènes d'acquisition de l'écriture comme, par exemple, l'évolution des compétences orthographiques et syntaxiques, l'évolution de l'acquisition des conjugaisons, celle de la cohérence des temps ou encore l'évolution de l'usage de la ponctuation. Dans un second temps, le projet *Scoledit* devrait permettre d'élaborer des séquences et des dispositifs didactiques à destination des enseignants.

Nous nous inscrivons dans une perspective similaire à celle de C. Boré et M.-L. Elalouf (2017, p. 32) qui, tout comme C. Bonnet, C. Corblin et M.-L. Elalouf (1998, citées par Boré & Elalouf, 2017) et C. Fabre-Cols (2000), font l'hypothèse « que la lecture d'un nombre important de textes d'élèves répondant à une même consigne permet au formateur – et à travers lui à

l'enseignant – de se constituer une culture de ces textes ». Mettre les textes d'apprenants en regard les uns avec les autres devrait ainsi permettre de connaitre de manière plus juste les compétences que peuvent acquérir les élèves à un niveau donné.

3. Inscription dans un dialogue multi-équipe

Le projet *Scoledit* est né au sein du laboratoire *Lidilem* et s'inscrit pleinement dans la continuité des travaux qui y ont cours. D'abord portés sur la didactique des langues, à l'initiative de L. Dabène, les travaux développés au sein du *Lidilem* se sont progressivement étendus à la linguistique, y compris de l'écrit avec les aspects orthographiques sous l'impulsion de V. Lucci et d'A. Millet. Progressivement, des études faisant le lien entre orthographe et acquisition / apprentissage ont émergé notamment grâce aux travaux de J.-P. Chevrot puis de C. Brissaud. Parallèlement, les membres du *Lidilem* témoignent d'un intérêt constant pour les écrits des apprenants, adultes comme enfants (travaux de C. Barré-De Miniac, F. Boch., C. Fabre-Cols, C. Frier, M.-C. Guernier, F. Rinck, notamment).

Par ailleurs, il existe au sein du *Lidilem* une certaine habitude des études sur corpus et des travaux de constitutions de corpus (travaux de G. Antoniadis, F. Grossmann, M.-P. Jacques, O. Kraif, T. Lebarbé, C. Ponton, A. Tutin notamment) et sans doute sous l'impulsion de la petite communauté de TAListes présente au *Lidilem*. Cette communauté a produit, spécifiquement, un certain nombre de travaux visant à assister la constitution de corpus à l'aide d'outils automatiques et plus particulièrement des corpus peu normés (corpus d'apprenants, corpus de tweets, corpus de SMS, etc.). Récemment, une habilitation à diriger des recherches a été soutenue par M.-P. Jacques au sein du *Lidilem* autour des corpus écrits, de leurs outils et des analyses qui s'y rapportent (Jacques, 2017). En 2018, une thèse a également été soutenue par E. Kogkitsidou sur le thème de la normalisation d'un corpus de SMS. Tous ces travaux, tant didactiques et linguistiques qu'informatiques, ont nourri la réflexion qui a mené à l'émergence du projet *Scoledit*.

Enfin, ce projet s'inscrit dans un contexte d'émergence des corpus d'écrits scolaires et a donc pu se nourrir et se nourrit toujours de nombreuses collaborations avec différentes équipes de recherche. La phase d'élaboration de la transcription a particulièrement été guidée par les interactions avec plusieurs équipes de recherches françaises qui, pour la plupart, se rejoignent désormais au sein du projet *E-CALM*[1] (financé par l'Agence nationale de la recherche et

[1] Écritures scolaires : Corpus, Analyses Linguistiques, Modélisations didactiques, http://e-calm.huma-num.fr/ [consulté le 21/08/2019].

coordonné par Claire Doquet). La normalisation proposée dans ce travail a beaucoup évolué et de façon significative lors des échanges générés au moment du séjour de recherche de Catherine Brissaud, membre du projet, accueillie par Lilia Teruggi et son équipe au sein du Dipartimento di Scienze Umane per la Formazione "Riccardo Massa" de l'université de Milan. Cette équipe développe actuellement un projet italophone similaire au projet *Scoledit*. Enfin, le travail réalisé sur l'étude des graphèmes a notamment émergé grâce à une collaboration avec les centres de recherche belges CENTAL et MiiL lors de notre séjour de recherche encadré par Cédrick Fairon et Louise-Amélie Cougnon.

4. Problématique

Un tel corpus ne révèle tout son intérêt que s'il est accompagné d'outils permettant son exploitation. Nous faisons l'hypothèse que les méthodes de traitement automatique des langues peuvent constituer une aide à cette exploitation.

Les outils de traitement des corpus écrits existant à l'heure actuelle ont majoritairement été développés pour des corpus de scripteurs experts et ne conviennent pas au traitement des écrits scolaires, très différents des types d'écrits pour lesquels ils ont été conçus.

Les spécificités des écrits scolaires, et particulièrement leur écart à la norme, représentent un véritable défi lorsqu'on cherche à élaborer des outils de traitement spécifiques. Ceci constitue sans doute une des raisons pour lesquelles les écrits scolaires ont encore fait l'objet de peu de travaux dans le domaine du TAL. Cependant, nous pensons que les outils et méthodes de ce domaine peuvent présenter un véritable apport pour les corpus scolaires, tels que le corpus *Scoledit*. C'est ce questionnement, l'étude de l'apport du traitement automatique des langues à la constitution et à l'exploration de corpus scolaires, qui nous a guidée tout au long de la réalisation de ce travail de thèse, au cours duquel deux questions principales ont été étudiées : comment constituer un corpus numérique longitudinal d'écrits scolaires ? comment outiller son exploitation et quel apport peut représenter le TAL dans ce processus ?

Ce travail représente donc un enjeu pour la linguistique et la didactique puisqu'il doit permettre la mise à disposition de la communauté scientifique et enseignante d'un corpus longitudinal d'écrits scolaires et d'outils d'exploration. Mais il représente également un enjeu pour le traitement automatique des langues puisqu'il doit permettre d'explorer un type d'écrits encore peu exploré dans ce domaine.

5. Plan de la thèse

L'exposé de cette thèse se divise en quatre parties. La première partie (Chapitres 1 à 3) s'attache à présenter le contexte général dans lequel s'inscrit cette thèse. Le chapitre 1 revient sur la définition de la notion de corpus et sur la linguistique de corpus. Le chapitre 2 présente le domaine du traitement automatique des langues et revient sur les apports de ce domaine aux corpus et à la linguistique de corpus. Le chapitre 3 présente un état des lieux du champ des corpus d'écrits scolaires, et plus particulièrement des corpus d'écrits scolaires francophones.

La deuxième partie (Chapitres 4 à 6) présente plus en détail le corpus *Scoledit* et la méthodologie adoptée pour l'élaboration de ce corpus et pour la constitution d'outils d'exploration de celui-ci. Le chapitre 4 présente la méthodologie d'un point de vue général, tandis que le chapitre 5 présente les choix de numérisation effectués pour le corpus *Scoledit* et que le chapitre 6 présente les choix d'annotation et d'enrichissements adoptés pour ce corpus.

La troisième partie (Chapitres 7 à 9) est consacrée à la présentation et à l'évaluation d'un outil de traitement automatique conçu pour le corpus *Scoledit*, un aligneur de formes appelé *AliScol*. Le chapitre 7 est consacré à un état de l'art des méthodes d'alignement. La conception et la réalisation de l'outil *AliScol* sont présentées au chapitre 8, tandis que son évaluation est présentée dans le chapitre 9.

La quatrième et dernière partie (Chapitres 10 à 12) propose d'explorer le corpus *Scoledit* à différents niveaux linguistiques grâce à l'usage d'outils de traitement automatique. Le chapitre 10 traite de l'étude des spécificités des écrits scolaires en terme d'écarts à la norme. Le chapitre 11 présente l'outil *AliScol_Graph*, qui permet l'exploration des graphèmes dans les écrits scolaires. Enfin, le chapitre 12 développe une étude sur la morphographie verbale, permise par l'usage de méthodes de traitement automatique des langues.

Une conclusion clôt l'ensemble et présente quelques perspectives pour poursuivre ce travail.

Partie 1 - Linguistique de corpus, TAL et corpus scolaires

Chapitre 1 - Corpus et linguistique de corpus 9

Chapitre 2 - Traitement automatique des langues et apport à l'exploitation des corpus 22

Chapitre 3 - Les corpus d'écrits scolaires : un domaine en pleine dynamique 45

Chapitre 1 - Corpus et linguistique de corpus

1. Qu'est-ce qu'un corpus ? 9
2. Linguistique de corpus, rappels historiques 15
3. La constitution de la linguistique de corpus en tant que discipline 18
4. Conclusion 22

1. Qu'est-ce qu'un corpus ?

Les corpus sont des ressources qui ont évolué et leur usage s'est largement répandu, de manière conjointe avec le développement de l'informatique, facilitant leur développement et leur diffusion. De ce fait, la notion de *corpus* a également évolué et, à l'heure actuelle, elle renvoie parfois à des réalités bien distinctes.

Commençons par définir un corpus comme un « ensemble de données linguistiques », que ces données soient écrites, orales, gestuelles, etc. Prenons quelques exemples qui, à l'exception du corpus *Scientext*[2], sont tous disponibles sur la plateforme *Ortolang*[3]. Un corpus peut rassembler des éléments aussi divers que :

- 1561 cartes et lettres postales échangées pendant la première guerre mondiale entre des soldats et leurs familles (*Corpus14* ; Steuckardt, 2014) ;
- Plus de 88 000 SMS authentiques en français collectés à Montpellier (Corpus *88milSMS* ; Panckhurst, Détrie, Verine, Lopez, Moïse, & Roche, 2014) ;
- 520 commentaires évaluatifs de relecteurs pour un colloque de jeunes chercheurs en sciences du langage (extrait du corpus *Scientext* ; Boch, Grossmann, & Rinck, 2007) ;
- 8 ouvrages de littérature jeunesse en français et en anglais alignés avec leur traduction dans l'autre langue du corpus au niveau des chapitres, paragraphes et phrases (Corpus *ParCoGLiJe* ; Stosic, Marjanović, & Miletic, 2018) ;

[2] Disponible sur la plateforme ScienQuest. https://corpora.aiakide.net/scientext20/ [consulté le 02/10/2019].
[3] Outils et Ressources pour un Traitement Optimisé de la LANGue, https://www.ortolang.fr/ [consulté le 09/08/2019]. Cette plateforme mutualise de nombreux outils et ressources pour la langue, dont de nombreux corpus.

- Des enregistrements de populations « jeunes » en région parisienne dans le but de documenter certains faits langagiers dans les banlieues et les zones urbaines sensibles (Corpus *Français Parisien Multiculturel* ; Gadet, 2017) ;
- Des dialogues de 4 personnes autour d'un jeu enregistrés en chambre sourde avec micro-casque (Corpus *jeux*[4]) ;
- Des captures et des vidéos de mouvements en langue des signes collectées dans l'objectif de réaliser des études pluridisciplinaires (Corpus *MOCAP1 ;* Benchiheub, Berret, & Braffort, 2016) ;
- Des séquences vidéo enregistrées lors de la description d'un dessin dans le but d'analyser le regard lors de cette tâche (Corpus *Analyse du regard lors de la description d'un dessin*[5]).

Mais cette définition seule n'est pas suffisante : pour être qualifié de corpus, un ensemble de données linguistiques doit correspondre à certains critères qui varient selon les domaines, les projets, les décennies et leurs innovations technologiques.

1.1. Critère linguistique

Le premier critère est le **critère linguistique**. Un corpus, pour être reconnu comme tel en linguistique, doit être constitué et organisé selon un ou plusieurs critères linguistiques explicites. On retrouve cette condition dans de nombreuses définitions :

> *Corpus: a subset of an ETL [electronic text library], built according to explicit design criteria for a specific purpose, eg the Corpus Révolutionnaire (Bibliothèque Beaubourg, Paris), the Cobuild Corpus, the Longman/Lancaster corpus, the Oxford Pilot corpus.* (Atkins, Clear, & Ostler, 1992, p. 1)

> *A* corpus *is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.* (Sinclair, 1996)

> *Un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications*. (Rastier, 2005, p. 32)

[4] Corpus jeux | ORTOLANG. (s. d.). Consulté 16 septembre 2019, à l'adresse https://www.ortolang.fr/market/corpora/sldr000773

[5] Analyse du regard lors de la description d'un dessin | ORTOLANG. (s. d.). Consulté 16 septembre 2019, à l'adresse https://www.ortolang.fr/market/corpora/sldr000837

Certains chercheurs plaident pour le caractère obligatoire du critère linguistique et excluent donc de l'appellation *corpus* les nombreuses bases textuelles, numérisées ou non, constituées par accumulation de ressources ou par extraction d'une seule source mais sans fondement linguistique ou sans visée de recherche précise.

Pour exemple, la base *Frantext*[6], développée à l'ATILF depuis les années 1970, regroupe des textes français principalement littéraires et philosophiques et dans une moindre mesure scientifiques et techniques. Ces textes n'ont pas été rassemblés selon un critère linguistique ou une hypothèse de recherche précise, mais dans le but de constituer des exemples pour le *Trésor de la Langue Française*[7]. Cet ensemble de textes n'est donc pas considéré comme un corpus mais comme une base textuelle.

Cependant, cette base permet d'extraire des textes sélectionnés dans le but de constituer un corpus, ce qui incitera B. Habert (2000) à appeler les bases textuelles similaires à *Frantext* des « réservoirs à corpus ».

Néanmoins, en certaines occasions, comme dans certains lieux de publication, une définition plus large de *corpus* peut être admise, regroupant également les bases textuelles sous l'appellation *corpus* afin de fournir un espace de médiatisation à ces bases. Ce fut le cas en 2002 lors des deuxièmes journées de la linguistique de corpus, où G. Williams (2005, p. 14) regroupa sous l'appellation *corpus* l'ensemble des ressources agglomérant des « textes stockés électroniquement sous un format texte ».

B. Habert (2000) pointe également le fait que les études sur le genre et la variation sont telles que souvent les caractéristiques linguistiques des textes et discours rassemblés au sein d'un corpus ne sont pas encore bien connues, ce qui empêche de connaitre la valeur linguistique de l'échantillon de données linguistiques sélectionnées. Notons par ailleurs que le recours à la notion d'*échantillon* peut sembler contestable en linguistique dans la mesure où un échantillon se définit souvent par rapport à un ensemble délimité. Or, il n'est pas possible de déterminer l'ensemble fini des données langagières d'une langue donnée.

B. Habert (2000) soutient donc, tout comme D. Biber (1993), l'élaboration d'une typologie des productions langagières basée sur des critères linguistiques dans le but de constituer des corpus homogènes, au regard de ces critères. Les typologies linguistiques et les outils de

[6] FRANTEXT (1992). *Autour d'une base de données textuelles ; témoignages d'utilisateurs et voies nouvelles*. Paris : Didier Érudition.
[7] Dendien J. (1991). *Access to information in a textual database: access functions and optimal indexes*. Oxford : Clarendon press.

typologie étant encore insuffisants, il propose de s'appuyer sur des critères extralinguistiques pour constituer des corpus (conditions de production, fonction communicative, thème ou domaine, genre et registre, etc.). Sa définition de *corpus* (ci-après) inclut donc à la fois les critères linguistiques et extra-linguistiques.

> *Collection de données langagières qui sont sélectionnées et organisées explicitement selon des critères linguistiques et extra-linguistiques explicites pour servir d'échantillon d'*emplois *déterminés d'une langue* (Habert, 2000, p. 13)

Cette absence de critère linguistique se retrouve encore dans certains corpus d'écrits scolaires, où les textes sont plutôt recueillis selon un critère extralinguistique, à savoir le contexte de recueil – par exemple en milieu scolaire –, qu'un véritable critère linguistique.

1.2. Critère d'authenticité

On associe souvent à la notion de corpus la notion de *données authentiques* ou *données attestées*, par opposition aux données obtenues par introspection, c'est-à-dire des exemples linguistiques inventés par des linguistes ou des locuteurs natifs en se basant sur l'intuition que possède tout locuteur vis-à-vis de sa langue. Certains auteurs comme T. McEnery (2003) intègrent cette notion à leur définition de corpus.

> *A corpus […] is simply described as a large body of linguistic evidence typically composed of attested language use* (McEnery, 2003, p. 449)

Selon les corpus, une distinction peut être faite entre les données produites en situation de communication non dépendantes d'une recherche et les données provoquées à des fins de recherche. Par exemple, dans le cadre de corpus de données issues du web (tweets, messages de forum, etc.), les données ont été produites de manière totalement extérieure à la recherche, sans même que les locuteurs aient conscience que leurs messages puissent être utilisés à des fins de recherche. En revanche, dans le cadre d'un corpus enregistré en chambre sourde ou avec des capteurs sensoriels, les locuteurs ne peuvent ignorer la présence d'appareils d'enregistrement à des fins scientifiques. Parfois même, les dialogues, les discours ou les écrits sont provoqués par et pour la recherche.

1.3. Critère de représentativité

Un autre critère important associé au corpus est le critère de représentativité. Beaucoup de corpus ont vocation à être représentatif d'un ensemble plus vaste de données langagières. Il peut s'agir d'une langue donnée, ces corpus sont alors appelés *corpus de référence* ou *corpus équilibrés*, ou d'ensembles plus restreints, comme un genre de textes ou une situation de communication spécifique, on parle alors de *corpus spécifiques*.

E. Tognini-Bonelli, qui envisage davantage les corpus comme des ressources pour l'apprentissage des langues et qui manipule principalement des corpus dits de référence, intègre cette notion à sa définition de corpus.

> *collection of texts assumed to be representative of a given language put together so that it can be used for linguistic analysis* (Tognini-Bonelli, 2001, p.2)

En raison de leur grande ambition, les travaux de constitution de corpus de référence sont souvent longs et peu fréquents. Un des premiers travaux conséquents dans ce domaine est le projet d'élaboration du *Brown Corpus* et du *British National Corpus* (ou *Brown University Standard Corpus of Present-Day American English*), dans les années 1960 (Kucera & Francis, 1967). Ce corpus anglophone de quelques millions de mots rassemble plusieurs centaines de textes de 15 genres différents (presse, scientifique, religion, par exemple). Deux décennies plus tard, le *British National Corpus* (*BNC*), corpus de plusieurs centaines de millions de mots, est élaboré au sein de l'Oxford University Press. Il rassemble des textes d'une grande variété de genres différents et vise la représentativité de l'anglais britannique de la fin du XXe siècle.

Cependant tous les acteurs manipulant les corpus ne sont pas en accord avec cette généralisation des résultats. En effet, certains considèrent que les corpus ne peuvent être représentatifs que de l'étude pour laquelle ils ont été construits. Puisqu'il n'est pas possible de connaitre l'ensemble plus vaste duquel il serait l'échantillon, à plus forte raison lorsqu'on s'intéresse à la « langue générale » dans le cadre des corpus dits équilibrés (Péry-Woodley, 1995), les résultats produits à partir d'un corpus ne peuvent s'appliquer qu'à celui-ci (Condamines, 2005).

1.4. Degré de numérisation

Constituer des corpus n'est pas une pratique nouvelle. Pour des raisons matérielles évidentes, les premiers corpus n'étaient pas numérisés, leur capacité à être partagés et répliqués était donc limitée. Au fur et à mesure des avancées technologiques, les corpus sont de plus en plus informatisés, permettant ainsi une diffusion à grande échelle de certains corpus.

Dans les années 1960, R. Quirk (1968) et ses collègues élaborent le *Survey of English Usage* (*SEU*), corpus d'environ un million de mots, considéré alors comme le premier grand corpus. Mais ce corpus n'est pas informatisé : pour le consulter il est donc nécessaire de se déplacer à l'*University College London*. Quelques années plus tard est élaboré le *Brown Corpus* que nous avons mentionné à la section précédente (1.3) et que d'aucuns considèrent comme une des formes de l'informatisation du *SEU* (Léon, 2008).

Au début du développement de l'ordinateur personnel, et par là même de l'informatisation des corpus, certains auteurs (McEnery & Wilson, 2001, entre autres) distinguaient les corpus informatisés (« machine-readable ») des autres corpus, admettant la possibilité de corpus non informatisés. En revanche, d'autres autres excluent de la définition de *corpus*, les ensembles linguistiques non informatisés, comme nous avons pu le voir précédemment dans la définition proposée par Atkins et ses collègues en 1992 (1.1).

Grâce à l'accès toujours plus important à des outils informatiques et au développement de ces outils, les corpus manipulables informatiquement sont désormais largement majoritaires. Mais aujourd'hui encore on trouve des corpus non informatisés ou non manipulables via les outils informatiques. C'est notamment le cas dans le domaine des corpus oraux où pour certains corpus les enregistrements audios sont disponibles mais aucune transcription ne permet de les exploiter informatiquement. C'est le cas également pour les corpus multimodaux, où la transcription des vidéos et autres supports d'enregistrement peut s'avérer bien souvent non triviale. Mais les corpus écrits ne sont pas exempts de ce problème, notamment dans le domaine des corpus scolaires où pour certains corpus il est possible d'accéder aux scans des productions mais aucune transcription n'est associée (cf. Chapitre 3 pour des exemples).

Enfin, puisqu'il est question d'informatisation des corpus, mentionnons que certains, notamment des spécialistes de TAL mais également des linguistes de corpus (Habert, Nazarenko, & Salem, 1997 ; McEnery, 2003), considèrent certaines annotations (morphosyntaxiques [8] ou syntaxiques le plus souvent) comme incluses dans le travail de constitution de corpus, comme le mentionne la définition donnée par F. Rastier (2005, voir 1.1). Ces annotations sont essentielles pour entreprendre ensuite des traitements informatiques plus élaborés.

Dans le cadre de ce projet, nous considèrerons comme corpus un ensemble de données rassemblées ou collectées selon un critère linguistique défini, ou dans le cadre d'une situation de communication unifiée, stocké électroniquement et de manière exploitable informatiquement sous un format texte, pouvant être agrémenté de données supplémentaires (annotations, version normée, représentation phonologique, etc.).

[8] L'annotation morphosyntaxique permet d'associer à chaque mot ou forme du corpus certaines informations morphosyntaxiques comme la classe grammaticale (également appelée partie du discours), le genre, le nombre, le temps verbal, etc. Le plus souvent, est associé à cette étape le processus de lemmatisation qui permet d'associer à chaque mot ou forme son lemme, c'est-à-dire sa forme non fléchie (l'entrée du dictionnaire la plupart du temps). L'annotation syntaxique permet d'associer à chaque phrase ou groupes de mots une structure syntaxique.

2. Linguistique de corpus, rappels historiques

Le projet que nous développons s'inscrit dans le champ de ce que nous appelons communément *linguistique de corpus*. Au regard du nombre de journées d'études, revues, colloques et autres publications qui ont trait à ce champ d'études, la linguistique de corpus est considérée comme une discipline des sciences du langage. Mais cette considération est récente et parfois encore controversée.

2.1. Premières approches sur corpus

Les années 1980, puis les années 1990, en France, virent un véritable engouement des linguistes pour les corpus (Habert *et al.*, 1997 ; El Kaladi, 2007), mais cet intérêt n'est pas nouveau. En effet, T. McEnery et A. Wilson (2001) mentionnent de nombreux projets basés sur des approches sur corpus dès la fin du XIXe siècle. Ces projets ont émergé de domaines aussi variés que l'apprentissage des langues (Fries & Traver, 1940 ; Bongers, 1947, cités par McEnery & Wilson, 2001), l'étude de l'acquisition du langage (Preyer, 1889 ; Stern, 1924 ; Ingram, 1978, cités par McEnery & Wilson, 2001) ou encore l'élaboration de conventions orthographiques (Käding, 1897, cité par McEnery & Wilson, 2001). G. Williams (2006) mentionne également les travaux lexicographiques français et britanniques (Lorge, 1949 ; Fries, 1952 ; Gougenheim, Michea, Rivenc, & Sauvageot, 1956, cités par Williams, 2006). Dans les années 1950, le courant structuraliste dominait aux Etats-Unis (Anthony M McEnery & Wilson, 2001). Ce courant, qui s'appuyait sur la théorie mathématique de la communication de Shannon, avait également adopté une approche sur corpus afin de rechercher les exemples nécessaires à ses travaux (Cori et al., 2008).

2.2. Linguistique introspective et apparition des premiers grands corpus

Peu de temps après, un nouveau courant linguistique est apparu aux Etats-Unis, à rebours de cette méthodologie basée sur les corpus. En effet, N. Chomsky a développé dans les années 1950 et 1970 la théorie de la grammaire générative, basée sur la méthode introspective. L'objectif des tenants de la théorie chomskyenne était alors de chercher les universaux des langues en travaillant sur la notion de « compétence » du locuteur, par opposition à la « performance » que l'on trouve dans les données attestées. Cette notion de compétence renvoie directement à la notion de grammaticalité et de possible de langue. Les exemples utilisés pour élaborer ces grammaires sont souvent construits par les linguistes, en se basant sur leur propre jugement de la grammaticalité ou sur le jugement des locuteurs de la langue :

c'est ce qu'on appelle l'introspection. L'essor de cette théorie linguistique marque alors le déclin des approches sur corpus aux Etats-Unis pendant deux décennies, durant les années 1960 et 1970.

Au même moment, un autre courant se développe en Grande-Bretagne, plus critique envers les théories génératives, celui des linguistes descriptivistes qui cherchent à décrire la langue *telle qu'elle est utilisée* dans des situations réelles de communication. Ce sont les linguistes issus de ce courant qui sont à l'origine des premiers travaux visant l'élaboration de corpus de grande taille. Mentionnons ainsi plusieurs projets précurseurs. Rappelons le projet *Survey of English Usage* (SEU), mené par le linguiste britannique R. Quirk et qui rassemble un million de mots (Quirk, 1968). À peu près au même moment, en France, émerge l'idée de *Frantext*, une base textuelle de français moderne comportant quatre-vingt millions d'occurrences. Quelques années plus tard, le *Brown Corpus* est élaboré à l'université américaine Brown par H. Kučera et N. Francis en 1963 (Kucera & Francis, 1967). Ce dernier se base sur une répartition équilibrée des textes qu'il contient initialement prévue pour le SEU, R. Quirk ayant en partie contribué à son élaboration.

Comme nous l'avons vu, le *Brown Corpus* contenant plusieurs millions de mots il est souvent considéré comme le premier corpus informatisé de grande taille (Léon, 2008), mais il marque aussi le début des corpus annotés. En effet, il est enrichi manuellement d'informations morphosyntaxiques. De même, apparaissent les premiers travaux pour l'ajout automatique d'informations morphosyntaxiques (Garside, Leech, and Sampson 1987, cités par McEnery, 2003).

2.3. Expansion des corpus et enrichissement des corpus

Dans les années 1990, grâce à l'utilisation massive de l'informatique et l'amélioration des techniques et outils associés, les corpus grossissent encore et atteignent alors quelques centaines de millions de mots. C'est le cas du *British National Corpus* et du *Bank of English*, corpus d'anglais britannique (Church & Mercer, 1993). Alors que les corpus élaborés dans les années 1970 et 1980 étaient conçus comme des corpus échantillonnés, certains des corpus élaborés dans la décennie suivante sont désormais conçus pour être des corpus de référence (Cori *et al.*, 2008). Deux positions s'opposent alors lors de la constitution des corpus.

Les tenants de la constitution de corpus de référence ou de corpus plus massifs suivent le mantra alors en vigueur *more data is better data*[9] qui visent une couverture la plus large

[9] Traduit par la formule *gros, c'est beau* par J. Habert (2000).

possible (Church & Mercer, 1993, cités par Péry-Woodley, 1995). Ils partent du double postulat que, d'une part, il n'est pas possible de cerner précisément les caractéristiques des différentes productions langagières ; d'autre part, qu'une augmentation continue de la masse de données amène sans cesse à une plus grande représentativité. Les corpus de référence visent donc à être suffisamment grands et à représenter le plus grand nombre de variétés possibles de la langue, de manière à pouvoir servir de base à l'élaboration de grammaires, de dictionnaires ou d'autres matériels linguistiques de référence (Sinclair, 1996).

À l'opposé, les tenants des corpus échantillonnés ou de corpus spécifiques promeuvent l'élaboration de corpus plus restreints en fonction de caractéristiques langagières ciblées et équilibrées au sein du corpus, privilégiant une construction raisonnée des corpus. Cependant, ces corpus sont donc plus sensibles aux variations des données linguistiques (Habert *et al.*, 1997).

Mais B. Habert et ses collègues (1997) considèrent qu'outre leur taille, ce qui est véritablement nouveau alors, c'est l'enrichissement de ces corpus. À partir des années 1980, les annotations se développent et les premiers corpus étiquetés[10], des corpus auxquels des informations morphosyntaxiques ont été ajoutées, apparaissent. Les années 1990, marquent le début des premiers corpus arborés[9], des corpus auxquels des informations syntaxiques ont été ajoutés.

De plus, dans cette même décennie, le nombre de langues disposant de corpus augmente également, ainsi que le nombre de corpus multilingues. Toutes ces nouveautés permettent le développement d'innovations technologiques et d'outils spécifiques marquant ainsi la jonction d'une partie du traitement automatique de langues avec la linguistique de corpus.

2.4. Des approches sur corpus à la linguistique de corpus

C'est au cours des années 1990 et 2000 que la linguistique de corpus se constitue véritablement en tant que domaine de la linguistique. En effet, les termes « corpus linguistics » sont employés pour la première fois en 1984 par J. Aarts et W. Meijs (cités par Léon, 2015), puis repris par de nombreux auteurs (par exemple McEnery & Wilson, 2001 ; Biber, Conrad, & Reppen, 1998 ; Tognini-Bonelli, 2001). En France, les termes *corpus linguistics* ont été traduits en français par *linguistique **sur** corpus* (Bilger, 2000), *linguistiques **de** corpus* au pluriel (Habert *et al.*, 1997 ; Condamines, 2005), et *linguistique **de** corpus* au singulier (Williams, 2005 ; Rastier, 2005 ; Condamines, 2005). La première de ces appellations met en avant l'usage des corpus par divers domaines de la linguistique (Williams, 2006). Les suivantes visent à faire de la linguistique qui

[10] Cette notion est davantage expliquée au chapitre 2.

emploie des corpus une discipline. Puis, cette dernière appellation, *linguistique **de** corpus*, s'est consolidée par de multiples parutions et colloques.

Au cours de cette décennie, les corpus se sont également diversifiés. En effet, face au problème de représentativité des corpus de référence, les corpus se sont peu à peu spécialisés et diversifiés (Habert, 2000), cherchant à recueillir des données issues de situations de communication de plus en plus diverses. En France, on observe par exemple l'élaboration de corpus journalistiques (Flintham, 1995 ; Abeille et al., 2001 ; Lecolle, 2001 ; par exemple) et de corpus d'écrits scientifiques (Bachschmidt, 1997 ; Gledhill, 1997 ; par exemple).

3. La constitution de la linguistique de corpus en tant que discipline

La linguistique de corpus est une discipline qui s'appuie sur des corpus, c'est-à-dire sur des données authentiques pour construire une description de cette langue et de ses usages (Williams, 2005). À ce titre, elle est basée sur une approche empirique (Tognini-Bonelli, 2001) de la langue et G. Williams (2005, p. 1) la décrit comme s'intéressant « à la langue en contexte sous la forme de grands ensembles, les corpus. ». Notons également que la linguistique de corpus a amené dans son sillage le développement de méthodes stochastiques (Léon, 2015). En effet, les faits de langues peuvent désormais être comptabilisés, ce qu'on va appeler la fréquence, il peut s'agir de la fréquence d'une forme ou d'un phénomène linguistique particulier.

Si le caractère méthodologique de la linguistique de corpus est très peu remis en cause, l'enjeu essentiel de la définition de ce champ est de déterminer s'il se restreint à cet aspect ou s'il s'agit également d'une nouvelle théorie linguistique et donc d'une nouvelle discipline (Tognini-Bonelli, 2001).

Le débat a fait rage au sein de la sphère linguistique dès l'apparition de corpus suffisamment grands pour opérer de profonds changements sur les observations des faits langagiers et s'est poursuivi durant plusieurs décennies. En effet, G. Leech (1992, cité par Léon, 2008) puis G. Williams (2005) ont affirmé que la linguistique de corpus constituait une discipline à part entière. Cependant, des auteurs comme M. Cori et S. David (2008) s'avèrent très critiques quant à cette affirmation : ils vont jusqu'à se demander si elle ne sert pas de légitimation à leurs travaux. C'est notamment la critique que fait J. Léon (2008) à G. Leech. De plus, F. Rastier (2005) rappelle qu'il ne s'agit en rien d'une discipline unifiée. M.-P. Jacques (2005) tout comme A. Condamines (2005) remettent en cause le singulier du terme *linguistique* dans cette appellation.

On peut estimer que les débats autour de l'émergence de la linguistique de corpus ont eu lieu en deux temps. Un premier débat, initié par les tenants de la linguistique chomskyenne, entre la linguistique sur corpus et la linguistique introspective, s'est penché sur la validité de ces deux courants et sur leur méthode d'obtention des données. Puis la volonté de certains linguistes d'instituer la linguistique de corpus en discipline a fait réémerger ce débat, accompagné d'un autre débat entre les tenants de l'approche *corpus-based* et les tenants de l'approche *corpus-driven*.

3.1. Contre la linguistique introspective

Dès l'apparition de la linguistique introspective, il est fait reproche à cette linguistique de s'appuyer sur un jugement subjectif de la validité ou non des énoncés employés. M.-P. Jacques (2005) cite ainsi les travaux de W. Labov (2001) et de P. Corbin (1980) qui montrent, pour le premier, l'écart conséquent qu'il y a entre le jugement des locuteurs et le jugement des linguistes, et, pour le deuxième, la grande disparité des jugements entre linguistes, au point que l'on peut se demander si ce qui est décrit représente la langue ou un idiolecte (Jacques, 2005). Rappelons également qu'un certain nombre de courants de la linguistique ne peuvent travailler sans corpus (sociolinguistique et linguistique de l'acquisition par exemple).

3.2. Apports supposés de la linguistique de corpus

L'usage de la linguistique de corpus devrait permettre, parce qu'elle travaille sur des données authentiques et de grande ampleur, de nuancer voire de corriger les intuitions des linguistes. Ces ajustements peuvent porter soit sur un phénomène connu mais dont la fréquence était sur ou sous-évaluée, soit sur la découverte d'un phénomène nouveau, que les analyses plus restreintes n'avaient pas permis de percevoir jusqu'alors.

De ce fait, les approches sur corpus revendiquent leur caractère réel, leurs propriétés d'objectivité, de représentativité, de vérifiabilité. C. D. Manning (2003, cité par Cori & David, 2008) avance même que ces approches devraient permettre de récuser des connaissances faussement établies. Ceci est permis notamment par les quantifications possibles qu'amène la linguistique de corpus. De plus, G. Kennedy (1998, cité par Léon, 2008), C. D. Manning (2003, cité par Cori & David, 2008), M.-P. Jacques (2005) estiment que les approches sur corpus permettent d'atteindre les fonctionnements linguistiques qui échappent à l'intuition. Ceci s'explique par le fait que ces approches permettent d'appréhender des termes ou des phénomènes réservés à des domaines spécifiques (médical, technique, etc.) ou à des usages moins accessibles par introspection, l'oral par exemple (Jacques, 2005) ou laissés de côté par

les courants dominants en linguistique, comme l'étude de l'acquisition ou de la variation (Cori & David, 2008) car le jugement d'acceptabilité y est plus compliqué.

Enfin, et c'est relativement nouveau en linguistique, la linguistique de corpus permet de s'intéresser à la variation (Condamines, 2005). En effet, peu de corpus, si ce n'est aucun, peuvent prétendre à la représentation de l'exhaustivité des usages langagiers ; en revanche, beaucoup ont pour objectif de représenter en profondeur un usage langagier ou un genre donné.

3.3. Critiques de la linguistique de corpus

À ces critiques, certains auteurs répondent que le travail de constitution de corpus est lui-même subjectif. En effet, un corpus ne contient pas l'ensemble des productions langagières attestées d'une langue, il est donc construit et les données langagières y sont choisies. Il y a donc là une première trace de subjectivité.

De plus, dans la pratique, beaucoup de corpus rassemblent des données de travail selon leur accessibilité davantage que selon un critère d'étude qui motive la recherche (Cori & David, 2008). Le besoin de représentativité est donc ici un réel enjeu. En effet, rien ne certifie que les données les plus accessibles soient représentatives.

Mais la subjectivité des concepteurs des corpus s'exprime également à un deuxième niveau de la constitution de corpus. En effet, une fois les données linguistiques assemblées, il est dans la plupart des cas nécessaire d'effectuer un travail de mise en forme afin de disposer de ces données dans un format unique, de les corriger, voire d'effectuer une deuxième sélection (Habert et al., 1998). Dans toutes ces étapes, et c'est particulièrement vrai dans l'étape de correction, la subjectivité du linguiste s'exprime. Et pour cause, ce travail de correction fait appel à son jugement introspectif. Dans la plupart des cas, les corpus n'échappent donc pas à l'introspection du linguiste (Raccah, 2018).

Enfin, si la linguistique introspective en se basant sur la seule intuition du linguiste ou du locuteur échoue à percevoir et rapporter certains phénomènes, jugés comme inexistants ou agrammaticaux, la linguistique de corpus n'est pas exempte de ce constat. En effet, il est couramment admis qu'un corpus si grand soit-il ne pourra jamais englober l'ensemble des productions possibles d'une langue. Ce constat fait dire à certains linguistes que la différence entre ces deux méthodologies linguistiques se situent dans le statut des données produites (par exemple Jacques, 2005 ; Cori & David, 2008). D'un côté, la linguistique introspective examine les énoncés en terme de possible de langue ; la tâche du linguiste ou du locuteur sera donc

d'évaluer la possibilité d'un énoncé en fonction de son intuition. De l'autre côté, la linguistique de corpus s'intéresse aux données en termes de « données attestées » *vs* « données non attestées ». Elle peut relever les énoncés existants, mais ne peut en aucun cas prédire les énoncés non existants : ce qui est non-attesté n'est pas forcément un impossible de langue (Fillmore, 1992).

Ces deux linguistiques adoptent des méthodologies différentes, s'intéressent à des productions langagières différentes et produisent des données au statut différent. On peut estimer que, d'une part, elles ne s'intéressent pas au même objet et n'observent donc pas les mêmes données et que, d'autre part, elles n'ont pas la même conception de ces données (Cori *et al.*, 2008) et de la langue. En effet, si la linguistique introspective s'intéresse à la langue de manière unifiée, la linguistique de corpus s'intéresse aux multiples attestations de la langue et à la variation. P. Corbin (1980) ; cité par Jacques, 2005) considère ainsi que la linguistique introspective s'intéresse à la langue en tant que « plus petit dénominateur commun à tous les membres d'une communauté linguistique » (Corbin, 1980, p. 155), tandis que la linguistique de corpus s'intéresse à la langue en tant que « la somme des énoncés dont on peut prédire qu'ils sont productibles par tout ou partie d'une communauté linguistique » (Corbin, 1980, p. 155).

3.4. Statut de discipline

L'apparition ou la réaffirmation de la linguistique de corpus a donc introduit de nouveaux objets en linguistique, et notamment des corpus de grande taille, informatisés et annotés, permettant ainsi à la linguistique de s'intéresser à de nouveaux champs d'étude (variation, genres nouveaux comme l'oral, etc.) mais également permettant un saut qualitatif de notre compréhension du langage (Halliday, 1993, p. 24, cité par Tognini-Bonelli, 2001). C'est en ce sens que de plus en plus d'auteurs lui attribuent le statut de discipline linguistique à part entière.

3.5. Courants internes

À ses débuts, les débats qui entourent la linguistique de corpus se sont essentiellement cristallisés sur le changement de paradigme qu'elle représentait par rapport à la linguistique introspective. Cependant, dès les années 1960, un second débat, interne, se déroulait entre les tenants d'une linguistique *corpus-based* et le courant *corpus-driven*, ceux que d'aucuns traduisent en français respectivement par « linguistique sur corpus », « approche sur corpus » ou « approche basée sur le corpus » et « linguistique de corpus » ou « approche dirigée par le corpus » (Mayaffre, 2005). Actuellement, les deux approches se distinguent toujours.

Selon l'historienne J. Léon (2008), les deux courants prendraient leur source dans la *London School* et les travaux de Firth, que ce soit revendiqué ou non par les initiateurs de ces courants. Le courant *corpus-based* envisage le travail sur corpus comme une approche déductive où les corpus constituent un réservoir d'exemples destiné à tester ou vérifier des hypothèses établies au préalable et aurait été à l'origine initié par G. Leech de l'université de Lancaster (Léon, 2008). Le second courant *corpus-driven*, initié autour des travaux de John Sinclair de l'université de Birmingham, adopte une démarche inductive qui vise à aborder le corpus sans présupposé théorique de manière à faire émerger les hypothèses de l'observation du corpus.

L'approche adoptée par le courant *corpus-based* peut être rapprochée d'une démarche linguistique classique (Tognini-Bonelli, 2001) où le linguiste élabore un modèle linguistique et où l'accès aux données (par introspection ou via des corpus) permet d'ajuster le modèle ou de fournir des exemples. Dans cette approche, les catégories ou variations qui apparaissent dans les corpus mais qui ne sont pas ressorties au moment de l'élaboration du modèle sont jugées peu importante. À l'inverse, l'approche *corpus-driven* permet de faire émerger des phénomènes linguistiques qui échappent à l'intuition et prend en compte ces phénomènes.

Le développement de l'informatique a permis l'émergence de collections de données linguistiques de plus en plus grandes et de plus en plus diversifiées, désormais constituées en corpus. De l'étude de ces corpus a émergé une nouvelle discipline des sciences du langage, la linguistique de corpus. Pour accompagner l'essor de cette nouvelle discipline, diverses méthodes et de nombreux outils de traitement automatique des langues ont été développés. Ces outils sont présentés au chapitre suivant (cf. Chapitre 2).

4. Conclusion

L'accès massif à l'informatique, l'accroissement des ressources numérisées et le développement d'outils de traitement automatique spécifiques ont permis l'élaboration de vastes corpus de données linguistiques. Avec l'accessibilité d'un grand nombre de données numérisées, la problématique des linguistes travaillant sur corpus s'est déplacée de la question du recueil du corpus à la question de la conversion des données dans un format exploitable linguistiquement.

Cependant, il est encore des domaines d'études où l'effort de collecte des données n'est pas anodin, c'est notamment le cas dans le champ des écrits scolaires. La taille des corpus qui y sont constitués (à l'exemple des corpus disponibles sur *Ortolang* et présentés en section 1 de ce chapitre) peut donc paraitre ridiculement petite au regard des vastes corpus constitués par ailleurs. Néanmoins, cela ne doit pas empêcher de mener des études linguistiques sur ces

corpus. En effet, comme le rappelle M.-P. Péry-Woodley (1995), il est parfois préférable de constituer des corpus de manière raisonnée et selon des critères définis et maitrisés, plutôt que d'assembler des vastes ensembles de données dont on ne maitrise pas toujours le contenu.

Enfin, notons que si la constitution de corpus est nécessaire dans de nombreux domaines et qu'il existe donc de nombreux corpus de recherche de taille restreinte, ces corpus sont souvent dans des formats non exploitables. Le processus de rendre son corpus de recherche exploitable par d'autres est couteux et de ce fait n'est pas si fréquent, il constitue donc un des enjeux de la recherche *Scoledit* dans laquelle s'inscrit cette thèse.

Chapitre 2 - Traitement automatique des langues et apport à l'exploitation des corpus

1. Comprendre le traitement automatique des langues 25
2. Apport du traitement automatique des langues à l'exploitation de corpus 30
3. Apports du traitement automatique des langues aux corpus d'écrits peu normés 35

Le traitement automatique des langues, parce qu'il applique certains traitements automatiques sur les langues, constitue souvent une aide à la manipulation d'ensembles de données trop grands pour être manipulés manuellement. C'est souvent le cas lorsqu'on mène des études linguistiques à partir de corpus. Ces dernières décennies, l'intérêt des linguistes de corpus pour le TAL s'est donc accru.

Dans ce chapitre, nous revenons sur l'explication de ce qu'est le traitement automatique des langues[11] et sur la façon dont ses méthodes ont évolué, avant de nous intéresser plus précisément à ce que le TAL a pu apporter à l'exploitation de corpus[12].

1. Comprendre le traitement automatique des langues

> *Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things. (Chowdhury, 2003)*
>
> *The goal of this new field is to get computers to perform useful tasks involving human language, tasks like enabling human-machine communication, improving*

[11] L'écriture de la section 1 s'inspire principalement de la lecture de deux ouvrages et d'un article :
- Bouillon, P., (1998). *Traitement automatique des langues naturelles*. Louvain-la-Neuve : De Boeck Supérieur.
- Fuchs, C., (1993). *Linguistique et traitements automatiques des langues*. Paris : Hachette Supérieur.
- Cori, M., (2008). Des méthodes de traitement automatique aux linguistiques fondées sur les corpus. *Langages* 171, 2008 (3), p. 95-110.

[12] L'écriture de la section 2 s'inspire principalement de la lecture de trois ouvrages :
- Lallich-Boldin, G., Maret, D., (2005). *Recherche d'information et traitement de la langue: fondements linguistiques et applications*. Villeurbanne: Presses de l'Enssib.
- Mitkov, R. (Ed.). (2004). *The Oxford handbook of computational linguistics*. Oxford: Oxford University Press.
- Pierrel, J. M. (2000). *Ingénierie des langues*. Paris : Hermes.

> *human-human communication, or simply doing useful processing of text or speech.*
> *(Jurafsky & Martin, 2016)*

Né à la fin des années 1940 avec les premiers projets de traduction automatique (Cori, 2008), le traitement automatique des langues est un domaine d'étude qui vise à comprendre et manipuler le langage naturel et les données linguistiques à l'aide de méthodes informatiques, formelles ou statistiques. Comme le mentionne les définitions données par G. G. Chowdhury et D. Jurafsky et J. H. Martin, citées ci-dessus, le TAL est à la fois un domaine de recherche et un domaine d'application.

1.1. Domaines d'application

Lorsqu'il a une visée applicative, le TAL se retrouve dans de nombreux domaines :

- La traduction automatique : traduire un texte dans une autre langue sans ou en minimisant l'intervention humaine ;
- La génération automatique de textes : produire un texte – comme un bulletin météo ou un diagnostic – à partir de données brutes ;
- La correction orthographique : corriger un texte ou proposer des corrections à l'utilisateur (les modules de correction sont inclus dans la plupart des logiciels de traitement de texte et à certains navigateurs Internet) ;
- La synthèse vocale : générer un flux de parole à partir d'un texte écrit ;
- La reconnaissance de la parole : générer un texte écrit à partir d'un discours oral ;
- La recherche d'informations : retrouver des informations à l'intérieur d'un ensemble de textes ;
- L'analyse d'opinions : analyser un ensemble de données et de discours pour évaluer l'opinion des auteurs ;
- etc.

Au-delà de ces aspects applicatifs, le TAL est également très lié à la description linguistique. Celle-ci est à la fois nécessaire pour élaborer des méthodes et des traitements automatiques, mais elle est aussi permise par les outils de traitements automatiques. Le TAL a donc également pour rôle d'aider la linguistique à comprendre le fonctionnement des langues naturelles. Cet aspect est particulièrement significatif lorsqu'on introduit du TAL dans le processus d'exploitation d'un corpus.

1.2. Fonctionnement d'un système de traitement automatique des langues

Que ce soit dans une visée applicative ou dans une visée descriptive, les méthodes de traitement automatique développées sont souvent similaires et reposent sur les mêmes éléments :

- Des **données linguistiques** (texte, discours, dialogue, etc.) ou des **données brutes** (matrice de données, représentation linguistique) à partir desquelles vont être réalisés les traitements ;
- Un **modèle linguistique** (règle, formalisme) ou **stochastique** (statistique et probabiliste) : ce modèle peut contenir des règles linguistiques, des règles d'apprentissage ou encore des calculs statistiques qui détermineront la nature des transformations à opérer. Un modèle linguistique se base sur des connaissances linguistiques généralement classées en niveaux d'analyse, tels que la phonétique, la phonologie, la morphologie, la syntaxe, ou encore la sémantique ;
- Une **séquence d'instructions** (algorithme) qui permettent d'opérer une transformation des données linguistiques ou des données brutes à partir des informations du modèle.

À partir de ces éléments d'entrée et après calcul algorithmique, une sortie est produite (Figure 1). Celle-ci peut être de différentes natures. Il peut s'agir d'un texte en langage naturel écrit ou oral (on parle alors de génération) ou d'une représentation des données linguistiques d'entrée comme des annotations ou des arbres syntaxiques (on parle alors d'analyse). Certains domaines d'application, comme la traduction automatique, peuvent nécessiter les deux approches. L'analyse permet alors de transformer la langue source en représentation puis la génération permet de transformer cette représentation en langue cible.

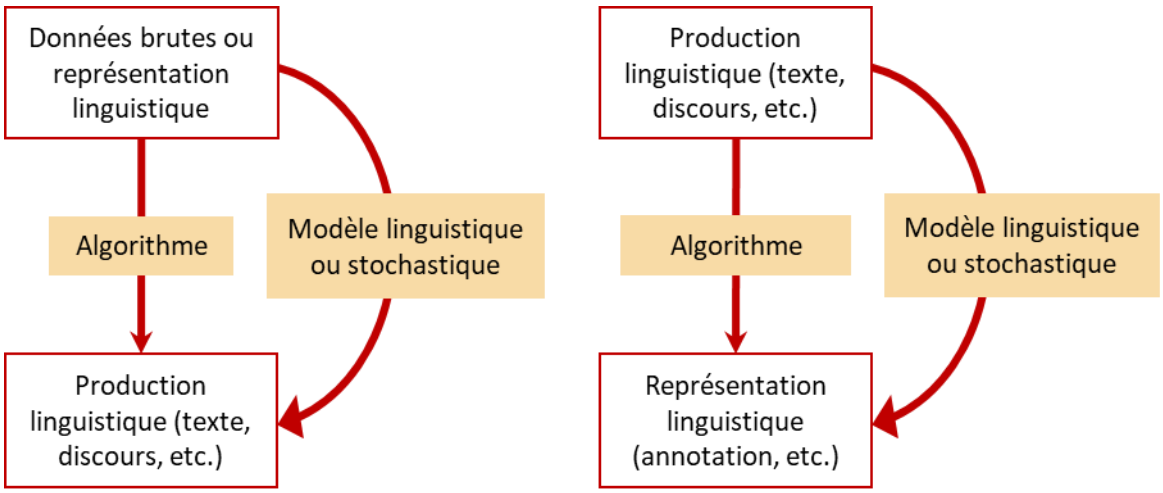


Figure 1 : Processus de génération (à gauche) et d'analyse (à droite)

Dans le cadre de l'utilisation du TAL pour l'exploitation de corpus, l'approche généralement nécessaire est celle de l'analyse : on analyse les données langagières, c'est-à-dire le corpus, pour en produire une représentation (annotations, calculs statistiques, etc.).

1.3. Méthodes du TAL

Dans ce domaine, il est d'usage de distinguer principalement deux types de méthodes (Poibeau, 2014) :

- Les **méthodes à base de règles** (dites méthodes **symboliques**). Ces méthodes sont principalement utilisées pour retrouver des informations de surface, accessibles à partir de règles simples ou de règles contextuelles (qui tiennent compte du contexte). Elles sont par exemple utilisées pour étiqueter des textes, étape qui consiste à donner la catégorie grammaticale et le lemme aux formes d'un texte (Cori, 2008) ;
- Les **méthodes statistiques** (dites méthodes **stochastiques**). Elles sont fondées sur des calculs statistiques effectués à partir de corpus. Elles sont particulièrement efficaces pour découvrir de nouvelles informations (Poibeau, 2014), par exemple des mots fréquemment co-occurrents ;

À ces méthodes s'ajoutent celles **basées sur des automates à états finis[13] ou des expressions régulières** qui, par rapport aux méthodes par règles, permettent une prise en compte plus large du contexte (Cori, 2008). Ces méthodes peuvent par exemple être utilisées dans des tâches de *chunck parsing*, tâche qui consiste à découper un texte en syntagmes.

Ces méthodes, et principalement les méthodes statistiques, peuvent être combinées avec de l'**apprentissage automatique**. Ce procédé consiste à apprendre un modèle linguistique ou statistique à partir d'un grand ensemble de données, que ce soit pour générer par apprentissage automatique les règles d'un système, pour modifier les valeurs des probabilités ou pour pondérer les transductions (les liens d'un état à un autre) des automates à états finis. On parle alors d'entrainer un système. Cependant, ce procédé implique plusieurs contraintes, à savoir la nécessité de disposer de suffisamment de données pour permettre l'entrainement

[13] Un automate à états finis est un outil formel qui se présente sous forme de graphe orienté et qui permet principalement de déterminer si une séquence de caractères est acceptée ou non par un langage donné. Dans le cadre du TAL, un tel outil a de nombreuses applications, comme l'étiquetage morphosyntaxique (assignation d'une catégorie grammaticale selon le mot identifié).

du système. Il est également nécessaire que ces données soient annotées et disponibles dans un format commun et suffisamment bien défini.

Enfin, ces dernières décennies de nouvelles méthodes sont apparues, basées sur de l'apprentissage profond (généralement) ou des réseaux de neurones qui permettent de modéliser des tâches complexes. Cependant, ces modèles peuvent être considérés commé problématique puisqu'ils sont souvent vus comme des *boites noires* rendant difficile l'interpréation des résultats (Allauzen & Schütze, 2018).

1.4. Evolution du TAL

Le domaine du traitement automatique des langues porte donc en son sein deux tensions inhérentes. Une première tension se manifeste entre les méthodes dites symboliques et les méthodes dites stochastiques. La seconde se traduit par la volonté de constituer une aide à la compréhension du langage combinée à la nécessité de constituer des applications robustes. Ces tensions se sont traduites à différents moment de l'histoire de ce domaine.

Les premiers traitements automatiques sur les langues sont nés avec le besoin, pendant la guerre froide, de traductions entre les langues anglaises et russes. Mais dès le milieu des années 1960, le TAL est jugé trop théorique et ne produisant pas suffisamment de résultats (rapport ALPAC, 1966[14]).

Le TAL, initialement centré sur la traduction automatique, cherche alors à se diversifier et à se constituer comme discipline scientifique à part entière, en adoptant une vision plus utilitariste et en pratiquant une recherche plus fondamentale davantage tournée vers la linguistique. De nouveaux champs de recherche sont explorés (grammaires formelles, systèmes de dialogue homme-machine, etc.), principalement basés sur des méthodes formelles. C'est à ce moment-là que le TAL trouve un nouvel essor grâce aux recherches théoriques en syntaxe de N. Chomsky et au développement de la théorie de la grammaire générative.

À partir des années 1990, suite au développement de l'informatique et d'Internet, les corpus se développent (cf. Chapitre 1) et les ressources textuelles disponibles augmentent, ainsi que la puissance de calcul (Pierrel, 2005). On assiste alors à un véritable engouement pour les méthodes statistiques et probabilistes (Habert *et al*., 1997). La disponibilité de ces données massives a permis également l'essor de l'apprentissage automatique, qui nécessite l'utilisation d'un grand nombre de données (Poibeau, 2014). Mais certains reprochent à ces paradigmes de

[14] National Research Council (US). Automatic Language Processing Advisory Committee. (1966). *Language and machines: computers in translation and linguistics; a report* (Vol. 1416). National Academies.

perdre de vue les objets et les unités manipulés, à contrario des méthodes symboliques, s'éloignant ainsi de la description linguistique.

Cette vision du TAL répond également au besoin de performance suite à l'introduction de ses méthodes dans des outils grand public. Émerge alors la notion de *TAL robuste*, plus applicatif, dont l'objectif est de pouvoir fournir des applications plus résistantes à la complexité des langues. Ce nouveau paradigme se fonde sur trois critères (Cori, 2008) :

- fonctionner avec de « vraies » productions langagières ;
- fournir une solution et une seule ;
- évaluer (quantitativement) les performances de l'outil développé.

Cependant, un certain nombre d'auteurs ont pu faire le reproche au TAL robuste de perdre de vue la recherche scientifique fondamentale, au profit de la recherche de performances applicatives.

1.5. TAL et corpus

Il y a donc entre le TAL et les corpus une interaction réciproque. D'une part, les linguistes de corpus ont besoin du TAL pour pouvoir manipuler de grands ensembles de données (Condamines, 2005) et en extraire les connaissances linguistiques et le contenu informatif (Pierrel, 2005). Ce point nous intéresse particulièrement et fera l'objet d'un exposé plus détaillé dans la section suivante (section 2). D'autre part, le TAL a particulièrement besoin des corpus, tant pour constituer de nouveaux outils, à partir des données textuelles et linguistiques contenues dans les corpus et des ressources (dictionnaires, lexiques, ontologies, etc.) constituées à partir de ces ressources, que pour évaluer les nouveaux outils ainsi constitués[15]. L'introduction de grands corpus a donc profondément changé le TAL.

2. Apport du traitement automatique des langues à l'exploitation de corpus

> *Comme chacun le sait, un texte ne saurait être assimilé à une masse de connaissances directement exploitable par la machine. Il faut dans un premier temps prévoir des traitements complexes pour identifier l'information pertinente, la normaliser, la catégoriser et éventuellement la mettre en contexte. Alors seulement l'ordinateur ou l'expert sera capable d'en tirer parti pour mener à bien ses analyses. [...] Elle [L'exploitation] exige la collaboration de spécialistes de*

[15] T. McEnery (2003) qualifie les données issues des corpus de « carburant pour le TAL » ("raw fuel of NLP", p. 448).

> *différents horizons, capables de traiter les données, de fournir les outils pour extraire l'information pertinente et d'ajuster de manière collaborative les traitements.* (Poibeau, 2014)

Des données linguistiques, quelles qu'elles soient ne sont quasiment jamais exploitables directement et nécessitent différentes interventions (numérisation, formatage, normalisation, enrichissement, etc.) pour les rendre exploitables (Poibeau, 2014). Le traitement automatique des langues fournit des outils et des méthodes pour certaines de ces interventions. En cela, le TAL représente une aide à l'élaboration des corpus.

En fournissant des outils d'exploration textuelle, le TAL assiste le travail d'analyse des linguistes en facilitant l'exploration de ces corpus, en permettant des analyses quantitatives plus conséquentes et par ce biais, en permettant aux linguistes d'observer des faits nouveaux qu'ils n'auraient pu observer manuellement (Pincemin, 2006). M.-Jacques (2017) considère ainsi que la constitution de ces outils fait partie intégrante de la recherche en linguistique. Le TAL représente donc également une aide à l'exploitation des corpus.

Dans la suite de ce chapitre, nous développerons ces deux aspects avant de nous concentrer plus spécifiquement sur l'apport du TAL aux corpus peu normés.

2.1. Elaboration de corpus et traitement automatique des langues

Bien qu'il soit possible de réaliser des analyses linguistiques à partir d'un corpus numérisé, celui-ci n'est véritablement avantageux que lorsqu'il est découpé en unités lexicales et annoté (étiquetage morphosyntaxique, arborescence syntaxique, annotation sémantique, etc.). Plus le corpus est volumineux, plus cette étape peut s'avérer fastidieuse. Le TAL fournit différentes méthodes pour assister ou automatiser cette étape d'annotation. Plus généralement, il fournit des méthodes qui constituent une aide à la création de ressources, comme des outils de normalisation pour les corpus peu normés (cf. section 3) ou encore des outils d'alignement pour les corpus multilingues. Nous détaillons certaines de ces méthodes ci-dessous.

2.1.1. Formalisme et codage des corpus

Pour être exploitable informatiquement, un corpus doit être dans un formalisme uniforme et défini. La plupart du temps, le TAL est absent de cette étape, mais l'usage des outils de traitement automatique est conditionné par celle-ci. Même indirectement, le TAL influe donc sur cette étape.

Selon les corpus, les formalismes utilisés peuvent avoir été créés spécialement au sein du projet de constitution. Mais de plus en plus, les projets tendent vers un formalisme commun. On citera

par exemple le formalisme *TEI*[16], particulièrement adapté aux corpus de transcription de manuscrits. Les formalismes sont souvent écrits dans des langages de balises comme *XML*.

2.1.2. La segmentation

Au cours de l'étape de segmentation, le texte est découpé en segments linguistiques afin de permettre son analyse. Généralement, on distingue deux types de segmentation : la segmentation en phrases et la segmentation en mots, aussi appelé tokenisation. La segmentation en mots ou en formes est généralement nécessaire à la plupart des traitements et analyse linguistique automatiques, notamment ceux d'ordre lexical, tandis que la segmentation en phrases, elle aussi nécessaire pour un grand nombre de traitements, l'est particulièrement pour des traitements d'ordre syntaxique.

2.1.3. La lemmatisation et l'étiquetage morphosyntaxique

L'étape de lemmatisation consiste, après tokenisation, à assigner à chaque forme du corpus un lemme c'est-à-dire la forme canonique d'une unité lexicale. Par exemple, le lemme des formes *mangea*, *mangèrent* et *manges* est MANGER.

L'étape d'étiquetage morphosyntaxique va généralement de pair avec la lemmatisation et ce sont souvent les mêmes outils (*LIA_TAGG*[17], *MElt*[18], *TALismane*[19], *TreeTagger*[20], par exemple) qui réalisent ces deux opérations. L'étape d'étiquetage morphosyntaxique, aussi appelée *PoS tagging*, permet d'assigner des informations de type morphosyntaxique comme la catégorie grammaticale et certains traits flexionnels (temps verbal, genre, nombre, etc.) à chaque forme du corpus.

Ces étapes sont relativement fréquentes lors de la constitution de corpus dit annotés et sont relativement bien maitrisées par les outils actuels. Contrairement à l'étape de segmentation, qui repose essentiellement sur des indices graphiques (ponctuation et espaces), elles

[16] Text Encoding Initiative (https://tei-c.org/) [consulté le 09/08/2019].

[17] Nasr A., Béchet F., Volanschi A. (2004). Tagging with Hidden Markov Models Using Ambiguous Tags. In Proceedings of COLING 2004, Stroudsburg, PA, USA. https://gitlia.univ-avignon.fr/jean-francois.rey/otmedia/tree/2ce72581f10ddc1cbcdbcf2bae0a8c9599c542ae/tools/lia_ltbox/lia_tagg [consulté le 22/09/2019].

[18] Denis, P., Sagot, B. (2010). Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français. http://gforge.inria.fr/projects/lingwb/ [consulté le 22/09/2019].

[19] Urieli, A., Tanguy, L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions: études de cas avec l'analyseur Talismane. *20e conférence du Traitement Automatique du Langage Naturel (TALN)*. http://redac.univ-tlse2.fr/applications/talismane.html [consulté le 22/09/2019].

[20] Helmut S. (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods. *Language Processing*, Manchester, UK. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ [consulté le 22/09/2019].

nécessitent un certain nombre de données linguistiques comme des dictionnaires de formes fléchies et des analyses syntaxiques de surface ou des modèles statistiques.

Cette étape constitue un premier niveau d'annotation. Un corpus présentant un lemme et une étiquette morphosyntaxique pour chaque forme est dit **corpus étiqueté**. À partir d'un tel corpus, il est possible de construire un lexique ou des listes de fréquence.

2.1.4. L'annotation syntaxique

Au cours de cette étape, une analyse syntaxique du corpus, au moins partielle, peut être réalisée. À l'issue de cette étape, des annotations représentant la structure syntaxique des énoncés peuvent être apposées aux corpus : on parle alors de **corpus arboré**, en référence aux structures syntaxiques sous forme d'arbre parfois utilisées. L'analyse syntaxique ne porte pas toujours sur les phrases complètes mais parfois uniquement sur des constituants plus petits de la phrase. Un corpus arboré permet de nombreuses analyses comme l'étude des cooccurrences et des dépendances syntaxiques.

2.1.5. L'annotation sémantique

L'annotation sémantique est « le processus qui consiste à établir des liens entre une ressource (le texte) et une autre ressource (ressource sémantique) » (Zargayouna et al., 2017). Elle permet ainsi de fixer l'interprétation d'un document (Ma et al., 2009). C'est une étape particulièrement délicate à réaliser automatiquement. J. Véronis (2000) distingue deux types d'annotations sémantiques : l'étiquetage du sens des mots et l'étiquetage des relations entre les mots. Ces annotations permettent par la suite d'analyser le contenu informatif du corpus.

2.1.6. L'alignement multilingue

L'alignement multilingue comprend généralement un alignement au niveau de la phrase ou au niveau des mots entre un texte et sa traduction dans les corpus multilingues. Cette étape est essentielle pour l'usage d'un corpus multilingue.

Les types d'annotations cités précédemment sont des exemples d'annotations les plus courantes, mais il en existe d'autres types (annotations prosodiques et annotations phonétiques notamment). La nature des annotations ajoutées détermine les traitements et les analyses linguistiques possibles sur le corpus.

Ces annotations représentent de formidables aides à l'exploration d'un corpus ; cependant elles sont souvent trop peu lisibles pour être utilisables par des non-spécialistes sans être accompagnés d'outils spécifiques.

2.2. Outils d'exploration des corpus

Outre les méthodes mentionnées ci-dessus qui ont à la fois un rôle dans la constitution et l'exploration d'un corpus, le TAL a également un rôle à jouer dans la constitution d'outils facilitant le parcours de ces corpus. Les formes que peuvent prendre ces outils sont variées, des outils les plus génériques aux plus spécifiques.

2.2.1. Outils génériques et outils spécifiques

Les outils d'exploration génériques (*Antconc*[21], *Anatext*[22], *Le Trameur*[23], *TXM*[24], par exemple), souvent disponibles sous forme de logiciels ou de sites Web, ont pour but d'être manipulés par des linguistes et des non-spécialistes des outils informatiques tout en permettant d'être utilisables sur des corpus de différentes natures. Ils doivent donc pouvoir s'adapter à différents types d'annotations. Notons cependant que ces outils n'incluent pas toujours du TAL ou seulement dans quelques uns de leurs modules.

Les outils spécifiques sont souvent conçus sous formes d'interfaces dans le but de rendre interrogeables les corpus pour lesquels ils ont été développés. On citera par exemple l'interface d'accès au corpus *Frantext*[25] ou encore celle permettant d'accéder au *British National Corpus*[26] (*BNC*).

À mi-chemin entre ces deux opposés, il existe également des outils semi-génériques, c'est-à-dire des outils conçus pour s'adapter à différents corpus, mais dont l'intégration ne peut être réalisée que par un des membres du projet. À titre d'exemple, nous pouvons citer les outils en ligne *le Lexicoscope*[27] et *Scienquest*[28] développés dans le cadre des projets *Emolex* et *ScienText*.

[21] Anthony, L. (2005, July). AntConc: design and development of a freeware corpus analysis toolkit for the technical writing classroom. *IPCC 2005. Proceedings. International Professional Communication Conference, 2005*. (pp. 729-737). IEEE. http://www.laurenceanthony.net/software.html [consulté le 09.08.2019]

[22] http://olivier.kraif.u-grenoble3.fr/anaText/ [consulté le 09.08.2019]

[23] Fleury, S. (2007). Le métier textométrique: le trameur, manuel d'utilisation. Secteur TAL Sorbonne nouvelle. Consulté le 09.08.2019. http://www.tal.univ-paris3.fr/trameur/

[24] (2007) Projet Textométrie, *Projet Textométrie*. Consulté le 09.08.2019. http://textometrie.ens-lyon.fr/

[25] https://www.frantext.fr [consulté le 09.08.2019]

[26] http://www.natcorp.ox.ac.uk/using/index.xml [consulté le 09.08.2019]

[27] Kraif, O., Diwersy, S. (2012). Le Lexicoscope: un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques. *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012*, volume 2: TALN (pp. 399-406). http://phraseotext.univ-grenoble-alpes.fr/emoBase/ [consulté le 09.08.2019]

[28] Falaise, A., Tutin, A., Kraif, O. (2012). Une interface pour l'exploitation de corpus arborés par des non informaticiens: la plate-forme ScienQuest du projet Scientext. *Traitement Automatique des Langues*, 52(3), 201-228. https://corpora.aiakide.net/scientext20/ [consulté le 09.08.2019]

2.2.2. Types d'exploration

Chacun de ces outils a ses spécificités qui permettent la prise en compte de certaines annotations (corpus étiqueté ou corpus arboré par exemple) et présente un certain nombre de fonctionnalités. Les fonctionnalités proposées par ces outils sont multiples mais la plupart d'entre elles relèvent du champ de l'analyse textométrique et lexicographique ou de la syntaxe. Parmi les fonctionnalités principales, on retrouve :

- L'extraction de **listes de fréquences**, qui permet d'extraire la liste des tokens du corpus et les fréquences associées ;
- La fonction de **concordancier**, outil qui permet de rechercher un token ou une séquence de caractères et d'afficher les contextes d'apparition de ceux-ci ;
- La recherche de **cooccurrents**, tokens qui apparaissent fréquemment dans le même contexte ; de nombreux outils permettent de les identifier ;
- Le calcul des **spécificités**, fonctionnalité qui permet de faire ressortir les spécificités lexicales du corpus étudié par rapport à un corpus de référence ;

Il est néanmoins important de rappeler que ces fonctionnalités et la qualité de leurs sorties dépendent en partie du type et de la qualité des annotations apposées sur le corpus.

Enfin, outre l'exploration du contenu linguistique des corpus, de plus en plus de travaux issus du TAL s'intéressent au contenu informationnel des corpus. Il s'agit de travaux portant sur l'extraction de données, la recherche d'information, l'analyse d'opinion, etc. Mais ces usages ne relèvent plus tant du champ de l'analyse linguistique que de l'analyse discursive.

3. Apports du traitement automatique des langues aux corpus d'écrits peu normés

Les outils précédemment présentés ont majoritairement été développés pour des corpus standards proches d'une certaine norme identifiable. Mais ces méthodes peuvent échouer ou paraitre peu efficaces face à des corpus d'écrits peu normés, c'est-à-dire éloignés de la norme.

On citera, en guise d'exemples de corpus peu normés, les corpus de communication électronique écrite[29], comme les corpus de SMS, de tweets, de messages issus de tchat, d'applications de messagerie ou encore de forums, et les corpus d'apprenants de langue

[29] Il existe un grand nombre de termes pour désigner ce type de corpus. R. Panckhurst (2009) et E. Kogkitsidou (2018) en présentent un certain nombre. Nous reprenons à J. Anis (2003) les termes de « communication électronique ».

étrangère ou seconde. Dans le premier cas, les variations importantes par rapport à la norme d'élaboration des outils sont généralement dues au contexte de production. Dans le second, ces variations sont souvent dues à une maitrise imparfaite de la langue employée.

Pour chacun de ces cas, nous nous intéressons aux types d'outils TAL développés et utilisés et les apports de leur utilisation.

3.1. Le traitement automatique des langues et les corpus de communication électronique écrite

La communication électronique écrite regroupe sous un terme générique les productions écrites médiées par ordinateur ou par téléphone et qui présentent généralement des caractéristiques communes. Elles sont généralement courtes, que ce soit pour des raisons matérielles (taille de l'écran par exemple) ou par contrainte de la part des plateformes (à l'instar de *twitter*) et présentent à la fois des caractéristiques de l'écrit et de l'oral (Crystal, 2006). E. Kogkitsidou (2018) liste un certain nombre de procédés linguistiques présents dans le corpus de SMS qu'elle a étudié et qui peuvent se retrouver dans d'autres types de communication électronique écrite. Elle mentionne notamment un grand nombre de procédés qui permettent un gain de place, nécessaires au moment de l'élaboration du corpus où les caractères des SMS étaient comptés, comme les phénomènes de capitalisation (« G trouV », *j'ai trouvé*), d'abréviation (« jtm », *je t'aime*) ou encore de phonétisation (« biz », *bise*). Un certain nombre de SMS incluent également des émoticônes, symboles schématisant la plupart du temps une figure (« :-) » par exemple), et de l'alternance codique, l'alternance au sein d'une même production ou d'un même échange de plusieurs langues.

Tous ces phénomènes expliquent que la normalisation de ces productions soit un préalable à l'application d'autres traitements automatiques (Kobus et al., 2008b). La notion de normalisation sera rediscutée dans un chapitre ultérieur. Pour le moment et dans ce contexte, nous nous appuyons sur la définition de G. Jose et N. Raj (2014) selon laquelle la normalisation consiste à réécrire un écrit en utilisant une orthographe plus conventionnelle afin de rendre ce message plus lisible pour un humain et pour une machine.

Comme le montre cette définition, la normalisation des corpus de communication électronique écrite est la plupart du temps vue comme une tâche de normalisation orthographique parfois accompagnée d'un module de re-segmentation (Guimier de Neef & Fessard, 2007; Kobus *et al.*, 2008b ; Beaufort, Roekhaut, Cougnon, & Fairon, 2010b ; Han & Baldwin, 2011 ; Kogkitsidou, 2018 ; notamment). Parmi ces travaux, C. Kobus, F. Yvon et G. Damnati (2008a) distinguent

distinguent différents types d'approches ou « métaphores », reprenant chacune des méthodes issues d'autres domaines du TAL :

- La métaphore de la **correction orthographique** (Brill & Moore, 2000; Toutanova & Moore, 2002 ; Choudhury, Saraf, Jain, Sarkar, & Basu, 2007 ; Guimier de Neef, Debeurme, & Park, 2007 ; Cook & Stevenson, 2009 et Han & Baldwin, 2011, cités par Kobus *et al.*, 2008a ; Tarrade, 2017). Dans cette approche, on considère que les écrits électroniques comportent un certain nombre d'erreurs ou de variations qu'il convient de corriger. Il s'agit alors d'une correction mot à mot des formes considérées comme erronées par comparaison à un dictionnaire de formes fléchies. La méthode la plus utilisée repose sur l'identification des formes erronées, la génération d'une liste de formes candidates possibles et la sélection de la meilleure forme candidate. L'inconvénient majeur de cette méthode est de ne pouvoir envisager que les « formes hors vocabulaire » (*out of vocabulary*), c'est-à-dire les formes qui n'existent pas, par exemple « disa » pour *dit* ou « diser » pour *disait*. De nombreuses erreurs comme les erreurs grammaticales ou morphologiques ne peuvent donc être prises en compte, par exemple « il dis » pour *il dit*.
- La métaphore de la **traduction** (A w, Zhang, Xiao, & Su, 2006 ; Raghunathan & Krawczyk, 2009, cités par (Kobus *et al.*, 2008b). Cette approche envisage la normalisation comme une traduction d'une langue source vers une langue cible ; la langue source étant l'écrit électronique et la langue cible l'écrit normalisé (Beaufort et al., 2010a). Cette approche repose principalement sur des ressources produites par entrainement à partir d'un corpus aligné préalablement. De plus, cette approche est considérée comme dépassant largement les besoins en normalisation de SMS (Choudhury *et al.*, 2007). Enfin, elle ne permet pas d'anticiper la créativité linguistique de ce type d'écrits (Kobus et al., 2008a).
- La métaphore de la **reconnaissance vocale** (Kobus et al., 2008b). Cette approche se base sur l'hypothèse que les formes trouvées dans les écrits électroniques sont parfois plus proches de leur réalité phonémique que de leur forme graphique. Dans cette approche, les écrits étudiés sont entièrement convertis en phonèmes puis re-segmentés en tokens à l'aide de dictionnaires et d'un modèle de langue. L'avantage de cette approche est donc de permettre de mieux gérer les problèmes de frontières de mots.

Désormais, on trouve de plus en plus de travaux hybrides, mêlant ces différentes méthodes, à l'exemple des travaux de l'équipe de R. Beaufort et de ses collègues (Beaufort et al., 2010a) et des travaux d'E. Kogkitsidou (2018) qui mêlent méthodes de correction et méthodes issues du champ de la traduction.

Dans le cadre des corpus d'écrits peu normés tels que les corpus électroniques écrits, il est donc souvent fait appel au TAL pour produire une normalisation automatique ou semi-automatique de ces écrits avant de pouvoir leur appliquer les outils d'analyses classiques.

3.2. Apports du TAL au traitement des corpus d'apprenants

L'emploi du TAL pour l'analyse des productions d'apprenants n'est pas nouveau. Depuis le début des années 1980 (Šmilauer, 2011), de nombreux projets se sont intéressés à l'introduction de ces méthodes dans des environnements d'enseignement ou d'apprentissage des langues assisté par ordinateur (ELAO et ALAO). On citera par exemple les projets *ROBO-SENSEI* (Nagata, 2009) et *E-Tutor* (Heift, 2010) qui incluent différents outils de TAL pour analyser les productions des apprenants et leur fournir un retour diagnostique. Pour une revue plus complète de ce domaine, il est possible de se référer à T. Heift et M. Schulze (Heift & Schulze, 2007 ; Heift, 2017).

Des systèmes similaires ont été développés pour le français, notamment au sein du laboratoire *Lidilem*, comme l'outil *ExoGen* (Blanchard et al., 2009) qui vise la génération d'exercices pour les apprenants et la production d'évaluations diagnostiques de leurs réponses ou la plateforme *MIRTO* (Antoniadis, Echinard, Kraif, Lebarbé, Loiseau, & Ponton, 2004) qui propose de mettre à disposition des outils TAL pour élaborer du matériel pédagogique.

Plusieurs projets ont vu le jour également dans le domaine de l'apprentissage langue première, notamment dans le domaine de l'orthographe, comme la plateforme *PlatON*[30] (Beaufort & Roekhaut, 2011) qui permet de s'entrainer à la dictée en autonomie à l'aide d'un système de synthèse vocale et une analyse automatique des erreurs produites par l'apprenant.

Cependant, la majorité de ces projets reposent sur des réponses limitées et / ou contraintes de la part des apprenants, que ce soit des textes à trous, des réponses courtes attendues ou des exercices de dictée (Meurers, 2013).

À la suite des divers travaux d'élaboration de corpus que nous avons mentionnés au chapitre précédent (cf. Chapitre 1), plusieurs équipes ont cherché à développer de grands corpus d'apprenants. Dans cette section, nous nous appuyons sur la définition donnée par S. Granger, (2007) : « les corpus d'apprenants, également appelés corpus d'interlangue (IL), sont des recueils électroniques de données authentiques de langue étrangère ou de langue seconde ».

[30] Disponible désormais en version commerciale sur la plateforme *Ortalia* (https://www.ortalia.com/) [consulté le 22/09/2019].

Pour des raisons de proximité avec le sujet de cette thèse, nous nous limiterons à la mention des corpus écrits.

La plupart des projets les plus conséquents se sont portés sur la constitution de corpus d'écrits d'apprenants de l'anglais (cf. Pravec, 2002 et Tono, 2003 pour une revue de ces corpus). Mentionnons tout de même le projet *FreeText* (Granger et al., 2001) qui a permis l'élaboration du corpus *FRIDA* pour le français. Ce corpus contient des productions écrites d'apprenants adultes de langues maternelles diverses.

Comme nous l'avons montré précédemment (cf. section 2), le TAL peut être un réel atout pour l'élaboration et l'exploitation des corpus. Néanmoins, en raison du grand nombre de variations (que ce soient des erreurs ou des usages différents de certains phénomènes) présentes dans les corpus d'apprenants par rapport aux corpus de locuteurs / scripteurs natifs, appliquer des méthodes de traitement automatique des langues n'est pas aisé. Les productions contenues dans ces corpus sont généralement bien moins contraintes que dans les systèmes d'assistance pour l'apprentissage des langues mentionnés à la section précédente. Bien que les variations rencontrées soient similaires, les méthodes développées pour ces systèmes ne peuvent donc s'appliquer à l'identique pour les corpus d'apprenants.

Néanmoins, plusieurs projets de constitution de corpus d'apprenants se sont intéressés à l'emploi de méthodes issues du TAL. On peut distinguer deux principaux apports du TAL pour ce type de corpus : les outils TAL peuvent représenter une aide à l'annotation, ainsi qu'une aide à l'exploitation des annotations, et plus généralement à l'exploitation de ces corpus.

Nous reprenons à D. Meurers (2015) et S. Granger (2013) la distinction entre deux types majeurs d'annotations utilisés lors de l'élaboration de corpus d'apprenants :

- Les **annotations linguistiques**, qui permettent l'identification d'un certain nombre de propriétés langagières et se basent sur les catégories linguistiques générales, utilisées pour tous types de corpus (annotations lexicales, syntaxiques, sémantiques, morphologiques, etc.) ;
- Les **annotations des erreurs**, qui identifient les différences avec les écrits des locuteurs natifs et la façon dont les énoncés linguistiques sont altérés, par rapport à un écrit standard.

Y. Tono (2003) propose une revue des différents corpus et des annotations associées à ces corpus.

3.2.1. Annotations linguistiques et outils TAL

Plusieurs projets se sont penchés sur l'automatisation de la phase d'annotation linguistique des corpus d'apprenants. Ces travaux se sont principalement intéressés à l'étiquetage morphosyntaxique (catégorie grammaticale et lemme), car c'est une des méthodes les plus maitrisées et les plus utilisées en TAL pour les corpus. Certains travaux ont repris des étiqueteurs et systèmes d'annotation déjà existants (Pravec, 2002), mais ceux-ci ont été développés et entrainés à partir de corpus d'écrits produits par des locuteurs natifs. Ils ne permettent donc de prendre en compte ni les structures erronées produites par les apprenants, ni leurs spécificités (sur ou sous-emploi de certaines unités linguistiques par exemple). Le corpus *USE* (Axelsson, 2000) est un corpus anglophone d'écrits d'apprenants de langue étrangère de plus de 600 000 mots. Lors de l'élaboration ses concepteurs ont choisi d'utiliser l'outil d'étiquetage développé par E. Brill en 1992, appelé *Brill-tagger*[31], mais son usage a nécessité un ré-étiquetage à la main des productions.

Pour pallier ce problème, S. Verlinde (2010) évoque la possibilité d'utiliser des outils qui peuvent être modifiés et ré-entrainés, comme *TreeTagger*, mais cela nécessite de disposer de données d'apprentissage, c'est-à-dire de corpus préalablement étiquetés. De la même façon, certains projets se sont penchés sur l'adaptation aux productions d'apprenants de systèmes déjà existants (*CLAWS7* et *TOSCA-ICLE Tagger* par exemple). R. van Rooy et L. Schäfer (2003) ont comparé et évalué trois systèmes d'étiquetage morphosyntaxique, un système entrainé pour des locuteurs natifs, le système *Brill-tagger* (voir plus haut) et deux systèmes adaptés aux écrits d'apprenants, les systèmes *TOSCA-ICLE*[32] et *CLAWS7*[33]. Leur étude ne montre pas beaucoup d'amélioration pour le système *TOSCA-ICLE* mais une amélioration de 10 % pour *CLAWS7*.

Enfin, une autre méthode a été choisie lors du développement du système *eXXelant* (Antoniadis, Ponton, & Zampa, 2007) qui constitue une interface d'interrogation du corpus d'apprenants *Frida*. Ce système repose sur une double annotation : un balisage manuel des erreurs et un étiquetage morphosyntaxique automatique. Cette deuxième phase d'annotation repose sur les informations ajoutées lors de la première.

---

[31] Brill, E. (1992, March). A simple rule-based part of speech tagger. *Proceedings of the third conference on Applied natural language processing* (pp. 152-155). Association for Computational Linguistics.

[32] de Haan, P. (2000). Tagging non-native English with the TOSCA-ICLE tagger. *Language and computers*, 33, 69-80.

[33] Garside R, Smith N 1997 A hybrid- grammatical tagger: CLAWS4. In Garside R, Leech G, McEnery A (eds) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Essex, Addison Wesley Longman, pp 102-121. Pour une version antérieure de l'outil *CLAWS.*

3.2.2. Du traitement automatique des langues pour assister l'annotation d'erreurs

Comme le mentionne S. Granger (2013), un des intérêts d'étudier les productions des apprenants est d'analyser la façon dont elles diffèrent des productions des locuteurs natifs, à la fois en termes de sur ou sous-emplois de certaines structures linguistiques, ce que les annotations linguistiques discutées à la section précédente permettent de faire, mais aussi en termes d'erreurs orthographiques, lexicales, syntaxiques, etc. Pour étudier ces erreurs, une annotation préalable est souvent nécessaire. Ces annotations sont souvent réalisées de manière manuelle, mais certains travaux se sont intéressés à l'annotation automatique des erreurs dans les corpus d'apprenants.

De nombreux outils de correction automatique existent mais ils ont été principalement conçus pour corriger des productions de locuteurs natifs, comportant peu d'erreurs, ou pour des productions aux caractéristiques bien spécifiques, comme les analyseurs de SMS et de tweets mentionnés précédemment. Différentes études, comme celles de S. Granger et F. Meunier (1994) et de J.-B. Johannessen, K. Hagen et P. Lane (2002) sur les correcteurs grammaticaux, montrent que les outils classiques ne permettent de corriger qu'un petit nombre d'erreurs produites par les apprenants de langue étrangère. Il est donc nécessaire de modifier ces outils pour les adapter à ce type de productions ou d'en construire de nouveaux.

Dans le cadre du projet *FreeText* (corpus *FRIDA*), S. L'Haire et A. Vandeventer-Faltin (2003) ont par exemple choisi d'utiliser l'analyseur syntaxique *FIPS*[34] en relâchant certaines contraintes pour permettre la détection des structures erronées et réaliser un diagnostic des erreurs. Mais bien souvent, les travaux en ce sens se limitent à l'annotation de certains phénomènes particuliers comme certaines erreurs grammaticales ou encore l'absence de certains articles (Tono, 2003 ; Díaz-Negrillo & Domínguez, 2006). La difficulté à créer des systèmes automatiques d'annotations d'erreurs s'explique par le grand nombre d'erreurs contenues dans ce type d'écrits, par comparaison avec les écrits des locuteurs natifs. Cette difficulté s'explique également par l'absence de connaissances sur ces erreurs, à savoir leur fréquence, leur contexte d'apparition, etc., ou par l'absence d'un corpus de référence, où les erreurs seraient préalablement annotées manuellement ou semi-automatiquement avec vérification manuelle. Ces données permettraient d'établir des systèmes à base de règles ou d'entrainer des systèmes probabilistes et de les évaluer (Milton & Chowdhury, 1994, cités dans Tono, 2003 ; Meurers, 2015).

[34] Wehrli, E. (1997). *L'analyse syntaxique des langues naturelles: Problèmes et méthodes.* Paris: Masson.

Précisons qu'utiliser des outils de traitement automatique pour annoter des corpus nécessite de déterminer des schémas d'annotation relativement consensuels, développés à travers ces outils. La littérature fait état de nombreuses publications en ce sens (Dagneaux, Denness, & Granger, 1998 et Díaz-Negrillo & Domínguez, 2006 pour une revue des taxonomies et schémas d'annotations proposés).

L'absence de systèmes performants à l'heure actuelle explique que dans de nombreux projets, l'annotation d'erreurs est réalisée de manière manuelle. Dans certains projets, cette annotation est assistée par ordinateur à l'aide de logiciels facilitant l'insertion des annotations, comme l'outil *Error Editor*[35] (Dagneaux *et al.*, 1998) développé à Louvain-La-Neuve.

3.2.3. Exploitation automatique des productions d'apprenants

À plusieurs reprises, S. Granger (2004 ; 2007 notamment) met en avant l'intérêt des outils issus du TAL pour exploiter les corpus en eux-mêmes, à travers les annotations apposées de manière automatique ou manuelle. Elle distingue plusieurs apports principaux, la constitution de listes de fréquences et de lexiques d'erreurs, ou encore la comparaison des structures linguistiques utilisées par les apprenants par rapport à ce que produisent les locuteurs natifs ou les apprenants ayant des L1 différentes. Pour ce faire, les outils et logiciels les plus robustes du TAL (logiciels de textométrie, concordanciers, etc.) peuvent être utilisés. À titre d'exemple, l'outil *WordSmith Tool*[36] a été utilisé dans le corpus *FRIDA* (projet *FreeText*) pour analyser certaines données linguistiques (Granger, 2004).

Outre l'analyse linguistique des productions, Meurers (2015) mentionne différents usages des méthodes de TAL comme la détection automatique de la langue maternelle ou première des apprenants (Tetreault, Blanchard et Cahill, 2013, cités par Meurers, 2015) ou encore l'évaluation de leur niveau (Pendar & Chapelle, 2008; Yannakoudakis *et al*., 2011; Hancke & Meurers, 2013, cités par Meurers, 2015). Un autre exemple est donné par l'outil *eXXelant* (Antoniadis *et al.*, 2007), qui permet d'extraire des exemples d'un corpus d'apprenants de français (*FRIDA-bis*) en s'appuyant sur les annotations présentes dans celui-ci et représente ainsi un support pédagogique pour les enseignants et les apprenants.

Bien que de nombreux travaux aient essayé d'introduire des méthodes et des outils de TAL au sein des corpus d'apprenants, leur apport est encore limité et nécessite bien souvent des traitements manuels en amont ou en aval. Néanmoins, le TAL peut constituer une aide à

[35] Hutchinson, J. (1996). *UCL error editor*. Louvain-la-Neuve: Centre for English Corpus Linguistics, Université Catholique de Louvain.
[36] Scott, M. (1996). *WordSmith tools*. Oxford: Oxford University Press.

l'élaboration de ces corpus. De la même façon, nous pensons que le TAL pourra constituer à terme également une aide pour l'élaboration de corpus scolaires. Nous n'abordons pas ici l'usage encore très balbutiant du TAL dans l'élaboration et l'exploitation de ces corpus. Ce point est abordé dans le chapitre suivant (cf. Chapitre 3).

Chapitre 3 - Les corpus d'écrits scolaires : un domaine en pleine dynamique

1. Corpus scolaires : état des lieux 45
2. Nécessité de constituer des corpus d'écrits scolaires accessibles 53
3. Traitements automatiques de ces corpus 55
4. Conclusion 57

Le travail développé dans le projet *Scoledit* s'inscrit dans le champ de la linguistique de corpus en cela qu'il vise à caractériser linguistiquement les écrits produits par des apprenants d'âge scolaire à partir de l'étude d'un corpus. À cette fin, nous développons un **corpus d'écrits scolaires**[37] (que nous raccourcissons parfois en **corpus scolaire**).

Nous appelons corpus d'écrits scolaires des corpus rassemblant des textes produits par des élèves en milieu scolaire (enseignement primaire et secondaire). Peu nombreux, les corpus d'écrits réalisés par des enfants d'âge scolaire (de 3 à 18 ans) en dehors du milieu scolaire sont parfois également assimilés à des corpus d'écrits scolaires. D'autre fois, ils sont appelés corpus d'écrits extrascolaires. Une des caractéristiques principales de ces corpus est que les scripteurs sont généralement en cours d'apprentissage de l'écrit. Il en résulte des corpus souvent assez éloignés de la norme attribuée aux locuteurs/scripteurs experts. Nous limitons ici la notion de corpus d'écrits scolaires aux corpus d'apprentissage de la langue première ou langue de scolarisation par opposition aux corpus de langue seconde.

1. Corpus scolaires : état des lieux

Comme le rappellent C. Doquet et J. David (2018), les travaux de recherche s'appuyant sur les corpus d'écrits scolaires s'inscrivent dans différents paradigmes, dont les principaux sont la didactique de l'écrit, la linguistique de l'écriture et des textes et la psychologie cognitive. Les objectifs qui sous-tendent l'élaboration de ces corpus peuvent donc être de natures diverses (David, 2000 ; Garcia-Debanc, 2015) :

- décrire linguistiquement les productions des apprenants (Clanché, 1988 ; Penloup, 2000 ; Cappeau & Roubaud, 2018) ;

[37] Appellation empruntée à M.-L. Elalouf et C. Boré, (2007), notamment.

- étudier la progression des apprentissages en écriture des enfants d'âge scolaire et identifier les zones de fragilité linguistique persistantes au cours de leur scolarité (Aurnague & Garcia-Debanc, 2016 ; Elalouf & Boré, 2007) ;
- analyser la genèse du processus d'écriture (Fabre, 1990 ; David & Doquet, 2016 ; Similowski, Pellan, & Plane, 2018) ;
- étudier la diversité des pratiques scripturales et la diversité des genres auxquels les jeunes apprenants sont confrontés (Boré, 2007a, 2007b et 2007c ; Gunnarsson-Largy & Auriac-Slusarczyk, 2013 ; De Vogüé, Espinoza, Garcia, Perini, & Marzena Watorek, 2017) ;
- constituer un support à la formation des enseignants (Masseron, 2004 ; Elalouf, 2005 et 2011) ;
- etc.

Les deux tableaux suivants (Tableau 2 et Tableau 3) dressent la liste d'un certain nombre de corpus d'écrits scolaires ou extrascolaires constitués ou en cours de constitution. Il ne prétend pas à l'exhaustivité. Un certain nombre de corpus non disponibles ont été écartés, notamment des corpus de thèses non rendus accessibles numériquement. Le projet *Scoledit* visant la constitution d'un corpus de textes produits à l'école primaire (CP-CM2[38]), nous nous sommes principalement penchée sur les corpus incluant des écrits réalisés par des élèves de cycle 2 (CP-CE2) et de cycle 3 (CM1-6e) de l'école française.

Le premier tableau répertorie les projets non francophones les plus saillants réalisés dans le champ des corpus d'écrits d'apprenants langue première ou maternelle. Le second se concentre plus particulièrement sur les corpus de langue française. Bien que ces tableaux ne soient pas exhaustifs, nous souhaitons montrer la pluralité des corpus actuellement constitués et en dresser un bref état des lieux.

Dans ces deux tableaux, les corpus portent la mention « non disponible » lorsqu'aucun accès au corpus n'a été identifié et qu'il n'en est fait mention dans aucun article ou dans aucune référence d'un travail de numérisation dépassant les quelques productions utilisées en exemples dans des communications. Il est possible que certaines voies d'accès à certains corpus aient échappé à notre attention. Les corpus surlignés sont des corpus particulièrement marquants et sur lesquels nous nous sommes appuyée tout au long de notre recherche ; ils sont exposés plus en détail dans la suite de ce chapitre.

[38] Se référer au Glossaire pour une correspondance entre les classes et les âges.

Nom du corpus *Auteurs*	*Date* *Durée du projet*	*Langue*	*Classe* *Age*	*Nombre d'enfants*	*Nombre de textes*	*Nature du recueil*	*Numérisation et accessibilité du corpus*	*Publication de référence* *Lien vers la ressource*
C. Juel	4 ans	Anglais (États-Unis)	du 1er au 4e grade	54 enfants	200 - 300	Recueil longitudinal d'écrits narratifs	non disponible	Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of educational Psychology*, 80(4), 437.
K. Perera	1985	Anglais (Angleterre)	8, 10 et 12 enfants	48 enfants	quelques dizaines	Compte-rendu d'une activité de construction en Lego effectuée plus tôt	non disponible	Perera, K. (1985). Grammatical Differentiation between Speech and Writing in Children Aged 8 to 12.
Lancaster Corpus of Children's Project Writing R. Ivanic et T. McEnery	1994-1996 (trois ans)	Anglais (Angleterre)	9-11 ans	36 (11 en longitudinal)	Plusieurs centaines	Recueil de l'ensemble des productions d'une séquence thématique. Suivi de 11 élèves sur 5 séquences pendant 3 ans.	Scans et transcriptions sont disponibles en ligne	Smith, N., McEnery, A., Ivanic, R. (1998). Issues in Transcribing a Corpus of Children's Handwritten Projects. *Literary and Linguistic Computing*, Vol.13, No.4. Oxford: OUP. *https://www.lancaster.ac.uk/fass/projects/lever/index.htm*
N. Chipere, D. Malvern et B. Richards	Années 2000	Anglais (Angleterre)	7, 11 et 14 ans		918	Essais narratifs de 50 mots à partir d'une phrase donnée	non disponible	Chipere, N., Malvern, D., Richards, B. (2004). Using a corpus of children's writing to test a solution to the sample size problem affecting type-token ratios. *Corpora and language learners*, 139-147.
Oxford Children's Corpus K. Wild, A. Kilgarriff et D. Tugwel	2006	Anglais (Angleterre)	5 - 14 ans		73 875	Écrits issus de sites internet où les enfants postent des critiques, des poèmes ou des histoires	Corpus annoté disponible sur *SketchEngine*	Banerji, N., Gupta, V., Kilgarriff, A., Tugwell, D. (2013). Oxford Children's Corpus: a corpus of children's writing, reading, and education. *Corpus Linguistics* 2013, 315. *https://www.sketchengine.eu/oxford-childrens-corpus/*
H. Roessingh, S. Elgie et P. Kover	2011	Anglais (Canada)	8-9 ans	77 enfants	77	Textes argumentés rédigés dans un temps contraint (45 minutes)	Non transcrit et non disponible	Roessingh, H., Elgie, S., Kover, P. (2015). Using lexical profiling tools to investigate children's written vocabulary in grade 3: An exploratory study. Language Assessment Quarterly, 12(1), 67-86.

Karlsruhe Children's text Corpus J. Fay	2011-2013	Allemand	écoles primaires et secondaires	1700 élèves	14 000	Primaire : écrit narratif après lecture de deux histoires Secondaire : récit fictionnel de projection dans l'avenir	Scans et transcriptions disponibles en ligne	Lavalley, R., Berkling, K., Stüker, S. (2015). Preparing children's writing database for automated processing. In LTLT@ SLaTE (pp. 9-15). *https://catalog.ldc.upenn.edu/LDC2015T22*
H1, H2, E2, ERK1 Children's Writing Corpus K. Berkling	H1 : 2014/2015 Autres : 2016/2017 (2-4 mois)	Allemand	6 - 11 ans	H1 : 88 élèves Autres : 173 élèves	H1 : 996 Autres : 2 117	Rédaction d'une histoire ou description à partir d'une image pendant 15 minutes. Consigne répétée toutes les semaines pendant 9 à 16 semaines	Scans, transcriptions et normalisations disponibles en ligne	Berkling, K. 2016). Corpus for children's writing with enhanced output for specific spelling patterns (2nd and 3rd grade). In Nicoletta Calzolari, *et al*., editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portoroz, Slovnia, May 23-28, 2016. European Language Resources Association (ELRA). Berkling, K. (2018). A 2nd Longitudinal Corpus for Children's Writing with Enhanced Output for Specific Spelling Patterns. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018). *https://catalog.ldc.upenn.edu/LDC2016T01* *https://catalog.ldc.upenn.edu/LDC2018T05*

Tableau 2 : Liste non exhaustive de corpus scolaires non francophones existants[39]

[39] Nous avons écarté les corpus de dictées, mais il en existe quelques-uns qu'il nous parait intéressant de mentionner, au vu de leur ampleur ou de leur accessibilité.

- Le *Birkbeck spelling error corpus*. Ce corpus rassemble plusieurs centaines de dictées anglophones de tout âge et de taille variable. Il est accessible dans la base *Oxford Text Archive* (http://ota.ox.ac.uk/desc/0643).
  Référence disponible : Mitton, R. (1985). A collection of computer-readable corpora of English spelling errors. *Cognitive Neuropsychology*, 2(3), 275-279.
- Le *Hong Kong corpus*. Ce corpus rassemble entre autres 290 dictées de 13 mots composés de deux caractères en langue chinoise. Il n'est, à notre connaissance, pas aisément accessible.

Référence disponible : Yeung, P. S., Ho, C. S. H., Chik, P. P. M., Lo, L. Y., Luan, H., Chan, D. W. O., Chung, K. K. H. (2011). Reading and spelling Chinese among beginning readers: What skills make a difference?. *Scientific Studies of Reading*, 15(4), 285-313.

Nom du corpus *Auteurs*	*Date* *Durée du projet*	*Classe* *Age*	*Nombre d'enfants*	*Nombre de textes*	*Nature du recueil*	*Numérisation et accessibilité du corpus*	*Publication de référence* *Lien vers la ressource*
C. Fabre	1984 1 année	CP - CM2	Quelques centaines	450 (brouillons et copies)	Brouillons d'élèves, écriture en multiples jets	Non disponible	Fabre, C. (1990). Les brouillons d'écoliers ou l'entrée dans l'écriture. *Tem, texte en main*. Librairie de l'Université. Grenoble.
P. Clanché	1988 1 année	CP - CM2	200	7 500	Textes libres	Non disponible	Clanché, P. (2010). Anthropologie de l'écriture et pédagogie Freinet. Presses universitaires de Caen.
M. Charolles	1988	10-12 ans	64	64	Tâche de cohésion textuelle à partir d'un ensemble d'images	Non disponible	Charolles, M. (1988). La gestion des risques de confusion entre personnages dans une tâche rédactionnelle. *Pratiques*, *60* (1), 75-97.
C. Boré	1992-1996	10-12 ans (CM2 – 6e)	4 classes (90 élèves)	122	Brouillons scolaires d'écrits de fiction	Disponible en annexe de thèse (version papier)	Boré, C. (1998). Choix énonciatifs dans la mise en mots de la fiction: le cas des brouillons scolaires (Doctoral dissertation, Université Stendhal (Grenoble)).
M.-N. Roubaud	À partir de 1997, plusieurs années	5 - 11 ans		Plus de 1 500 textes	Écrits produits en classe, sans intervention de la recherche, genres variés	Disponible partiellement dans le *corpus ÉMA* (voir p. 50)	Roubaud, M. N. (2017). Le français écrit: transcription et édition. Le cas des textes scolaires. *Corpus*, (16).
Corpus Astrapi M.-C. Penloup	2000	7-11 ans	4000	4000	Courrier de lecteurs (écriture extrascolaire)	Non disponible	Penloup, M. C. (2001). De quelques propriétés d'une pratique de lecture extrascolaire: le courrier des lecteurs du journal Astrapi. *23*(1), 75-91.

C. Boré	Avant 2005	6e (11-12 ans)	26	142 textes	Récits narratifs à partir d'un conte	Non disponible	Boré, C. (2007). La métamorphose d'un genre: quelques descripteurs pour un genre scolaire de récit. In Boré, C. (dir.) Construire et exploiter des corpus de genres scolaires : Echos de la journée d'étude du 10 juin 2006 (Vol. 10). Presses universitaires de Namur.
M.-F. Fradet-Le Coz	de 2001 à 2006	classes de 6ème	130	380 textes (plusieurs jets par texte)	*Écrits* fictionnels en trois ou quatre jets	Disponible en version papier et numérisé sur CD Rom (annexes de thèse)	Fradet-Le Coz, M. F. (2009). La construction de la fiction dans l'écriture de textes narratifs à visée littéraire à l'entrée au collège: quand le dialogue pédagogique interfère avec le dialogue intérieur chez les jeunes scripteurs. Thèse de Doctorat, Université Paris Est–Paris XII Val de Marne. Document PDF de la thèse : *https://www.theses.fr/2009PEST0029*
Y. Reuter	à partir de 2001	CP-CM2		plusieurs centaines de textes	Récits et descriptions produits dans une perspective de pédagogie Freinet	Non disponible	Reuter, Y. (2006). Les récits sollicitant le vécu au CM2. Éléments d'analyse et de comparaison. *Repères. Recherches en didactique du français langue maternelle*, *34*(1), 111-139.
B. Lafourcade	2003 à 2005 (2 années)	Classe de seconde	77 lycéens + 7 élèves + 30 scripteurs experts (étudiants)	une centaine de copie	Deux versions successives d'une suite de nouvelles (traitement de texte ou support papier)	Disponible en version papier (annexes de thèse)	Lafourcade, B. (2008). Traitement des contraintes formelles liées au genre et au medium de production par des scripteurs novices: étude didactique (Doctoral dissertation, Paris Est). Document PDF de la thèse : *https://www.theses.fr/2008PEST0059/document*
M.-L. Elalouf	2005	CM2 (5 classes), 6e (2 classes) et 5e (1 classe)	8 classes	500 textes	Recueil de plusieurs versions de textes de types variés	Disponible sur CD Rom, transcrit et normalisé	Elalouf, M.-L. (dir.). (2005) *Ecrire entre 10 et 14 ans. Un corpus, des analyses, des repères pour la formation*. Paris, Scérén, CRDP Versailles, CDDP Essonne.
C. Ghienne	2013-2014	CM2 et 6e	3 classes		Rédactions et textes produits durant l'année, ainsi que les brouillons et les états intermédiaires	Non disponible	Ghienne, C. (2015). Corpus de brouillons d'élèves de CM2/6e: du recueil à l'analyse. In *SHS Web of Conferences* (Vol. 20, p. 01008). EDP Sciences.

Corpus Grenouille H. Andersen, C. Leblay et E. Auriac-Slusarcyk	2010	8-14 ans (CE2, CM2, 6e et 4e)	170 élèves (CE2 : 20 ; CM2 : 56 ; 6e : 48 ; 4e : 46)	584 textes (CE2 : 76 ; CM2 : 166 ; 6e : 200 ; 4e : 142)	Deux versions (un premier jet et une version définitive) de récits et comptes-rendus scientifiques réalisés à partir de six images en deux jets	Scans diffusés mais non transcrits	Gunnarsson-Largy, C., Auriac-Slusarczyk, E. (2013). Présentation du corpus Grenouille. In Ecriture et réécriture chez les élèves. Un seul corpus, divers genres discursifs et méthodologies d'analyse, Ed. Academia-Bruylant, pp. 7-14, 2013. *https://philosophemes.msh.uca.fr/corpus?field_corpus_tid=46*
C. Vénérin-Guénez	2011 - 2015	9 - 11 ans		230	Rappels de récit (reformulation) après écoute du même conte pour l'ensemble des élèves. Deux versions successives.	Non disponible	Vénérin-Guénez, C. (2018). À la recherche de stratégies scripturales dans des rappels de récit par des élèves de cours moyen. *Repères. Recherches en didactique du français langue maternelle*, (57), 83-98.
Corpus C'est pas moi E. Auriac-Slusarcyk et M.-H. Foulquier	2012, 2013	CP (6 - 7 ans)	38 élèves	70 textes	Deux jets successifs de la même histoire, rédigée à partir d'une histoire entendue et des images de l'album.	Transcriptions disponibles en ligne mais scans non diffusés	2012 : *https://philosophemes.msh.uca.fr/cafa98b5ff532fb31b37498142ef8ed8/c %C3 %A9critmoi-2012* 2013 : *https://philosophemes.msh.uca.fr/b9ff3fde6c15c837415c0b125c1850c9/c %C3 %A9critmoi-2013*
S. Castagnet-Caignec	2012, 2015 et 2016	CE1 - CM2, 5e, 3e, 1e L	161 élèves	469 textes	Tâche de novélisation (passage du récit filmique au récit écrit). Trois films utilisés dans différents niveaux scolaires	Non disponible	Castagnet-Caignec, S. (2018). Traitement du temps dans des récits à visée littéraire chez les élèves de primaire et secondaire. *Repères. Recherches en didactique du français langue maternelle*, (57), 35-56.
Corpus Ecriscol J. David et C. Doquet	en cours depuis 2013	CE1 - entrée à l'université		plusieurs milliers de textes	Tâche de continuation de récit	Scans, transcriptions et annotations en cours de diffusion	Doquet, C., Enoiu, V., Fleury, S., Maziotti, S. (2017). Problèmes posés par la transcription et l'annotation d'écrits d'élèves. *Corpus*, (16). *http://www.univ-paris3.fr/corpus-ecriscol-300513.kjsp?RH=1416243625396*

Extrait du projet *Dynascript* S.-D. Vogüé *et al.*	*Donnée inconnue*	enfants (7-11 ans), étudiants et adultes, entandants et sourds, locuteurs de LSF ou non	480 personnes (160 enfants)	2880 prévus (dont 960 produits par les enfants)	Écrits de 6 genres variés (récit fictionnel, récit d'expérience, description, exposition, argumentation, recommandation)	En cours d'élaboration	Vogüé, S. D., Espinoza, N., Garcia, B., Perini, M., Sitri, F., Watorek, M. (2017). Constitution d'un grand corpus d'écrits émergents et novices: principes et méthodes. *Corpus*, (16).
Corpus ÉMA C. Boré, M.-L. Elalouf	à partir de 2015	CP - CM2	15 classes	821 textes	Cinq sous-corpus de différents genres scolaires : écrits fictionnels, dialogue, brèves de journal et conte étiologique	235 textes disponibles sur *Ortolang*, transcrits et annotés (normalisés)	Boré, C. et Elalouf M.-L. (2017) : Deux étapes dans la construction de corpus scolaires : problèmes récurrents et perspectives nouvelles. In *Corpus 16*, 31-63. *https://hdl.handle.net/11403/ema-ecrits-scolaires-1/v2*
Corpus PreCPhi E. Auriac-Slusarcyk	2016	CM1, CM2, 4ème, 5ème (9-11 ans, 12-14 ans)	742 élèves (CM1 : 72 ; CM2 : 93 ; 5e : 112 ; 4e : 465)	951 brouillons et 951 textes	Productions de textes à partir d'une œuvre d'art ou d'un thème. Un deuxième jet à plusieurs mois d'intervalle a parfois été recueilli.	Scans diffusés mais non transcrits	Maire, H., Auriac-Slusarczyk, E., Slusarsczyk, B., Daniel, M. F., & Thebault, C. (2018). Does One Stand to Gain by Combining Art with Philosophy? A Study of Fourth-Year College (13/14 Years of Age) Philosophical Writings Produced within the" Precphi/Philosophemes" Corpus. *Journal of Education and Learning*, *7*(4), 1-19. *https://philosophemes.msh.uca.fr/corpus?field_ages_des_eleves_tid=All&field_niveau_tid=All&field_classe_tid=All&field_methode_tid=All&field_corpus_tid=126&field_modalite_tid=All*
Corpus Resolco C. Garcia-Debanc et K. Bonnemaison	en cours	9-16 ans (CM1, CM2 et collège)	400 élèves	400 textes	Tâche de cohésion textuelle à partir d'un ensemble de phrases	Transcriptions et normalisations en cours de diffusion	Garcia-Debanc C. & Bonnemaison K. (2014). « La gestion de la cohésion textuelle par des élèves de 11-12 ans : réussites et difficultés », *Actes du Congrès Mondial de Linguistique Française* (CMLF 2014), Berlin, Germany, 961-976.
K. Similowski, D. Pellan et S. Plane	*Donnée inconnue*	CM1 - CM2	6 classes	146	Récits narratifs, consigne différente selon les groupes, parfois rédigés en deux versions	Non disponible	Similowski, K., Pellan, D., Plane, S. (2018). Que révèlent les traces de réécriture dans les brouillons d'élèves produisant des récits à partir de sources littéraires ?. *Repères. Recherches en didactique du français langue maternelle*, (57), 15-34.

Tableau 3 : Liste non exhaustive de corpus scolaires existants pour le français langue première ou langue de scolarisation

2. Nécessité de constituer des corpus d'écrits scolaires accessibles

Les années 1970 et 1980 marquent le début d'un développement plus conséquent des travaux sur l'apprentissage de l'écriture (Simon, 1973, cité par David, 2000 ; Fayol, 1985 et Schnweuly 1988, cités par Bonnet & Gardes-Tamine, 1990 ; etc). Nombre de ces travaux s'appuient sur des corpus d'écrits scolaires (Clanché, 1988 ; Charolles, 1988, citée par Garcia-Debanc & Bonnemaison, 2014 ; Fabre, 1990 ; etc.). Ces corpus sont souvent de taille très restreinte et sont réalisés au sein des projets d'étude pour lesquels ils sont consultés. Il n'y a alors pas de volonté de les numériser en vue de leur diffusion.

A. Baron, P. Rayson, P. Greenwood, J. Walkerdine et A. Rashid (2012), ainsi que M.-L. Elalouf & C. Boré (2007), émettent l'hypothèse que cette absence de volonté est notamment dû à la trop grande complexité de la tâche de numérisation des corpus scolaires notamment en raison de certaines de leurs spécificités. Les écrits scolaires sont souvent manuscrits, ce qui nécessite une transcription manuelle pour être numérisés. De plus, ils sont souvent peu normés, tant en termes d'orthographe que de ponctuation ou de structures syntaxiques, etc. Enfin, il est souvent nécessaire pour comprendre et analyser les textes des apprenants de disposer à la fois des brouillons et des copies finales et de pouvoir les mettre en regard, ce qui complexifie la structure des corpus à élaborer.

M.-L. Elalouf en 2011, puis C. Garcia-Debanc et K. Bonnemaison en 2014 appellent à un renouvellement de la didactique du français grâce à la constitution de corpus d'écrits scolaires de grande taille, permettant d'introduire des méthodes issues de la linguistique de corpus. Comme l'estime C. Garcia-Debanc (2015, p. 200), « un nombre de textes suffisant pour constituer une masse critique de données est nécessaire pour dresser une cartographie des acquisitions » dans différents domaines (orthographe, syntaxe, cohésion textuelle, etc.).

Dès 2005, M.-L. Elalouf et ses collègues publient un premier corpus de plusieurs centaines de textes entièrement numérisés, transcrits et normalisés, ce qui représente le premier travail de ce type pour le français langue de scolarisation. À sa suite, d'autres projets, pour la plupart encore en cours, ont émergé. De tels corpus, sont qualifiés de grands par différents auteurs comme C. Garcia-Debanc et K. Bonnemaison (2014) non par comparaison aux corpus linguistiques déjà existants, mais par comparaison aux corpus scolaires de très petite taille existant jusqu'alors.

Avant les travaux précurseurs de M.-L. Elalouf, il n'existait, pour le français, que des corpus d'apprenants langue seconde[40] (Granger, 2007)). Quelques corpus d'apprenants en langue de scolarisation existaient déjà dans d'autres langues, à l'exemple du *Lancaster Corpus of Children's Project Writing* (Smith et al., 1998). Ce projet, considéré comme précurseur dans le domaine, rassemble un grand nombre de textes rédigés par un groupe d'élèves suivis pendant trois ans. Le corpus longitudinal permet de suivre l'ensemble des textes produits par 11 élèves au sein de cinq thématiques pendant ces trois ans.

À la suite des travaux de M.-L. Elalouf, d'autres projets francophones ont vu le jour. Citons par exemple le *Corpus ÉMA* (Boré & Elalouf, 2017), dont une partie est d'ores et déjà en ligne sur la plate-forme *Ortolang*[41]. Ce corpus rassemble l'ensemble des productions d'une classe donnée pour une séquence donnée. Pour chaque texte est donné un scan, une transcription, une annotation et des métadonnées. Ce corpus est toujours en cours de développement.

Parallèlement à ce travail, le projet E-CALM, initié par Claire Doquet, a vu le jour en 2016. Il vise à rassembler dans un large corpus, un certain nombre de corpus plus restreints déjà constitués ou en cours de constitution.

- Le corpus *Scoledit*, sur lequel repose ce travail de thèse, qui rassemble des textes recueillis de manière longitudinale du CP au CM2 ;
- Le corpus *Ecriscol* (Doquet *et al.*, 2017), élaboré au sein du laboratoire *Clesthia* et qui rassemble un grand nombre de textes scolaires (du CE1 jusqu'à l'entrée à l'université) ainsi que leurs avant-textes (brouillons, notes, etc.) ;
- Le corpus *Resolco* (Garcia-Debanc & Bonnemaison, 2014), élaboré au sein du laboratoire *CCLES*, qui rassemble des textes recueillis à partir d'une tâche de cohésion textuelle du CE2 jusqu'au master ;
- Le corpus *Littératie avancée* (Jacques & Rinck, 2017), élaboré au sein du laboratoire *Lidilem*, qui rassemble des textes recueillis en licence et en master.

À terme, le projet *E-CALM* rassemblera des textes d'apprenants, produits en milieu scolaire ou universitaire, du CP à l'université, et devrait permettre d'étudier l'évolution de certains phénomènes linguistiques emblématiques de l'acquisition de la maitrise de l'écrit.

[40] Pour plus de précisions, il est possible de se référer au chapitre 2, section 3.2., ainsi qu'à la liste de corpus anglophones d'apprenants élaborée par M. Weisser http://martinweisser.org/corpora_site/learner_corpora.html [consulté le 21/08/2019].

[41] ÉMA (École, Mutations, Apprentissages, ÉA 4507) (ÉMA) (2018). *Corpus ÉMA, écrits scolaires* [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, https://hdl.handle.net/11403/ema-ecrits-scolaires-1/v2.

Un des principaux objectifs est également la publication d'un vaste corpus de référence d'écrits scolaires et universitaires. La mise en commun, au sein du projet *E-CALM*, de l'ensemble de ces corpus est particulièrement importante puisqu'elle permet à l'ensemble de ces projets une certaine interopérabilité, grâce à des choix de transcriptions, de numérisation, de normalisation et parfois de recueil communs, et donc une plus grande comparabilité.

En 2010, E. Nonnon constate que, bien que la notion de progressivité soit une notion centrale en didactique, peu nombreux sont les travaux qui s'y penchent explicitement. À sa suite, C. Garcia-Debanc et K. Bonnemaison (2014) pointent le manque de matériau, c'est-à-dire de corpus de textes adéquats, disponible pour étudier l'évolution des acquis. Depuis lors, plusieurs travaux, à l'exemple des projets que nous venons de présenter, proposent un recueil à différents moments de la scolarité dans l'objectif de pouvoir étudier cette notion. Cependant, les corpus véritablement longitudinaux, c'est-à-dire proposant un suivi des élèves sur plusieurs années, sont encore peu nombreux.

3. Traitements automatiques de ces corpus

La numérisation de grandes quantités de textes scolaires est une première avancée pour le domaine de la didactique de l'écrit et lui permet de disposer d'un grand nombre de données partagées par toute la communauté. Mais, si elles ne sont pas accompagnées d'outils de traitement automatique, les analyses possibles sont souvent restreintes, soit en termes de complexité des requêtes possibles, soit en termes de taille des échantillons analysés. Plusieurs projets se sont donc intéressés à l'introduction de tels outils dans leurs analyses.

Le numéro 10 de la collection *Diptyque* (Boré, 2007a) relate plusieurs de ces expériences. Par exemple, des analyses ont été réalisées à partir du corpus publié en 2005 et coordonné par M.-L. Elalouf. Ces analyses (voir aussi Elalouf, 2005) ont été conduites à l'aide du logiciel *Tropes*, logiciel d'analyse de discours, et visent l'étude des modes et styles discursifs dans les productions scolaires. I. Fenoglio propose également d'utiliser le logiciel *MEDITE*[42] pour l'étude génétique de textes scolaires. Ce logiciel permet d'identifier les différences (déplacements, insertions, suppressions et remplacements de blocs de caractères) entre deux versions d'un même texte. Enfin, C. Boré, (2007b) et D. Malrieu présentent une étude menée sur les genres scolaires des écrits d'élèves à partir du logiciel *Cordial*[43], un correcteur orthographique dont les résultats des analyses linguistiques sont accessibles.

[42] Ganascia, J. G., Fenoglio, I., & Lebrave, J. L. (2004). EDITE MEDITE: un logiciel de comparaison de versions. *Actes de JADT*, 468-478.
[43] https://www.cordial.fr/ [consulté le 30/09/2019].

Dans le projet *Dynascipt* (De Vogüé et al., 2017), il est également prévu d'intégrer des analyses automatiques, à l'aide d'un logiciel de textométrie comme *TXM*[44] ou encore *le Trameur*[45]. Ce dernier est d'ores et déjà utilisé pour l'analyse de plusieurs corpus scolaires, dont principalement le corpus *Ecriscol*. Les outils de textométrie intègrent souvent des étiqueteurs morphosyntaxiques, comme *TreeTagger*, capables de donner les catégories grammaticales et les lemmes des formes d'un corpus. Pour le corpus *Resolco*, le choix s'est porté sur l'étiqueteur maison appelé *Talismane*[46].

Cependant, comme le mentionne C. Boré dans ce même ouvrage (Boré, 2007a), ces analyses ne peuvent se faire que sur des versions normalisées ou *réorthographiées* des productions (afin que les formes utilisées puissent être identifiées automatiquement par les différents outils) ou de manière manuelle et intuitive. Pour être plus explicite, on citera la recherche du lexème « accord » menée au sein de son corpus. Une recherche sous la forme lemmatisée *accord* lui permet de retrouver l'ensemble des occurrences fléchies orthographiées selon la norme attendue mais ne lui permet pas de retrouver les variantes orthographiques telles que « acor » ou « accors » que seule la connaissance de ce type d'écrits lui permettra de retrouver en allant spécifiquement chercher ces formes. Une certaine connaissance préalable de ces formes est donc nécessaire pour pouvoir les identifier.

Pour pallier ces problèmes, plusieurs projets de constitution de corpus scolaires intègrent des annotations contenant une proposition de graphie normée. Mais, à l'heure actuelle, ces annotations doivent être réalisées manuellement ou de manière aidée, à l'aide de logiciels d'annotations, tels que *Glozz*[47].

La conception d'outils permettant l'analyse automatique directe des productions d'apprenants ou, tout du moins, la transition de manière automatique d'une version transcrite, telle que produite par les enfants, à une version normalisée analysable par les outils précités reste à faire.

---

[44] Heiden, S., Magué, J. P., & Pincemin, B. (2010). TXM: Une plateforme logicielle open-source pour la textométrie-conception et développement. *10th International Conference on the Statistical Analysis of Textual Data - JADT 2010, June 2010*, Rome, Italie. p. 1021-1032.

[45] Fleury, S. (2013). *Le Trameur. Propositions de description et d'implémentation des objets textométriques*. Publication sur le site de l'Université Paris 3. Consulté le 03/10/2019. http://www.tal.univ-paris3.fr/trameur/trameur-propositions-definitions-objets-textometriques.pdf.

[46] Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. (Doctoral dissertation, Université Toulouse le Mirail-Toulouse II).

[47] Widlöcher, A., & Mathet, Y. (2009). La plate-forme Glozz: environnement d'annotation et d'exploration de corpus. In Actes de la 16e Conférence Traitement Automatique des Langues Naturelles (TALN'09), session posters, Jun 2009, Senlis, France, France. 10 p.

En 2011, K. Berkling, J. Fay et S. Stuker développent une première étude en ce sens, au sein du *Karlsruhe children writing corpus*. L'objectif est de mettre au point un outil de classification automatique des erreurs d'orthographe présentes dans les productions écrites d'apprenants. Pour ce faire, un alignement est généré automatiquement entre la production écrite et une version corrigée de cette production, produite manuellement ou de manière automatique. Cette étude a été menée pour un corpus d'apprenants en langue allemande. Des études similaires pour le français restent donc à faire.

4. Conclusion

Un renouvellement des méthodes est en train de s'opérer en didactique de l'écrit. Cependant, ce renouvellement a besoin d'être accompagné de matériaux et d'outils spécifiques qui font encore actuellement défaut tels que de vastes corpus de textes scolaires largement accessibles.

En effet, les corpus constitués sont encore peu numérisés et peu accessibles. De plus, si nombre de ces corpus réalisent des recueils transversaux, c'est-à-dire à différents moments de la scolarité, peu le font de manière longitudinale, c'est-à-dire en suivant les mêmes élèves. Ces corpus sont essentiels pour approcher la notion de progressivité.

Notons également que parmi les corpus accessibles en ligne, peu s'intéressent aux apprenants les plus jeunes (CP-CE1). Si un foisonnement de projets visant l'élaboration et la diffusion de grands corpus d'écrits scolaires peut être constaté, les besoins sont encore grands.

Enfin, comme nous l'avons montré au chapitre 2, le traitement automatique des langues peut être un réel apport pour l'exploitation de nombreux corpus. Or, les projets portant sur l'introduction d'outils de TAL pour l'exploration de ces types de corpus sont encore restreints.

Le projet *Scoledit* vise à poser les premières pierres pour répondre à certains de ces manques.

Partie 2 - Constitution du corpus *Scoledit*, ressource longitudinale d'écrits scolaires annotés

Chapitre 4 - Constitution et exploitation d'un corpus scolaire 61
Chapitre 5 - Conception et numérisation du corpus *Scoledit* 87
Chapitre 6 - La normalisation, une annotation déportée 121

Chapitre 4 - Constitution et exploitation d'un corpus scolaire

1. Numérisation d'un corpus 62
2. Enrichissement et annotation d'un corpus 73
3. L'approche par comparaison : un préambule à l'exploitation 83
4. Conclusion 85

Précédemment (cf. Chapitre 1), nous avons défini les corpus comme des ensembles de données rassemblées ou collectées selon un critère linguistique donné, ou dans le cadre d'une situation de communication unifiée, stockées électroniquement et de manière exploitable informatiquement sous un format texte, pouvant être agrémentées de données supplémentaires (annotations, version normée, représentation phonologique, etc.). Selon cette définition, quatre étapes majeures se distinguent dans l'élaboration d'un corpus :

1. La définition du protocole de recueil. Au cours de cette étape, la méthodologie de recueil des données est déterminée à partir des objectifs définis de la recherche ; se décident par exemple la nature des matériaux recueillis, les participants au recueil du corpus, le lieu et la durée du recueil, mais aussi le procédé de recueil (support d'écriture, mode d'enregistrement, etc.).
2. Le recueil ou la collecte des données. Au cours de cette étape, les différentes productions langagières qui composent le corpus sont rassemblées. Selon les types de données recueillies et les projets, cette étape peut nécessiter de se déplacer physiquement sur le lieu de la situation de communication (corpus oraux en milieux professionnels, corpus scolaires, etc.), de réaliser un travail de fouille et d'archivage à postériori de la situation de communication (corpus journalistiques, corpus d'écrits de Poilus, etc.), ou encore de réaliser un travail de captation et de sélection des données au moment de leur réalisation et de leur diffusion ou peu de temps après (corpus de tweets, d'échantillons radiophoniques, etc.).
3. La numérisation des données. Au cours de cette étape, les données brutes sont transformées en données numériques, manipulables à l'aide d'outils informatiques et, le plus souvent, dans des formats standards reconnus par différents outils (*TXM*, *Toolbox*, etc.) ou par différentes ressources (*ELRA*, *Ortolang*, *Pangloss*, etc.). Cette étape est fondamentale tant pour l'exploitation que pour la diffusion d'un corpus.

4. L'enrichissement des données. Au cours de cette étape, de nombreuses informations peuvent être ajoutées au corpus en vue d'une exploitation plus fine ou plurielle de celui-ci. Tous les corpus ne présentent pas d'enrichissements ou d'annotations ; néanmoins lorsque ceux-ci sont présents, ils représentent une réelle plus-value pour le corpus et sa diffusion.

L'étape de recueil est très différente d'un projet à l'autre ; elle a déjà été abordée pour les corpus scolaires dans le chapitre 3, nous ne la détaillons pas plus ici. Les choix de recueil effectués pour le corpus *Scoledit* sont explicités au chapitre suivant. En revanche, les enjeux sous-tendus par les étapes de numérisation et d'enrichissement méritent d'être explicités et seront développés et discutés au cours de ce chapitre. Nous verrons que la frontière entre ces deux étapes peut être parfois poreuse et nous circonscrirons le contenu de chacune de ces étapes dans le cadre du projet *Scoledit*.

1. Numérisation d'un corpus

Comme son nom l'indique, l'objectif de l'étape de numérisation est de transformer les données brutes en données numériques. Cette étape répond à plusieurs objectifs. Premièrement, il s'agit de sauvegarder et de pérenniser les données. Deuxièmement, la numérisation des données facilite leur partage, puisqu'une fois numérisées, elles peuvent être aisément répliquées et partagées en ligne. Enfin, numériser les données les rend exploitables informatiquement. Différents traitements peuvent être nécessaires pour cela :

- La modification du type de données. Les données brutes peuvent être sous des formes très diverses : enregistrements sonores ou télévisuels, copies d'élèves, cartes postales, livres de recettes, livres, manuscrits, documents au format PDF, pages html, etc. Dans de nombreux cas, une transformation du type de ces données est nécessaire, de données de type audio, vidéo ou iconographiques vers des données textuelles, exploitables informatiquement. Selon le type de données manipulées, il peut s'agir de transcrire des données orales, multimodales, ou manuscrites, de scanner des documents au format papier, d'utiliser des techniques de reconnaissance optique de caractères pour reconnaitre des textes imprimés, etc.
- La modification du format de données[48]. Il existe de nombreux formats textes (texte brut, texte mis en forme à l'aide de logiciels de traitement de texte, texte structuré avec

[48] Précisons que certains auteurs appellent cette étape l'étape de normalisation (Habert, Fabre, & Issac, 1998, par exemple). Le terme *normalisation* renvoie à l'idée d'encoder et de normaliser des données numériques dans un

des balises *XML*, etc.). Selon les outils d'exploitation et de diffusion visés, les formats admis ne sont pas les mêmes (par exemple l'outil *TXM* autorise tous types de balises *XML*, tandis que la plateforme *Ortolang* exige l'usage du standard *XML-TEI*). Le corpus doit donc être établi dans un format adéquat.

- La standardisation des données. Certaines données textuelles peuvent contenir des phénomènes non interprétables par les outils informatiques, à l'exemple des émojis ; une phase de standardisation de ces phénomènes peut donc être nécessaire. Selon les auteurs, cette phase peut être considérée comme une phase de numérisation ou d'enrichissement ; elle sera présentée plus en détail dans la section dédiée à l'enrichissement et l'annotation des corpus (section 2 de ce chapitre).

En ce qui concerne les corpus scolaires, la numérisation consiste généralement en deux étapes :

1. Le scan des productions, ce que P.-Y. Testenoire (2017) appelle « l'édition fac-similé », dont résulte une représentation photographique de la production qui donne à voir ses caractéristiques visuelles ;
2. La transcription des productions, dont résultent une ou plusieurs représentations textuelles de la production qui donnent à voir ses caractéristiques linguistiques.

L'étape de scan des productions

Cette étape peut généralement avoir un quadruple objectif. Premièrement, il s'agit de sauvegarder et de pérenniser les données. Deuxièmement, la version scannée facilite la transcription des copies : les scans sont directement accessibles sur écran et plus aisément transportables. Troisièmement, cette étape permet de donner à voir et de partager un aperçu des copies. Enfin, la version scannée permet un retour à la copie d'origine et offre ainsi une iconicité maximale (Boré & Elalouf, 2017).

L'étape de transcription

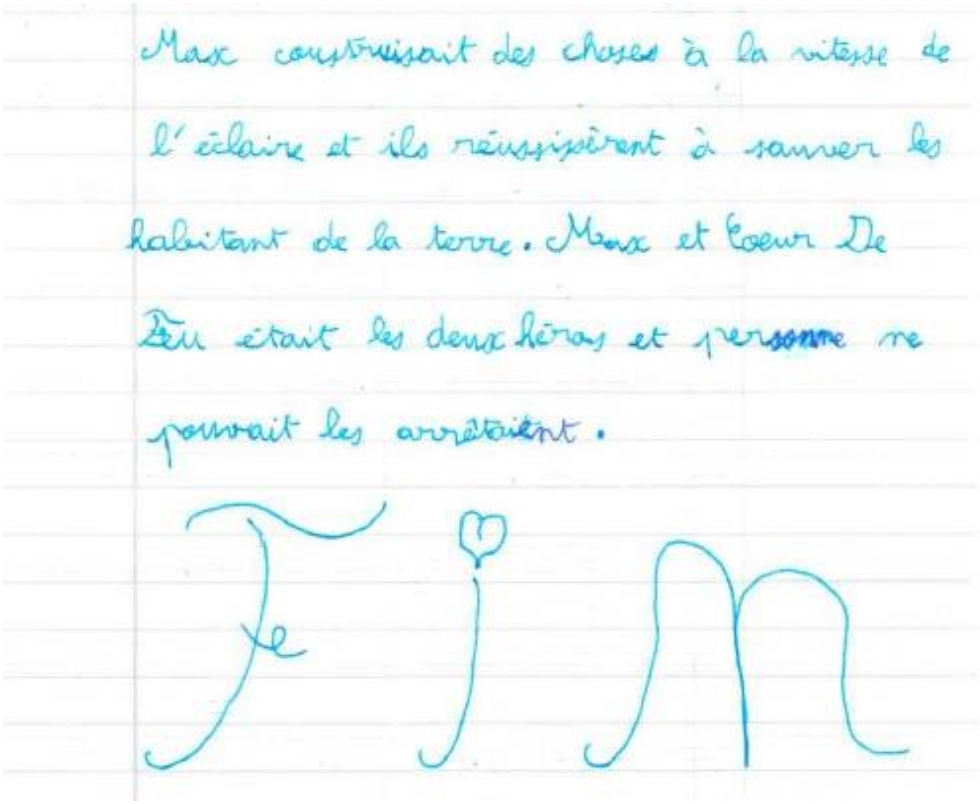
L'objectif principal de cette étape est la production d'une reproduction dactylographiée du texte original de l'élève. La plupart du temps, cette étape est manuelle et consiste en un travail de copie, les systèmes de reconnaissance optique de caractères ne permettant pas, à l'heure actuelle, de traiter les écrits de jeunes apprenants.

format commun. Bien que nous soyons en accord avec cette définiion de la notion, nous n'emploierons pas ce terme que nous réservons à la standardisation orthographique, morphologique et autre de notre corpus.

1.1. Enjeux de la transcription

L'objectif de l'étape de transcription est donc de reproduire numériquement le texte des scripteurs. L'ensemble des traitements et analyses qui sont effectués sur le corpus sont effectués sur ces transcriptions. Il est donc crucial de maitriser au mieux ce processus qui soulève deux problèmes principaux : *que transcrire ?* et *comment transcrire ?*

Il n'est pas possible de reproduire l'ensemble des informations contenues dans la version manuscrite des productions (couleurs des encres, formes des lettres, etc.), sous peine d'obtenir une transcription complètement illisible pour un lecteur non expert d'une part et de rendre le processus de transcription déraisonnablement couteux pour les transcripteurs d'autre part, sans compter les informations visuelles non reproductibles, comme les dessins ou les flèches.



Transcription : […] Max coustruisait des choses à la vitesse de / l'éclaire et ils réussissèrent à sauver les / habitant de la terre. Max et Coeur De / Feu était les deux héros et per<revision/>sonne ne / pouvait les arrêtai<revision/>ent. // F<dessin/>in

Normalisation : […] Max construisait des choses à la vitesse de l'éclair et ils réussirent à sauver les habitants de la terre. Max et Coeur De Feu étaient les deux héros et personne ne pouvait les arrêter. Fin

Figure 2 : Production de texte en fin de CM2 de l'élève 2973

Dans l'exemple de la figure 2, il n'est pas possible de rendre compte de tous les phénomènes iconiques : par exemple les changements de couleurs d'encre entre le texte initial et les révisions, la présence d'un cœur sur le « i » du segment « Fin », les variations de taille de caractères entre la dernière ligne et le reste du texte, etc., sans alourdir la transcription de manière conséquente.

Des choix doivent donc être effectués au cours de l'étape de transcription. Dès lors, les chercheurs se retrouvent « tiraillés entre deux exigences » (Roubaud, 2017, p. 118) :

- la reproduction des **caractéristiques visuelles signifiantes** du texte d'origine, que J. Anis (1983, cité par Testenoire, 2017) appelle la *vi-lisibilité* et qui permet de rester fidèle au texte et à la mise en page de l'auteur ;
- la **lisibilité** et la **compréhension** de la transcription pour les lecteurs.

Ces choix s'effectuent en cohérence avec les perspectives d'étude du corpus. Ainsi, la transcription d'un même corpus peut être très différente selon les questions de recherche pour lesquelles ce corpus est élaboré.

1.1.1. Comment transcrire ?

La génétique du texte distingue trois types de transcription (Testenoire, 2017) :

- La **transcription diplomatique**. Il s'agit de la transcription la plus iconique. Elle vise la reproduction mimétique d'une production à l'aide de procédés dactylographiques, comme l'usage de caractères barrés en cas de rature et la reproduction de l'emplacement de chacun des éléments de la page. Elle peut être considérée comme la transcription la moins interprétative.
- La **transcription linéaire**. Dans cette transcription, l'ensemble des évènements (ajouts, ratures, retours à la ligne, etc.) sont placés sur une ligne continue et un codage spécifique (utilisation de symboles ou de balises) permet de les mentionner. Ce procédé convient donc particulièrement à l'application de méthodes automatiques mais nécessite une certaine part d'interprétation (Lebrave, 1990, cité par Testenoire, 2017) (par exemple « le » ou « la » ou encore « révéié » ou « rivéié » dans la figure 3).
- La **transcription chronologique**. Cette transcription qui vise à rendre visibles les différentes étapes de rédaction est quasi exclusivement utilisée pour les projets de génétique textuelle. Elle explicite, par un système de strates, les différents moments d'écriture ou de réécriture et reconstitue ainsi les étapes chronologiques de la réalisation du texte.

Prenons l'exemple de la production de CP de l'élève 1637 (Figure 3).

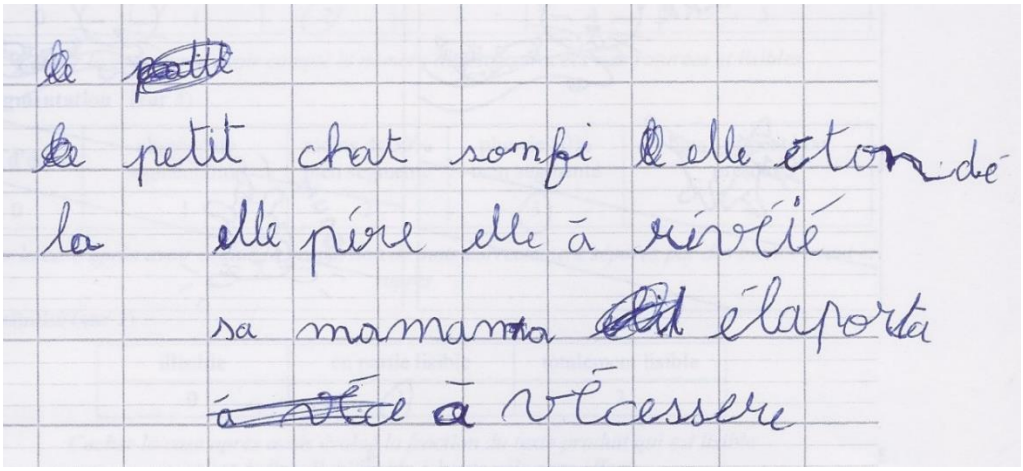


Figure 3 : Production de texte en fin de CP de l'élève 1637

Une transcription diplomatique de cette production pourrait être :

~~le~~ ~~patit~~

~~le~~ petit chat sonfi ~~l~~ elle étondé

la elle pére elle á révéié

sa maman~~ma~~ ~~élit~~ élaporta

~~á véce~~ à vécessere

Pour cette même production, une transcription linéaire pourrait être :

<rature>le</rature> <rature>patit</rature> // <rature>le</rature> <ajout>la</ajout> petit chat sonfi <rature>l</rature> elle étondé / elle pére elle á révéié / sa mama<rature>m</rature>na <rature>élil</rature> élaporta / <rature>á véce</rature> à vécessere

La transcription chronologique[49] permet de matérialiser les différents moments de révision d'une production. En ce qui concerne les productions de notre corpus, les révisions ont été faites soit en même temps que la rédaction initiale, soit très peu de temps après. Ces différents temps sont donc difficiles à dissocier et il est peu pertinent d'essayer d'élaborer une transcription chronologique d'une de ces productions.

[49] Pour un exemple de ce type de transcription, se reporter aux exemples donnés par Y. Testenoire (2017, p. 96-97).

Il y a encore peu, il n'existait aucun consensus de transcription des écrits scolaires et les projets qui s'y rapportaient relataient cette diversité de transcription. Certains projets incluaient même cette diversité en leur sein.

Dans leur article de 2017, par exemple, C. Boré et M.-L. Elalouf, proposaient deux transcriptions différentes : à la fois une transcription iconique des productions (Figure 4), proche de la transcription diplomatique, et une version commentée (Figure 5), proche d'une transcription linéaire. Le choix du type de transcription est effectué selon la nature des textes à transcrire, selon qu'ils contiennent ou non des interventions de l'enseignant.

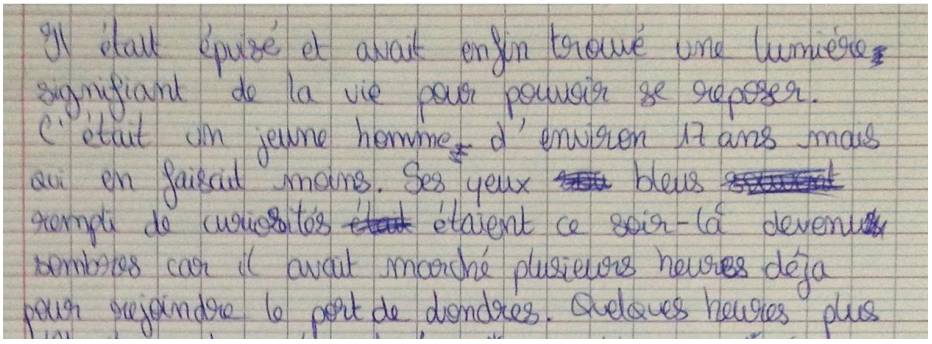
Sur le terrain de foot <du haut>, je vois :
A ma droite il y a des arbres à coté du grillage qui mène à la
cour du bas.
à ma gauche il y a le gymnase
devant moi il y a les classes
j'entent les maternel qui joue.
~~je sens la pollution~~ à toute heure et à midi on sent se quont va
manger

Figure 4 : Exemple de version iconique (Boré & Elalouf, 2017, p. 36)

Alors le joueur de flûte quitta la ville et tous les enfants
le suivirent
Un matin, [virgule rajoutée par le professeur (P)] il se
leva tôt en passant par la ruelle, avec sa flûte et les enfants [? de
P en marge]. Il joua un petit morceau et les enfants en [chantaient
souligné par P]. Une ou deux heures après 15 marke. [?
de P en marge, point ajouté par P]. [Il "l" surchargé par P en
"i"] se dit [« cet pa male » "cet" souligné par P, "s" rajouté à
"pas" par P, "e" biffé par P] et ils se [partagère souligné par
P] la mise [? de P en marge]. Et il [rentrer souligné par P] chez
eux [chaqun souligné par P]. Le lendemain matin il alla autre
[par "t" rajouté par P]. Il resta deux [heure souligné par P] et
[ses souligné par P] l'[exploi "t" ajouté par P, ? de P] 70 mark
[point virgule rajouté par P] il se [partaga souligné par P] la
somme et tout le monde est [contemps souligné par P] [? rajouté
par P au-dessous].

Figure 5 : Exemple de version commentée (Boré & Elalouf, 2017, p. 37)

D'autres ont fait le choix de mêler diverses caractéristiques issues de ces différents types de transcription en une seule transcription, parfois appelée transcription semi-diplomatique (Testenoire, 2017). Par exemple, la transcription (Figure 6) adoptée pour le corpus proposé par l'équipe de recherche *Ecriscol* et présentée en 2017 (Doquet *et al.*, 2017) pouvait être vue comme s'approchant d'une transcription diplomatique dans la mesure où elle respecte la mise en page choisie par l'auteur du texte : les retours à la ligne et certains alinéas sont reproduits. Mais elle se rapproche également d'une transcription linéaire dans le sens où les évènements sont souvent formalisés à l'aide de symboles ou d'annotations. Les ratures ne sont pas reproduites à l'aide de la typographie barrée mais sont identifiées par l'usage de crochets. Les informations chronologiques sont également insérées dans cette transcription et ne font pas l'objet d'une transcription particulière.

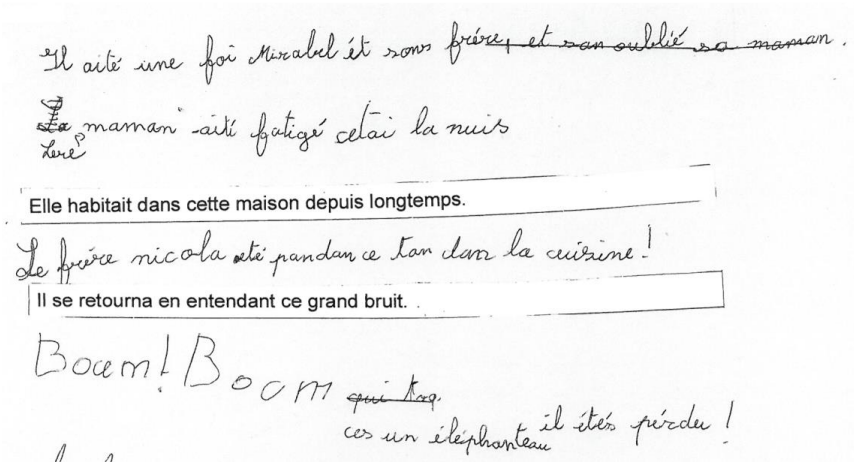


Transcription :

Il était épuisé et avait enfin trouvé une lumière [VIRGULE]
signifiant de la vie pour pouvoir se reposer.§
C'était un jeune homme [VIRGULE] d'environ 17 ans mais
qui en faisait moins. Ses yeux [#XXX#] bleus [#XXX#]
<rempli>_<remplis> de curiosités ®[était]// étaient//® ce soir-là devenu[s]
sombres car il avait marché plusieurs heures <déja>_<déjà>
pour rejoindre le port de Londres. Quelques heures plus

Figure 6 : Exemple de transcription du projet Ecriscol (Doquet et al., 2017, p. 144)

C. Garcia-Debanc, L.-M. Ho-Dac, M. Bras et J. Rebeyrolle, (2017) proposaient également une transcription (Figure 7) que l'on pourrait qualifier de semi-diplomatique dans la mesure où les retours à la ligne sont marqués visuellement, tandis que les révisions et l'insertion de fragments de texte contraints sont marqués par des balises.



Transcription :

Il aité une foi Mirabel ét sons frére <del rend="strikethrough">, et san oublié sa maman</del>.<lb/><del rend="hatched">La </del><add place="below">Lere</add> maman aité fatigé cetai la nuis<lb/>

<surface attachment="glued">

<zone>Elle habitait dans cette maison depuis longtemps. </zone>

</surface>

Figure 7 : Exemple de transcription (Garcia-Debanc et al., 2017, p. 164)

1.1.2. Que transcrire ?

Les possibilités de transcription sont donc variées et plusieurs questions se posent : quels éléments transcrire ? Comment les transcrire ? Souvent, ces choix sont guidés par les objectifs d'étude des projets mais aussi par l'historique personnel des porteurs de projet. Par exemple, M.-N. Roubaud (2017), qui a participé à la mise en place de conventions pour la transcription de l'oral au sein du *Groupe aixois de recherches en syntaxe* (GARS), propose des éléments de transcription analogues aux conventions adoptées pour la transcription de l'oral, tant dans le format de codage que dans les phénomènes codés. Les syllabes indéchiffrables sont notées * ou *** par analogie avec la transcription des syllabes incompréhensibles de l'oral à l'aide de ces mêmes symboles. Une transcription des amorces graphiques est également proposée, par analogie avec les amorces phoniques, phénomène relativement fréquent à l'oral, alors même que ce phénomène ne semble pas particulièrement prépondérant dans les écrits scolaires. Enfin, la transcription étant réalisée dans la même perspective que la transcription de l'oral, une priorité est donnée à l'exploitation ; les graphies n'y sont donc que mentionnées [entre crochets] mais c'est la graphie normée qui prime. En cela, cette transcription est déjà une annotation et introduit une certaine interprétation.

L'équipe du projet *Ecriscol* (Doquet *et al.*, 2017) choisit en revanche d'adopter un codage similaire à celui retenu par l'Institut des Textes et Manuscrits (ITEM) dont l'attention porte

notamment sur des manuscrits d'écrivains tout en mettant l'accent sur les traces de révision des textes, en cohérence avec leur approche génétique des textes. Dans ce projet, les chevrons permettent de signifier <les segments ajoutés> et les crochets marquent [les segments supprimés], à l'instar du codage proposé par l'ITEM. La dimension génétique étant fondamentale dans le projet *Ecriscol*, les initiateurs de ce projet choisissent également de spécifier à travers un codage spécifique un changement de scripteur (l'enseignant ou l'enseignante dans la plupart des cas) ou un changement de moment d'écriture (David & Doquet, 2016).

Avant le projet *Ecriscol*, C. Fabre (1990) s'inscrivait déjà dans une perspective génétique à travers son étude des brouillons d'écoliers. Tout comme dans ce projet, l'accent était mis sur les traces de révision des textes, que ce soient les ajouts, les éléments supprimés ou les éléments déplacés.

Un élément important également des écrits scolaires est la présence dans de nombreux écrits de marques de correction de la part des enseignants. Généralement, la présence ou non de ces marques dépend de la nature des écrits : s'agit-il de brouillons ou de textes finaux ? Le recueil était-il écologique ou provoqué par la recherche ? En d'autres termes les textes sont-ils produits pour des besoins pédagogiques et sur décision de l'enseignant ou initiés par des chercheurs ? Les textes ont-ils été retravaillés en classe ? Lorsque ces révisions, notes ou remarques sont présentes, il convient de les transcrire de telle sorte qu'elles soient identifiables comme non produites par l'élève, ce que font la plupart des projets de transcription.

De nombreux éléments peuvent être reproduits lors de l'étape de transcription : les éléments de mise en forme, les marques de paragraphe, les traces de révisions, les interventions de l'enseignant, parfois même certaines informations orthographiques et certaines autres informations iconiques comme les changements de stylo ou de tracé. Néanmoins, certains éléments échappent à toute transcription, notamment les fléchages, régulièrement utilisés pour réviser un texte, les tableaux ou encore les dessins ; ils peuvent être signalés mais difficilement reproduits. Face à cette perte d'information inévitable, C. Garcia-Debanc et ses collègues (2017) préconisent un retour systématique à la version papier ou scannée de la production avant toute analyse.

Toute transcription est donc un processus au cours duquel s'effectuent à la fois une perte d'information, notamment une perte des informations visuelles et manuscrites, et un ajout d'information, puisqu'elle introduit une certaine part d'interprétation et d'analyse (Testenoire, 2017). A. Grésillon (1994) note ainsi que toute transcription est à la fois plus riche et plus pauvre

que le manuscrit dont elle est issue. Bien que non désirés, ces deux processus sont, comme nous venons de le voir, inévitables.

Un des objectifs du projet *Scoledit* est l'exploitation informatique et automatique des écrits d'élève. La transcription adoptée au sein du projet doit donc être en conformité avec les contraintes liées à cet objectif. Cela signifie qu'aucune information ne peut être relevée par des procédés typographiques, qui essaient de reproduire l'iconicité de la production, comme les retours à la ligne, le soulignement de caractères, etc. Les informations à relever doivent l'être par des marqueurs établis par convention, souvent des symboles ou des balises, qui peuvent complexifier la lecture du texte mais présentent l'avantage d'être identifiables informatiquement. Par exemple, les conventions de transcriptions de l'ITEM proposent deux notations concurrentes pour identifier les ratures (David & Doquet, 2016) : une notation typographique ~~rature~~ et une notation symbolique [rature] ; seule la notation symbolique pourra être aisément identifiable par un outil de traitement automatique. C'est donc cette notation symbolique, utilisant des marqueurs conventionnels, que nous avons adoptée au sein du projet *Scoledit*.

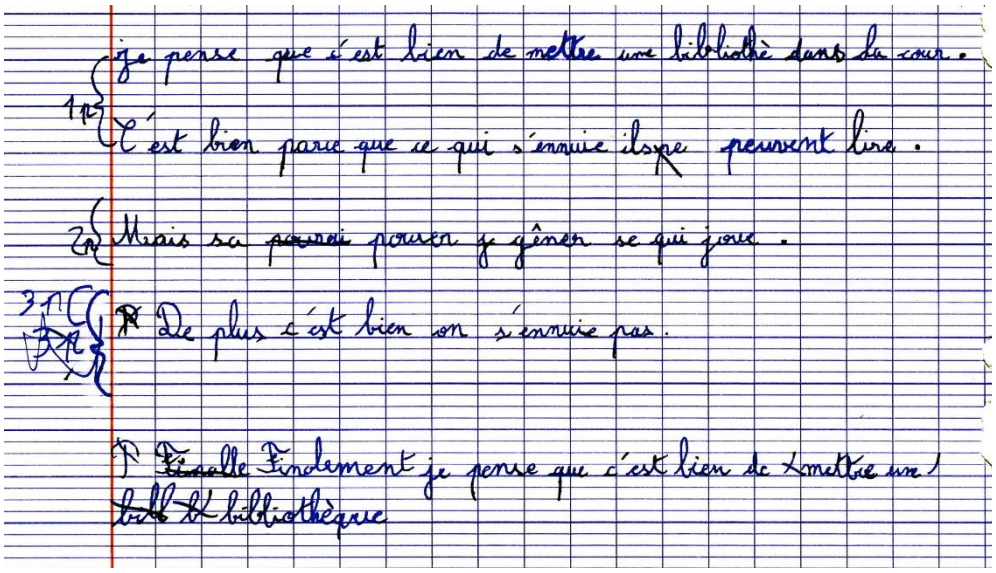
La transcription du corpus *Scoledit* est principalement une transcription linéaire, exploitable par des outils informatiques. De plus, en accord avec les perspectives d'étude du projet, à savoir la description linguistique des écrits d'élèves, la transcription effectuée vise principalement à rendre visible le contenu textuel de la production finale de l'élève, indépendamment des aspects génétiques (révision du texte par ajout ou suppression par exemple) et visuels (sauts de ligne, changement de stylo, etc.) des productions, ces aspects y sont donc peu détaillés.

1.2. Réflexion nationale autour des conventions de transcription

Alors même que le nombre de corpus scolaires transcrits et diffusés ou en cours de diffusion est relativement restreint, les conventions de transcription de ces écrits sont nombreuses et un même codage peut traduire plusieurs réalités différentes. Par exemple, l'usage de crochets dans les conventions adoptées par M.-N. Roubaud permettent de signaler une graphie non normée, tandis que dans les conventions adoptées par l'équipe du projet *Ecriscol*, il traduit une rature.

Face à ce constat et à cette diversité des pratiques, une réflexion nationale autour d'une convention de transcription commune a émergé suite à la journée d'étude *Analyse linguistique de grands corpus d'écrits scolaires, problèmes de transcription, d'annotation et de traitement*, organisée par le groupe *Écriture Scolaire* du laboratoire *Clesthia* le 18 mars 2015. Cette réflexion a abouti à l'écriture de conventions pour la transcription de corpus scolaires (disponible en Annexe 2). Suite à ce travail, C. Boré, M.-L. Elalouf et M.-N. Roubaud ont adopté

les conventions de ce manuel pour le corpus ÉMA (pour un exemple, voir Figure 8), modifiant ainsi les choix présentés précédemment. Le processus d'élaboration de conventions communes se poursuit désormais au sein du projet *E-CALM* financé par l'ANR.



Transcription :

1)Je pense que c'est bien de mettre une bibliothè dans la cour. £

1)

C'est bien parce que ce qui s'ennuie ils [pe] <peuvent> lire. §

Mais sa [pourai] <pouver> [j] <Gêner> se qui joue. §

De plus c'est bien on s'ennuie pas. §

[F] [Finalle] Finalement je pense que c'est bien de mettre une £ [bill] [b] bibliothèque. §

Figure 8 : Exemple de transcription du corpus ÉMA

Dans un souci de se conformer à un format déjà existant pour de nombreux corpus écrits et compatible avec la plateforme *Ortolang*, le format *XML-TEI* a été choisi. Le *XML* est un langage de description et de structuration sous forme de balises. Les balises sont choisies parmi le répertoire de balises définies par la *TEI* (*Text Encoding Initiative*) qui propose un grand nombre de balises adaptées aux documents textuels. En l'occurrence, les balises ont été choisies de telle sorte qu'elles correspondent aux besoins des transcriptions des corpus scolaires issus des différents projets de recherche.

Les réflexions pour la transcription de notre propre corpus ayant eu lieu en même temps que cette réflexion nationale, il ne nous a pas été possible de les adopter pour notre projet. Cependant, dans le formalisme que nous avons choisi, un mélange de balises et de symboles, les éléments transcrits et annotés sont suffisamment similaires pour que les deux formalismes soient interopérables sans trop de difficultés. Cela signifie qu'il est possible de passer de notre

convention de transcription à la convention de transcription commune et inversement de façon automatisée.

2. Enrichissement et annotation d'un corpus

En 1997, R. Garside, G. Leech et A. McEnery décrivent le processus qui aboutit à un enrichissement des corpus comme un processus en trois étapes. Au cours de la première étape, le corpus est recueilli, enregistré ou récupéré et stocké électroniquement. Puis vient la seconde étape où émerge la volonté d'extraire de l'information de ce corpus, que ce soit pour l'analyser ou pour construire des ressources telles que des dictionnaires ou des grammaires. Au cours de la troisième étape apparait alors le besoin d'ajouter des informations linguistiques explicites pour répondre aux volontés exprimées lors de la seconde étape. Cette phase d'ajout est appelée phase d'enrichissement du corpus.

Deux types d'enrichissement peuvent être distingués : les enrichissements extralinguistiques, généralement stockés sous forme de **métadonnées**, et les enrichissements linguistiques, qui prennent souvent la forme d'**annotations**.

Les métadonnées permettent généralement d'apporter des informations contextuelles (conditions de recueil par exemple), des informations sur les locuteurs et / ou scripteurs ou encore des informations sur les types de traitement opérés sur les données.

Les enrichissements linguistiques prenant souvent la forme d'annotations, la phase d'enrichissement linguistique est souvent appelée phase d'annotation du corpus. Cette dernière appellation est la plus courante dans la littérature (Garside *et al.*, 1997). Pour T. McEnery et A. Wilson (2001), l'étape d'annotation est l'étape qui permet de rendre explicites des informations jusqu'alors restées implicites. Pour reprendre les termes de C. Poudat et F. Landragin (2017, p. 36), « une annotation est un ajout qui vient repérer, souligner, expliciter, clarifier, définir, commenter ou interpréter » un élément du texte initial. Cette étape, qui permet l'ajout d'informations de nature interprétative (Leech, 1997, cité par Pierrel, 2000), apporte donc une réelle plus-value au corpus (Pierrel, 2000 ; McEnery & Wilson, 2001).

Ainsi, dans le cadre d'un corpus annoté, le matériau observable sur lequel s'appuie l'analyse linguistique est composé à la fois du texte – brut ou transcrit –, des métadonnées et des annotations, encore appelées enrichissements (Poudat & Landragin, 2017). L'étape d'annotation des corpus n'est donc pas si triviale puisqu'elle est extrêmement dépendante de la phase d'exploration et des phénomènes linguistiques qu'on souhaite y observer. C'est pourquoi, la décision de la nature des enrichissements se fait généralement à partir de l'observation du corpus.

La phase d'enrichissement et d'annotation nécessite deux éléments principaux :

- un **modèle d'annotation**, déterminé par l'observation du corpus et selon les objectifs de recherche ;
- une **méthodologie,** outillée ou non, **d'annotation** du corpus, qui doit permettre un enrichissement homogène du corpus afin de faciliter son exploitation.

L'étape d'annotation ou d'enrichissement peut être réalisée de manière automatique ou semi-automatique, à l'aide de logiciels d'annotation, ou de manière manuelle. Cependant, malgré l'existence de ces outils, annoter est rarement une activité complètement automatique (Née, 2017). Cette étape a donc un cout non négligeable, ce qui limite à la fois le nombre de corpus annotés (Pierrel, 2000) mais également la finesse et la richesse des informations explicitées.

2.1. Types d'enrichissements linguistiques

Comme nous l'avons mentionné au chapitre 2, il existe différents types d'enrichissement et les informations ajoutées peuvent être de natures multiples. Pour la plupart des corpus, la première étape d'annotation effectuée est un découpage en unités lexicales associé à un étiquetage morphosyntaxique au cours duquel une catégorie grammaticale est associée à chaque forme[50] du corpus. Au cours de cette phase, un lemme peut également être attribué à chaque forme, on appelle cela la lemmatisation.

Un enrichissement peut prendre des formes très variées. Le plus souvent, il s'agira d'annotations introduites à l'intérieur du corpus qui portent sur des formes (étiquetage morphosyntaxique par exemple) ou des groupes de formes (étiquetage syntaxique par exemple).

Plus généralement, on distinguera à l'instar de C. Poudat et F. Landragin (2017) et de B. Habert (2005) les enrichissements intégrés ou embarqués des enrichissements déportés ou débarqués. Les premiers sont insérés directement dans le texte qui compose le corpus (un exemple est donné dans la figure 9). L'avantage de cette méthode est qu'elle permet de travailler sur le corpus, sans traitements préalables à l'aide d'outils d'exploration de corpus qui prennent en compte les balises d'annotations (*TXM* pour ne citer qu'un exemple). En revanche, cette méthode conduit rapidement à une annotation très lourde et peu lisible pour un lecteur humain.

[50] La notion de *forme* est définie dans le Glossaire.

Transcription :
<P#Avez-vous envie de retravailler le texte de votre autoportrait ? Pourquoi ?> §
J'ai envie de retravailler le texte de mon autoportrait £
car j'ai fais [des] un texte cour et je veux s' £
ameliorer §

Normalisation :
//P#Avez-vous envie de retravailler le texte de votre autoportrait ? Pourquoi ?// §
J'ai envie de retravailler le texte de mon autoportrait £
car j'ai <fais>_<fait> [des] un texte <cour>_<court> et je veux <s'-£ameliorer>_<s'améliorer> §

Figure 9 : Exemples d'annotations embarquées, corpus Ecriscol

Dans le second cas, les enrichissements sont encodés sur des fichiers séparés, à l'exemple des transcriptions orthographiques de l'oral élaborées dans de nombreux projets de corpus oraux, comme le *Corpus of Interactional Data* (CID, Bertrand *et al.*, 2006). Ce type d'enrichissement nécessite d'introduire un moyen pour relier les informations ajoutées aux éléments du texte ou du signal sonore ou visuel original. Le lien peut se faire via des coordonnées d'annotations qui identifient les séquences de caractères ou les séquences sonores auxquelles les enrichissements s'appliquent, comme c'est le cas pour les corpus annotés à l'aide de la plateforme *Glozz*[51]. Le lien peut également se faire via une étape d'alignement forme à forme ou phrase par phrase comme c'est souvent le cas pour les corpus multilingues. Cette dernière méthode permet d'envisager plusieurs couches d'annotations, souvent indépendantes les unes des autres, sans alourdir la lecture du corpus.

2.2. La transcription est-elle une annotation ?

Nombreux sont les corpus qui nécessitent une transcription, c'est-à-dire une représentation des données brutes, afin de pouvoir accéder informatiquement aux données linguistiques. C'est le cas par exemple des corpus oraux et multimodaux, qui nécessitent que soient transcrits textuellement la parole, les gestes ou les actions afin de pouvoir en extraire de l'information. C'est également le cas des corpus composés de scans d'ouvrages non disponibles en version numérique (le *Dictionnaire de l'Académie Française*[52] par exemple) ou encore des corpus

[51] Widlöcher, A., & Mathet, Y. (2009, June). La plate-forme Glozz : environnement d'annotation et d'exploration de corpus. http://glozz.free.fr/ [consulté le 09/08/2019]

[52] Le Dictionnaire de l'Académie Française qui prévoit l'intégration des 9 éditions dans sa nouvelle édition numérique, mais seules les deux dernières versions seront accessibles textuellement.

manuscrits, que ce soient des brouillons d'auteurs (les manuscrits de Stendhal[53] par exemple) ou des productions écrites manuelles d'élèves, comme dans notre corpus.

Selon les auteurs et les projets, cette transcription peut être considérée soit comme une première étape d'annotation, soit comme le texte brut à partir duquel le corpus sera étudié, donc inhérent au corpus. G. Leech (1997, cité par Pierrel, 2000) distingue ainsi l'annotation de la représentation. Dans cette vision, la représentation présente le contenu textuel et linguistique, tandis que l'annotation apporte des informations métalinguistiques, à propos de la langue du texte (Garside *et al.*, 1997). La transcription tend donc davantage à être considérée comme une représentation. G. Leech reconnait toutefois que cette distinction peut ne pas être aussi évidente, notamment dans le cas de corpus oraux, où la transcription peut contenir une grande part d'interprétations. Il va néanmoins considérer l'enregistrement sonore comme les données primaires et la transcription brute comme des données secondaires. Appliqué au domaine des corpus d'écrits scolaires, cela reviendra à considérer les écrits manuscrits comme les données primaires et les transcriptions uniquement comme des aides à la compréhension.

À rebours de cette vision, J. Véronis considère « tout apport d'information aux données brutes originales » (Véronis, 2000b, p. 112). L'étape de transcription est, pour cet auteur, comprise dans le processus d'annotation. De la même manière, les concepteurs du corpus *of Interactional Data* (CID, Bertrand *et al.*, 2006) envisagent un premier niveau d'annotation constitué de trois éléments : une transcription orthographique enrichie, une représentation phonétique et un alignement entre la représentation phonétique et le signal sonore, puis entre la représentation phonétique et la transcription.

Dans le cadre du corpus *Scoledit*, seul le contenu linguistique des productions nous intéresse, nous faisons très peu de cas de l'iconicité de la production[54]. Nous travaillons donc quasi exclusivement à partir des transcriptions des textes des élèves. Les scans de ces productions ont pour fonction principale d'illustrer les transcriptions, mais aussi d'en permettre une meilleure compréhension et éventuellement de permettre une vérification de ces transcriptions. Nous considèrerons donc que la transcription est une représentation informatiquement exploitable des productions de notre corpus et qu'à ce titre celles-ci font partie intégrante du corpus brut. Ce choix renvoie également à la définition de « corpus » que nous avons adoptée dans ce projet et qui est énoncée au chapitre 1. En raison de ce choix, l'étape de transcription, malgré la part d'interprétation qu'elle comporte, nous semble

[53] Meynard C. et Lebarbé T. http://www.manuscrits-de-stendhal.org/ [consulté le 09/08/2019]
[54] Exception faite des retours à la ligne que nous avons relevés notamment dans le but de permettre un affichage similaire à la production manuscrite et ainsi faciliter la lecture de ces productions.

inhérente à la constitution de corpus et a donc été explicitée au chapitre précédent portant sur la méthodologie de constitution de corpus.

2.3. Normalisation : qu'en est-il de la correction et de la standardisation ?

Dans de nombreux cas, l'exploitation des données d'un corpus nécessite une correction, une normalisation ou une standardisation préalables. On citera notamment les corpus d'apprenants, que ce soit de langue seconde ou de langue maternelle, mais aussi les corpus produits par des scripteurs peu lettrés, à l'exemple du *Corpus14* (Steuckardt, 2014) qui rassemble des lettres écrites par des Poilus ou leur famille pendant la guerre 1914-1918. Sont également concernés les corpus issus de moyens de communication numériques donnant souvent lieu à des échanges peu formels tels que les corpus de SMS, de tweets ou encore de tchat. Pour ces derniers particulièrement, on parlera plus souvent de *standardisation* que de *correction*. En effet, bien que certains phénomènes relèvent d'erreurs orthographiques, d'autres relèvent plus de variations volontaires des scripteurs, à l'exemple des abréviations ou encore des répétions multiples[55] de caractères pour accentuer le propos. Ces écrits présentent également des phénomènes nouveaux, tels que les émoticônes, qu'il est nécessaire de standardiser ou d'identifier afin de pouvoir les exploiter.

Selon les projets considérés, la façon de gérer et d'envisager cette correction n'est pas la même. En effet, tant les formats choisis pour conserver ces informations que le statut donné à celles-ci sont susceptibles de variation.

2.3.1. Statut de la correction

La première question à se poser dans le cas d'ajouts d'informations de correction, normalisation ou standardisation est de savoir quel statut on donne à ces informations. S'agit-il d'un complément d'informations pour le corpus ou est-ce à partir de ces informations que le corpus va être analysé ? Et, pour reprendre les termes de G. Leech (1997, cité par Pierrel, 2000), s'agit-il d'une représentation des données de notre corpus ou d'une interprétation ? En effet, selon les projets, les informations orthographiques ou de standardisation peuvent être vues comme des aides à la lecture pour le lecteur (comme dans le cadre du corpus *ÉMA*), une aide pour le traitement des données ou les seules données sur lesquelles reposera l'analyse (comme dans le cadre de certains projets fondés sur les corpus de SMS). Dans ce dernier cas, le texte initial pourra donc être contenu dans une annotation ou un enrichissement qui sera vu comme

[55] Exemple « son meiller meiller amie » (2439, CE2).

une sauvegarde (Poudat & Landragin, 2017). Le choix de réaliser les analyses à partir d'une version standardisée du corpus est souvent effectué pour des raisons techniques. En effet, comme nous l'avons explicité au chapitre 2 (section 3), les logiciels de traitement automatique étant généralement élaborés à partir de corpus standardisés, ils échouent souvent à traiter des corpus non standardisés.

2.3.2. Quel format pour la correction ?

En second lieu, il est nécessaire de se demander sous quel format conserver ces informations. Les auteurs de certains projets choisissent ainsi d'encoder ces informations à l'aide d'annotations embarquées, à l'exemple des corpus *Ecriscol* (Figure 9), où, pour chaque forme erronée, on observe une annotation donnant à la fois la forme produite et la forme normée. De manière assez similaire, dans le sous-corpus du projet BFM[56] (Guillot, Lavrentiev, & Marchello-Nizia, 2007 ; Heiden, Guillot, & Lavrentiev, 2010) chaque forme est transcrite de plusieurs façons dont une version normalisée (Figure 10).

```
<w type="NOMcom" xml:id="w106_000023">
<choice>
<me:norm>servisse</me:norm>
<me:dipl>seruisse</me:dipl>
<me:facs>&slong;erui&slong;&slong;e</me:facs>
</choice>
</w>
<w type="CONcoo" xml:id="w106_000024">
<choice>
<me:norm>et</me:norm>
<me:dipl><ex>et</ex></me:dipl>
<me:facs><bfm:mdvAbbr>&et;</bfm:mdvAbbr></me:facs>
</choice>
</w>
<lb n="7"/>
<w type="PROind" xml:id="w106_000025">
<choice>
<me:norm>l'en</me:norm>
<me:dipl>l'en</me:dipl>
<me:facs>len</me:facs>
</choice>
```

Figure 10 : Exemple issu du corpus BFM (Heiden et al., 2010)

Dans d'autres projets, le choix a été fait d'une normalisation déportée, à l'instar de différents corpus issus du projet *SMS pour la science* ou *SMS4Science* (Fairon, Klein, & Paumier, 2006, Figure 11).

SALUT COMAN SA VA	Salut comment ça va
ztadorrre trop fort	Je t'adore trop fort

[56] Certains exemples de cette partie sont issus de l'ouvrage *Explorer un corpus textuel* de C. Poudat et F. Landragin (2017). S'y référer pour plus de détails sur ces exemples.

Figure 11 : Exemples issus du corpus SMS4Science (Fairon et al., 2006)

Au sein des corpus scolaires, ce choix avait également été réalisé par M.-L. Elalouf et ses collègues (2005) qui pour chaque écrit recensaient plusieurs versions d'un même texte[57] dont une version orthographiée selon la norme. Cependant, quelques années plus tard, ce choix a été modifié et les chercheurs ont opté pour des annotations embarquées du même type que celles adoptées par *Ecriscol* afin de pouvoir utiliser les logiciels d'exploitation.

C. Poudat et F. Landragin (2017) évoquent aussi des projets comme le projet *Corpus14*, que nous avons déjà mentionné, où le choix a été fait de ne pas apporter de correction mais plutôt de permettre des requêtes suffisamment larges dans l'exploitation pour englober même les cas erronés.

2.3.3. Correction automatique ou manuelle ?

En troisième lieu, il est nécessaire de choisir entre annotations automatiques, semi-automatiques ou manuelles. Dans la plupart des travaux, on opte pour une correction manuelle des productions ou des transcriptions car peu d'outils suffisamment performants existent.

Il est cependant un domaine dans lequel beaucoup de travaux se sont penchés sur la question de la production d'une standardisation, c'est celui des corpus issus des nouveaux modes de communication (SMS, tweets, forums, etc.). Les méthodes utilisées dans ce domaine ont été traitées à la section 3.1. du chapitre 2.

Dans le domaine des corpus scolaires, une première étude avait été réalisée dans le cadre de nos travaux de mémoire de master (Wolfarth, 2015) visant la production de corrections semi-automatiques. Cependant, au vu du fort taux d'ambigüités générées, ces travaux n'ont pas été poursuivis et une autre approche détaillée dans les sections suivantes a été adoptée.

2.3.4. Quelles variations annoter ?

La principale question que nous nous posons est celle de savoir quels phénomènes de variation annoter et à partir de quelle norme les phénomènes de variations sont définis. Comme le mentionnent C. Fairon et ses collègues (2006), définir cette norme peut être particulièrement délicat dans le cadre des projets portant sur les SMS. En effet, celle-ci peut varier selon divers facteurs et notamment selon le registre et la situation de communication (Fairon *et al.*, 2006), très différents selon les types de SMS étudiés. Dans le cadre des corpus scolaires, cette question pourrait sembler plus simple car la norme scolaire, attendue dans ce type d'écrit, est

[57] L'ouvrage *Ecrire entre 10 et 14 ans* (2005) fait état de 4 versions : une version scannée, une version iconique, une version commentée et une version orthographiée selon la norme.

un objet sans doute mieux identifié et à cet égard, on est souvent tenté de corriger ces écrits selon la norme enseignante. Cependant, celle-ci n'est pas unifiée. De plus, comme le signalent J. David et C. Doquet (2016), apposer une correction aux productions risquerait d'altérer l'authenticité de celles-ci et les procédés linguistiques étudiés risquent de ne plus être ceux produits par l'élève apprenant. D'un autre côté, la présence d'un grand nombre d'écarts à la norme et particulièrement de formes erronées empêchent certains procédés automatiques comme l'étiquetage morphosyntaxique. Il est donc nécessaire de normaliser certains écarts et d'associer ces écarts avec leur normalisation afin de pouvoir appliquer des traitements automatiques d'une part et de pouvoir retrouver la version originelle d'autre part. Ce n'est donc pas la norme enseignante qui est nécessaire. L'enjeu est alors de déterminer quelle est cette norme nécessaire, autrement dit : quels écarts doivent être normalisés ?

Dans ce domaine, les choix sont très différents selon les projets et les contextes de recherche. Dans l'exemple précédent, les corpus de SMS, l'objectif de cette standardisation est généralement de permettre d'étudier les procédés linguistiques utilisés dans ce type d'écrit à partir d'outils d'exploration automatique. Les phénomènes standardisés sont donc principalement ceux qui empêchent cette exploration automatique, comme les erreurs d'orthographe, les variations graphiques, abréviations, troncations et allongements de voyelles par exemple, les émoticônes et les symboles, etc. L'objectif n'est souvent pas tout à fait le même lorsqu'on s'intéresse aux corpus d'apprenants. En effet, dans bien des cas, il s'agit à la fois de produire un corpus exploitable par des outils de traitement automatique mais sans perdre de vue les productions originales des apprenants pour pouvoir en étudier les écarts à la norme (définie au préalable). M.-N. Roubaud (2017) explique cette tension entre la volonté d'identifier la norme attendue et la conservation des choix effectués par l'apprenant. Selon le type d'écarts étudiés, les choix de normalisation ne vont pas être les mêmes.

La plupart de ces projets (le projet *ÉMA* et le projet *Ecriscol* par exemple) intègrent une correction orthographique et rétablissent la segmentation en mots qui peut être altérée. C'est là un des seuls consensus. En effet, sur de nombreux points les projets divergent. Par exemple, en ce qui concerne la ponctuation, certains vont faire le choix de n'apposer aucune correction ou modification à la production originale (Roubaud, 2017), d'autres vont choisir de rétablir certaines ponctuations fortes lorsque des indices, comme la présence de majuscules, le permettent (Doquet *et al.*, 2017). Certains auteurs évoquent la possibilité d'autres corrections, morphosyntaxiques ou syntaxiques, par exemple. M.-N. Roubaud (2017) évoque ainsi la possibilité de rétablir le bon usage des pronoms, mais pour des raisons de description de la grammaire des élèves, choisit de conserver les pronoms utilisés par les élèves. Là encore aucun

consensus n'existe et peu de projets présentent des corrections autres qu'orthographiques ou morphographiques.

Pour conclure, les possibilités pour encoder les données de standardisation d'un corpus assez éloigné de toute norme sont nombreuses et la méthode d'analyse du corpus sera très différente selon le format et la méthode de correction choisie. De même, les outils qui pourront être invoqués ne sont pas les mêmes selon qu'ils nécessitent une version linéaire et standardisée des productions ou selon qu'ils peuvent intégrer des annotations sous différentes formes. Nous reviendrons au chapitre suivant sur ces outils et sur la méthode d'exploitation choisie pour le corpus *Scoledit*. Pour l'heure, nous présentons dans la suite de ce chapitre les choix effectués pour l'encodage des informations de correction.

2.4. Approche du corpus *Scoledit*

La première année de recueil du corpus *Scoledit* coïncide avec la première année d'apprentissage formel de l'écriture, à savoir le CP. Les textes recueillis lors de cette année comportent une très grande variété d'écarts à la norme qu'il convient de prendre en compte dans la conception de méthodes d'exploitation du corpus.

Dans un premier temps, le traitement nécessaire à cette tâche a été vu comme une tâche de correction orthographique[58]. Pour ce faire, une liste de formes normées candidates était calculée à partir de différentes listes de formes spécialisées (*Manulex*[59]) ou non (*Lexique 3*[60]). Bien que cette approche ait amené quelques résultats intéressants, on note également un certain nombre de points critiques :

1. Cette méthode permet de détecter uniquement les formes produites non existantes dans les listes de formes utilisées. Elle laisse donc de côté les formes qui, tout en présentant une orthographe existante, n'en sont pas moins erronées, à l'exemple de l'orthographe *chah* utilisée pour la forme *chat*, ou encore un grand nombre d'erreurs de morphographie verbale, par exemple *tomber* au lieu de *tombé*.
2. L'objectif de ces traitements est de produire une liste de formes candidates qui contienne la forme normée attendue. Afin que les probabilités de contenir cette forme

[58] Le détail de cette démarche a été explicité dans un mémoire de master (Wolfarth, 2015). Nous nous contenterons de reporter ici quelques résultats et conclusions importants. Se référer à la section 3.2. du chapitre 2 pour une revue des méthodes de correction orthographique développées en TAL.

[59] Lété, B., Sprenger-Charolles, L., & Colé, P. (2004). Manulex: A grade-level lexical database from French elementary-school readers. *Behavior Research Methods, Instruments, & Computers, 36*, 156-166. http://www.manulex.org [consulté le 06/10/2019].

[60] New, B., Pallier, C., & Ferrand, L. (2005). Manuel de Lexique 3. *Behavior Research Methods, Instruments, & Computers*, 36(3), 516-524. http://www.lexique.org/ [consulté le 06/10/2019].

avoisine les 100 %, il est nécessaire d'utiliser plusieurs ressources, ce qui génère un grand nombre de formes candidates, et par conséquent un fort taux d'ambigüité.

3. Une seconde étape est alors nécessaire : le choix de la forme attendue parmi les formes candidates calculées. Cette étape peut être réalisée manuellement, ce qui entraine un cout humain important, ou automatiquement, mais un fort taux d'ambigüité risque de générer un fort taux d'erreur.

Face aux difficultés rencontrées lors de la mise en œuvre de cette première approche et à l'absence de travaux majeurs mentionnant l'utilisation du TAL pour l'étude de corpus scolaires et donc de l'absence de méthodes adaptées, nous choisissons d'adopter une approche similaire à celle proposée par O. Kraif et C. Ponton (2007) appelée la stratégie « moins-disante ».

Cette approche, développée pour une tâche de détection et d'analyse d'erreurs en ALAO (Apprentissage des Langues Assisté par Ordinateur), consiste à s'appuyer sur un ensemble de technologies relativement fiables en TAL, comme les étiqueteurs morphosyntaxiques, les lemmatiseurs, etc. et la connaissance du contexte de production et des réponses attendues pour produire une détection, puis une analyse des erreurs. À partir de ces ressources, les erreurs sont identifiées par comparaison avec les réponses des apprenants et les réponses attendues. Les technologies de TAL permettent alors d'analyser les erreurs et de produire une rétroaction. On citera par exemple l'utilisation de dictionnaires de formes pour déterminer les erreurs de lemmes ou de traits morphologiques.

Ces traitements sont réalisés de manière déclarative et modulaire pour permettre une évolution facilitée du système et de ses ressources. Cette approche étant développée dans un contexte didactique, elle est conçue davantage comme un traitement semi-automatique où les enseignants ont la possibilité d'intervenir que comme un traitement entièrement automatique. Concrètement, en cas d'ambigüités dans les analyses, le choix est laissé aux enseignants.

Dans le contexte qui nous intéresse, le corpus *Scoledit*, il n'existe pas de réponse préétablie à l'énoncé ; cependant, il est possible de générer un texte de comparaison qui fasse office de « réponse attendue ». Pour cela, nous proposons de produire une version dite « normalisée » de chaque transcription. Ces normalisations, qu'on pourrait définir en première approximation comme des réécritures proches d'un attendu, ont pour objectif de rendre l'écrit plus standard et de se rapprocher de normes d'écriture afin de permettre le traitement de ces productions.

Une fois en possession d'une version transcrite des productions, d'une part, et d'une version normalisée, d'autre part, il est alors possible de comparer ces deux versions, ce que nous appelons *approche par comparaison*.

Nous choisissons donc de ne pas réaliser d'annotations embarquées des formes, c'est-à-dire une annotation *in situ* des formes erronées, mais de réaliser une annotation débarquée de la production, la normalisation des productions.

Notre démarche se rapproche de celle adoptée par K. Berkling et ses collègues (2011) pour réaliser un diagnostic automatique des erreurs orthographiques dans le *Karlsruhe Children Writing Corpus*, un corpus d'écrits scolaires de langue allemande[61] rassemblant des textes libres. Confrontés à l'absence de « target text », un texte attendu auquel se référer, ils décident de constituer manuellement[62] ces versions attendues des textes afin de pouvoir les comparer aux textes produits dans le but de générer une catégorisation automatique des erreurs.

3. L'approche par comparaison : un préambule à l'exploitation

3.1. Genèse de l'approche

De nombreux corpus en linguistique sont annotés *in situ* et manuellement ou automatiquement à l'aide de balises (projets *Ecriscol* et *ÉMA*). Grâce à cette méthode, le lien est immédiat entre la forme produite et son annotation, en l'occurrence sa normalisation. Dans notre approche, nous choisissons de produire une version normalisée des productions ; celle-ci fait office d'annotation déportée et est donc conçue comme une interface entre les transcriptions des productions et les outils d'exploration automatique. Il est donc nécessaire de construire un lien entre ces transcriptions et leurs normalisations permettant d'associer entre elles les différentes unités linguistiques, normalisées et non normalisées. Pour cela, nous nous appuyons sur une comparaison entre les transcriptions des productions et leur normalisation, ce que nous appelons *approche par comparaison*. Cette approche comporte deux avantages principaux :

- adaptabilité à d'autres tâches du milieu scolaire, comme les dictées où la normalisation s'appuie sur le texte dicté, les exercices de réécriture où la forme attendue est connue, etc. ;
- applicabilité à différents niveaux d'exploitation pour comparer différents types d'éléments linguistiques (syntagmes, formes, graphèmes, morphèmes, etc.).

Quel que soit le niveau d'applications, dont nous verrons différents exemples au chapitre 8 et dans la partie 3 qui suit, l'approche est toujours la même et inclut des procédés similaires que nous présentons dans ce chapitre.

[61] Cf. Chapitre 3 pour plus de détails sur ce corpus.
[62] Un essai d'automatisation de ce processus a été réalisé (Stüker et al., 2011).

3.2. Exposition de la méthode

3.2.1. Étapes

L'approche par comparaison permet d'extraire une catégorisation à partir de deux types d'éléments : les éléments transcrits, c'est-à-dire produits par les élèves, et les éléments normés, qui correspondent à un attendu, et au travers d'un alignement entre ces deux types d'éléments. À partir de ces entrées, différentes étapes sont nécessaires :

1. **Les prétraitements** : Les éléments sont mis en forme de façon à pouvoir être traités par les outils de traitement automatique. Il peut s'agir, par exemple, de la suppression de balises. Cette étape peut être facultative.
2. **L'alignement** : Cette étape constitue le cœur de la méthode. L'alignement permet d'associer les éléments transcrits et les éléments normés. Ces éléments peuvent être des formes, des morphèmes, des graphèmes, etc.
3. **Les post-traitements et enrichissements** : Diverses informations peuvent être ajoutées pour les besoins de l'interprétation. Après un alignement des formes par exemple, l'ajout d'étiquettes morphosyntaxiques peut constituer un post-traitement nécessaire.
4. **Les comparaisons et interprétations** : Après alignement de différents éléments linguistiques, il est possible de comparer leur version transcrite et leur version normée en vue de produire des interprétations. Il peut s'agir d'analyser le type d'erreurs produites, le taux de réussite de certains phénomènes, etc. Comme nous le verrons par la suite, les interprétations réalisées sont dépendantes des méthodes utilisées pour aligner les différents éléments.

La figure 12 ci-après résume chacune de ces étapes.

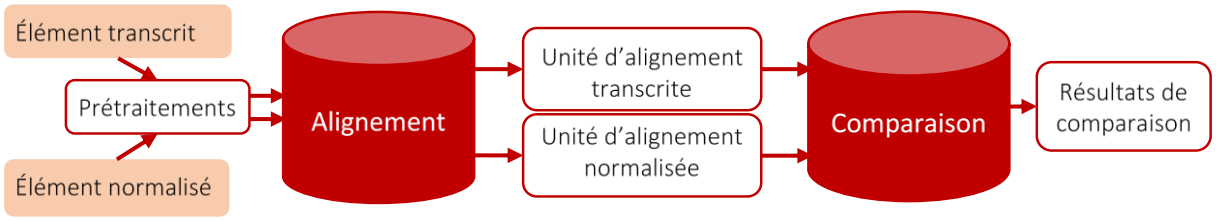


Figure 12 : Approche par comparaison

À l'issue de ce processus, des analyses sont possibles sur trois types de données :

- les **éléments transcrits**, issus de la transcription de la production, fidèles à la production de celui ou celle qui écrit ;
- les **éléments normés**, issus de la normalisation de la production, parfois étiquetés et que peuvent analyser des outils de traitement automatique ;

- les **résultats** issus de la comparaison de ces deux types d'éléments (catégorie d'erreur, par exemple).

3.2.2. Niveaux d'applications

Comme nous l'avons déjà mentionné, cette approche peut être opérée à différents niveaux de traitement. Le tableau 4 exemplifie quelques possibilités. Par exemple, l'approche par comparaison peut être appliquée au niveau du graphème en prenant appui sur un découpage et un alignement préalables des segments transcrits et des formes normées et sur un outil d'alignement en graphèmes. L'approche par comparaison peut permettre alors de réaliser des analyses orthographiques plus fines, à l'échelle des correspondances phonographiques.

Par la suite (cf. Partie 4), nous verrons quelques-uns de ces niveaux d'application.

Niveau de comparaison	*Eléments transcrits*	*Eléments normalisés*	*Prétraitements*	*Unités d'alignement*	*Post-traitements*	*Résultats de la comparaison*
Lexical	Production transcrite	Production normalisée	Suppression des balises de transcription	Formes	Étiquetage morphosyntaxique	Type d'erreurs orthographiques
Morphologique (verbes)	Verbe transcrit	Verbe normalisé	/	Bases et désinences	Étiquetage du temps et du mode	Lieu de l'erreur (base, désinence)
Syllabique	Forme transcrite	Forme normalisée	/	Syllabes		Lieu de l'erreur (attaque, coda)
Graphémique	Forme transcrite	Forme normalisée	/	Graphèmes	/	Type d'erreurs orthographiques
Syntaxique	Production transcrite	Production normalisée	Suppression des balises de transcription	Fragments textuels	/	Compréhension des règles de ponctuation

Tableau 4 : Exemple d'éléments manipulables au moyen de l'approche par comparaison et des résultats attendus

4. Conclusion

Le processus d'élaboration d'un corpus est un processus complexe qui inclut différentes phases : une phase de recueil, une phase de numérisation et une phase d'enrichissement. La façon la plus courante de procéder à l'enrichissement linguistique d'un corpus est de recourir à des annotations. Il s'agit d'un enrichissement embarqué. Cependant, au vu du grand nombre de variations à la norme présentes dans notre corpus, nous faisons le choix de recourir à un enrichissement déporté, à savoir une version normalisée du corpus. Une comparaison entre la transcription du corpus et sa normalisation devrait permettre son exploitation linguistique, c'est ce que nous appelons l'*approche par comparaison*.

Chapitre 5 - Conception et numérisation du corpus *Scoledit*

1. Enjeux et motivations 87
2. Recueil du corpus 89
3. Caractérisation et structure du corpus de textes 101
4. Métadonnées 103
5. Numérisation du corpus 104
6. Diffusion du corpus 117
7. Conclusion 120

Comme nous l'avons mentionné en introduction, le travail présenté dans cette thèse s'inscrit dans le cadre d'un projet plus large appelé *Scoledit* qui vise l'élaboration, la diffusion et l'exploitation linguistique et didactique d'un corpus de textes scolaires outillé, c'est-à-dire accompagné d'outils de traitement et d'exploration. Après une présentation des enjeux qui ont motivé la réalisation d'un tel projet, nous exposons dans ce chapitre la méthodologie adoptée pour la conception du corpus, de son recueil à sa diffusion, sans oublier l'étape cruciale de sa numérisation.

1. Enjeux et motivations

Comme nous avons pu le voir au chapitre précédent, la constitution de corpus scolaires est une des préoccupations qui animent désormais la sphère des linguistes et des spécialistes de la didactique du français et plusieurs corpus d'apprenants émergent au niveau national et international. L'objectif du projet *Scoledit*, et plus particulièrement du corpus éponyme, est de contribuer à cet effort d'élaboration de corpus en vue d'un changement de paradigme dans l'étude de l'apprentissage du français.

Dans ce contexte, la description des caractéristiques linguistiques des textes scolaires produits en français par des élèves de 6 à 11 ans et de leur évolution à partir d'un corpus de textes conséquent reste donc à faire. Un tel travail permettrait d'une part de mettre à la disposition des linguistes des écrits ordinaires d'apprenants, d'autre part de nourrir le travail des didacticiens par la compréhension des dynamiques d'écriture à l'œuvre dans les écrits scolaires. Ce travail permettrait par ailleurs de soutenir, par la constitution d'une banque de textes accessible à tous les professeurs, l'enseignement de l'écriture à l'école. Dans un contexte

où un certain nombre d'observateurs s'accordent à dire que cet enseignement est insuffisant, et ce dès le cours préparatoire, ce travail semble donc nécessaire.

Pour répondre à ces enjeux, nous travaillons à la collecte et à l'édition d'un grand corpus numérique longitudinal de textes narratifs scolaires et de dictées produites par des élèves de 6 à 11 ans (CP-CM2) [63] et rencontrés à plusieurs reprises lors de leur scolarité élémentaire. Dans un premier temps, l'objectif de ce projet est de réaliser une description linguistique des structures utilisées par les élèves au cours de la construction de leurs apprentissages de l'écrit (morphographie, syntaxe, lexique, orthographe, acquisition des conjugaisons, structuration du discours), ainsi que de l'évolution des procédés d'écriture à différents moments de la scolarisation à l'école primaire. Dans un deuxième temps, ce projet devrait également permettre d'élaborer des séquences et des dispositifs didactiques à destination des enseignants.

Tout comme la plupart des corpus scolaires francophones actuellement émergents, le recueil a lieu en milieu scolaire et l'élaboration du corpus est guidée par la volonté de constituer une ressource standardisée[64] suffisamment large de manière à servir d'appui à diverses études en didactique du français. Il se distingue cependant de ces projets par deux caractéristiques essentielles. La première de ces caractéristiques est son caractère unifié, tous les élèves ayant reçu la même consigne, et sa dimension longitudinale, la même consigne ayant été donnée plusieurs années de suite. La deuxième caractéristique, et c'est celle qui fonde notre travail de thèse, est le caractère outillé du corpus *Scoledit*. Cela signifie que, dès sa conception, nous avons fait le choix d'élaborer, en parallèle du recueil du corpus et de sa numérisation, des outils informatiques d'exploitation de ce corpus.

En effet, à terme le corpus encore en cours d'élaboration devrait contenir plusieurs milliers de productions. Un tel corpus ne peut être finement analysé manuellement, c'est pourquoi nous proposons de faciliter cette exploitation à travers différents outils élaborés grâce à des méthodes issues du traitement automatique des langues (TAL), dont un module d'alignement en vue de l'enrichissement des données. Ce module devrait permettre une aide automatique à l'annotation de nombreux phénomènes linguistiques (orthographiques, syntaxiques, lexicaux, etc.), permettant une grande variété d'utilisation du corpus. Par ailleurs, le recours au TAL

[63] Une table de correspondance entre la classe, l'année d'apprentissage et l'âge moyen des élèves au moment du recueil est disponible dans le glossaire.
[64] Le terme *standardisé* renvoie ici à la notion de format standard, commun à d'autres ressources.

devrait permettre à terme une interrogation fine du corpus par les chercheurs et les enseignants.

Ce projet représente donc aussi un véritable enjeu pour le TAL, puisque, comme nous avons pu le voir au chapitre 3, il s'agit d'un type de corpus encore peu étudié dans ce domaine. De plus, la grande variété présente dans les productions scolaires constitue un défi à l'automatisation de leur traitement.

Loin de cantonner le processus de constitution de corpus à la tâche de recueil, à travers le projet *Scoledit* nous proposons donc de considérer l'élaboration d'un corpus comme un processus plus large qui comprend à la fois le recueil et la numérisation de ce corpus, la construction et la réutilisation d'outils d'exploitation et la diffusion de ces données et outils.

L'enjeu du projet global est donc triple : 1/ un enjeu linguistique de constitution d'une ressource outillée pour la recherche en linguistique ; 2/ un enjeu pour le TAL, de caractérisation et de modélisation de types d'écrits souvent très éloignés de la norme ; 3/ un enjeu pédagogique et didactique appuyé par la connaissance fine des acquis et difficultés, accessibles au travers d'un outil d'interrogation du corpus.

2. Recueil du corpus

Comme nous venons de le préciser, la tâche fondatrice du projet *Scoledit* est la conception d'un corpus longitudinal de productions scolaires (productions de texte + dictées) recueillies en école primaire. Ce corpus qui rassemble des productions recueillies à chaque fin d'année scolaire du CP au CM2 auprès des mêmes élèves, a nécessité un long travail de recueil de cinq ans. Il s'est effectué en deux phases distinctes : un premier recueil de deux ans (CP-CE1) dans le cadre du projet « Lire - Écrire au CP » de l'Institut Français de l'Éducation (IFé) ; un second recueil de trois ans (CE2-CM2) dans le cadre du projet *Scoledit*.

Les productions sont recueillies par les chercheurs membres du projet. Le recueil est effectué au sein de l'école sans que les enseignants aient connaissance des sujets avant les passations. Ni la dictée, ni la production de texte ne sont donc préparées en amont dans les classes.

2.1. Phases de recueil

2.1.1. Première phase de recueil : le recueil du projet « Lire- Écrire au CP »

À l'origine, le projet *Scoledit* prend appui sur le projet de recherche national « Lire - Écrire au CP », coordonné par Roland Goigoux et financé par la direction générale de l'enseignement scolaire (DGESCO), l'Institut français de l'Éducation (IFé) et le laboratoire Acté (Clermont-

Ferrand)[65]. Cette recherche étudie l'influence des pratiques d'enseignement de la lecture et de l'écriture sur la qualité des premiers apprentissages en s'appuyant sur une vaste enquête effectuée entre septembre 2013 et juin 2015 dans 131 classes réparties dans 13 académies sur l'ensemble de la France. Au cours de cette enquête, de nombreuses données ont été recueillies, dont trois dictées par élève, en septembre 2013, juin 2014 (classe de CP) et en juin 2015 (classe de CE1), et deux productions de texte, en juin 2014 et juin 2015. Au total, 2 507 dictées et 2 507 productions de texte ont pu être recueillies en fin de CP et 2049 dictées et 2 049 productions de texte en fin de CE1.

2.1.2. Deuxième phase de recueil : le recueil du projet *Scoledit*

L'idée du projet *Scoledit* est née suite au projet « Lire - Écrire au CP ». Face à l'importance et à l'intérêt des données, particulièrement des productions écrites des élèves, il a semblé judicieux de poursuivre ce recueil auprès des élèves concernés par cette première recherche tout au long de leur scolarité à l'école primaire. Cependant, si le projet « Lire - Écrire au CP » a rassemblé de nombreux acteurs issus de différents laboratoires, le projet S*coledit* est un projet interne au *Lidilem* qui concerne un nombre bien plus restreint de chercheurs (3 enseignants chercheurs et une doctorante principalement). C'est pourquoi, il a été nécessaire de restreindre l'étendue du recueil.

Certains membres du projet entretenant des liens plus particuliers avec différents collègues du sud de la France (organisation de colloques, travaux de recherche communs, etc.), nous nous sommes concentrés sur les écoles réparties dans cinq académies de la moitié sud de la France : Bordeaux, Clermont-Ferrand, Grenoble, Lyon et Toulouse. Nous avons contacté les écoles ayant participé au projet « Lire - Écrire au CP » issues de ces cinq académies, en écartant certaines d'entre elles, peu accessibles. Parmi les écoles contactées, 37 ont répondu favorablement à notre demande. Dès la fin de l'année 2015 (recueil de CE1), nous avons ajouté trois écoles n'ayant pas participé au projet « Lire - Écrire au CP » mais avec lesquelles nous avions un contact particulier, ce qui, en cas de refus des autres écoles, nous assurait une poursuite du projet plus certaine sur le long terme. La répartition des écoles et leur caractérisation est visible sur la carte et dans le tableau ci-après (Figure 13 et Tableau 5).

[65] Goigoux, R. (2016). *Lire et écrire au CP. Étude de l'influence des pratiques d'enseignement de la lecture et de l'écriture sur la qualité des apprentissages*. Synthèse du rapport de recherche. Paris: MEN-ESR, Lyon: ENS-Lyon. http://ife.ens-lyon.fr/ife/recherche/lire-ecrire/rapport/rapport-lire-et-ecrire [consulté le 03/10/2019].

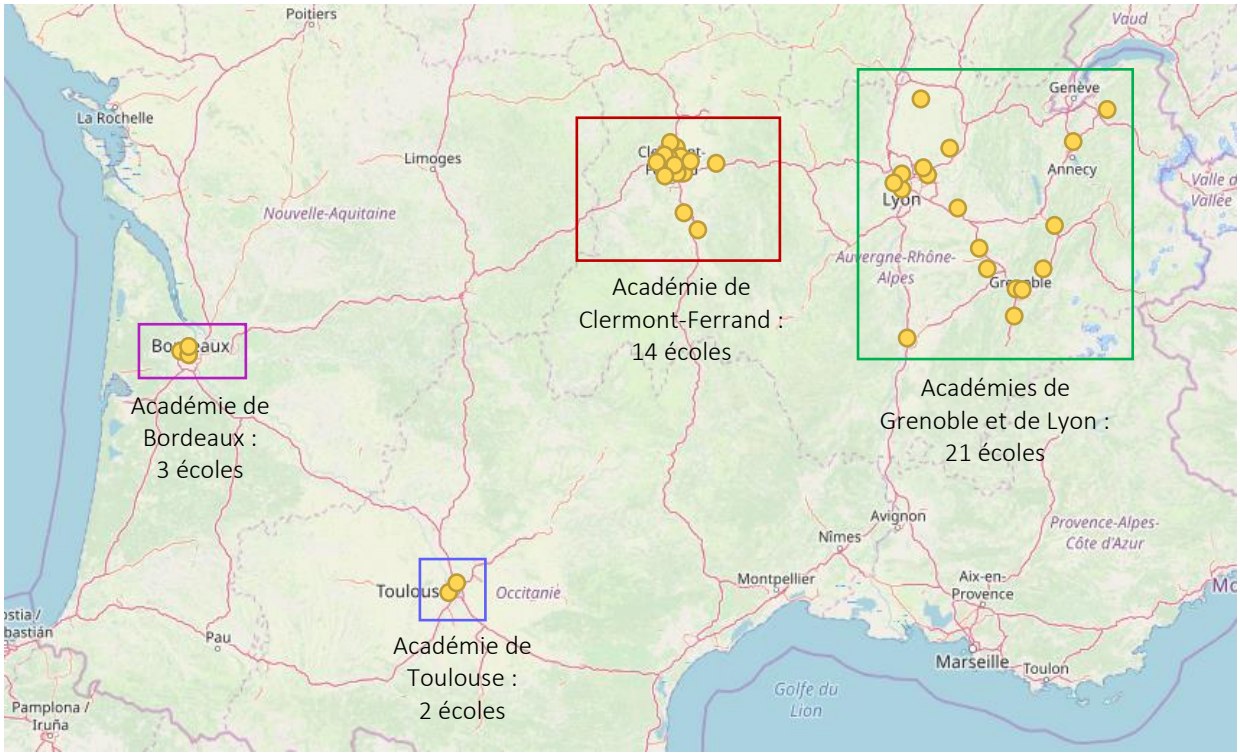


Figure 13 : Répartition des écoles du projet Scoledit

Académie	*Numéro de l'école*	*Nombre d'élèves*[66]	*Environnement de l'école*
Bordeaux	1	38	Milieu urbain grande ville
	2	51	Milieu urbain grande ville
	3	20	Milieu urbain grande ville (anciennement éducation prioritaire)
Clermont-Ferrand	4	37	Milieu urbain ville moyenne (éducation prioritaire)
	5	35	Milieu péri-urbain
	6	18	Milieu péri-urbain
	7	17	Milieu péri-urbain
	8	10	Milieu urbain
	9	27	Milieu urbain
	10	17	Milieu urbain
	11	74	Milieu urbain
	12	18	Milieu péri-urbain
	13	32	Milieu péri-urbain
	14	35	Milieu rural
	15	32	Milieu rural
	16	28	Milieu urbain ville moyenne

[66] Nombre d'élèves présents sur les listes en fin de CE2.

	17	31	Milieu rural
Grenoble	**18**	29	Milieu rural
	19	39	Milieu urbain ville moyenne
	20	20	Milieu rural
	21	35	Milieu urbain
	22	19	Milieu rural
	23	27	Milieu urbain
	24	24	Milieu rural
	25	14	Milieu rural
	26	119	Milieu urbain
	27	69	Milieu rural
	28	34	Milieu rural
	29	49	Milieu urbain
	30	39	Milieu urbain
	31	22	Milieu urbain
Lyon	**32**	47	Milieu rural
	33	52	Milieu rural
	34	55	Milieu rural
	35	24	Milieu urbain
	36	32	Milieu urbain
	37	72	Milieu urbain
	38	15	Milieu urbain

Tableau 5 : Écoles participant au recueil Scoledit

Ce recueil complémentaire, effectué d'avril 2016 à juin 2018, a permis de recueillir 1 135 productions de texte en classe de CE2[67] (2016), 1 132 productions de texte et 813 dictées en classe de CM1 (2017) et 1 030 productions de texte et 784 dictées en classe de CM2 (2018). En raison de contraintes liées au processus de recueil (nombre de classes concernées dans les écoles, temps accordé par les enseignants, etc.), nous n'avons pas pu recueillir les dictées de l'ensemble des élèves en classe de CM1 et de CM2. La dictée étant une tâche beaucoup plus contrainte que la tâche de production de texte, les variations entre les productions dictées sont bien plus faibles qu'entre les productions de textes des élèves, il nous a donc semblé qu'un nombre plus restreint de dictées était suffisant. Nous avons donc privilégié le recueil des productions de texte.

[67] En raison des démarches administratives conséquentes qu'a nécessitées la mise en place de la deuxième phase du recueil S*coledit*, aucune dictée n'a pu être recueillie cette année-là.

Un nombre conséquent de productions a été recueilli entre le CP et le CE1 au sein du projet « Lire - Écrire au CP ». Seules celles qui sont issues des écoles impliquées dans la deuxième phase du recueil (projet *Scoledit*) ont été intégrées au corpus. À l'issue de ces deux phases de recueil et après sélection parmi les productions recueillies lors de la première phase, nous disposons donc des productions suivantes :

- 1 151 dictées à l'entrée du CP ;
- 975 dictées à la fin du CP ;
  - 763 productions de texte à la fin du CP ;
- 736 dictées à la fin du CE1 ;
  - 871 productions de texte à la fin du CE1 ;
  - 1 135 productions de texte à la fin du CE2 ;
- 813 dictées à la fin du CM1 ;
  - 1 132 productions de texte à la fin du CM1 ;
- 784 dictées à la fin du CM2 ;
  - 1 030 productions de texte à la fin du CM2.

2.2. Consignes de recueil

Afin d'obtenir un corpus qui permettent de comparer les productions des élèves entre elles, le choix a été fait d'uniformiser la consigne à l'ensemble des élèves participant au recueil. Dans la mesure du possible, nous avons également essayé de l'uniformiser dans le temps, afin de permettre une comparaison entre les années. Mais nous verrons qu'à cet égard, il nous a fallu prendre en compte le degré de maitrise de la langue des élèves ainsi que les choix effectués lors du projet « Lire-écrire CP » dont sont issues une partie de nos données. Il n'a donc pas été possible d'utiliser la même consigne de production de textes en CP et en CE1. Les consignes de dictées ont également varié plusieurs fois au cours du recueil.

L'ensemble des consignes de recueil est disponible en Annexe 3, tandis que l'Annexe 4 donne un exemple de l'ensemble des productions recueillies pour un élève présent tout au long du recueil (élève 96).

2.2.1. Productions de texte

Au cours du projet « Lire-écrire CP », deux consignes différentes ont été présentées aux élèves selon l'année d'apprentissage : une consigne de production de texte à partir d'images séquentielles, en classe de CP et une consigne de production plus libre, ne contraignant que le

choix des personnages, en classe de CE1. Pour la suite du recueil, nous avons fait le choix de conserver la consigne donnée en CE1. Ce choix devrait faciliter les comparaisons entre les productions produites au fil des années.

2.2.1.1. Consigne de production de texte en classe de CP

Lors de la phase de collecte de l'année de CP, quatre images (Figure 14) étaient présentées aux élèves, qui disposaient ensuite de 15 minutes pour répondre à la consigne suivante : « *Aujourd'hui vous allez écrire chacun l'histoire d'un petit chat. Je vais vous montrer ce qui arrive à ce petit chat. Regardez bien les images. Vous allez écrire cette histoire ici. Si vous avez oublié l'histoire, vous pouvez retourner la feuille pour retrouver les dessins. Vous avez 15 minutes pour ce travail. Vous allez travailler seul ; personne ne vous aidera, par exemple à écrire un mot*. ». Lors de la rédaction, les élèves pouvaient, au besoin, consulter les images au tableau ou au dos de leur feuille [68]. Les élèves ont rédigé sur un support à très grands carreaux conçu spécifiquement pour le projet. Pour plus de détails, il est possible de se référer aux livrets en Annexe 3.

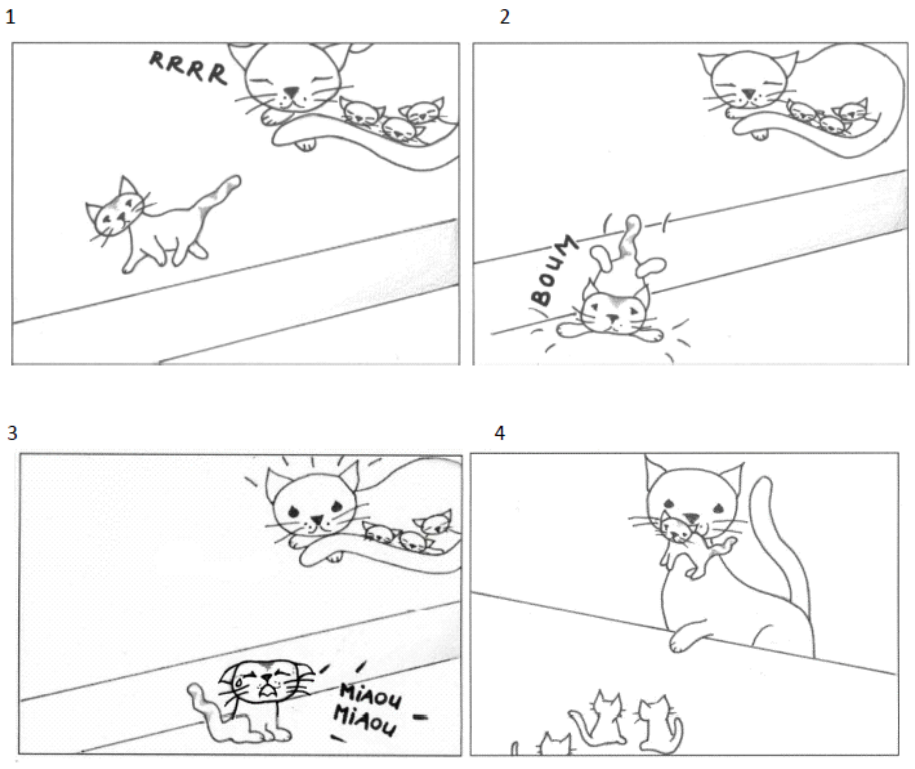
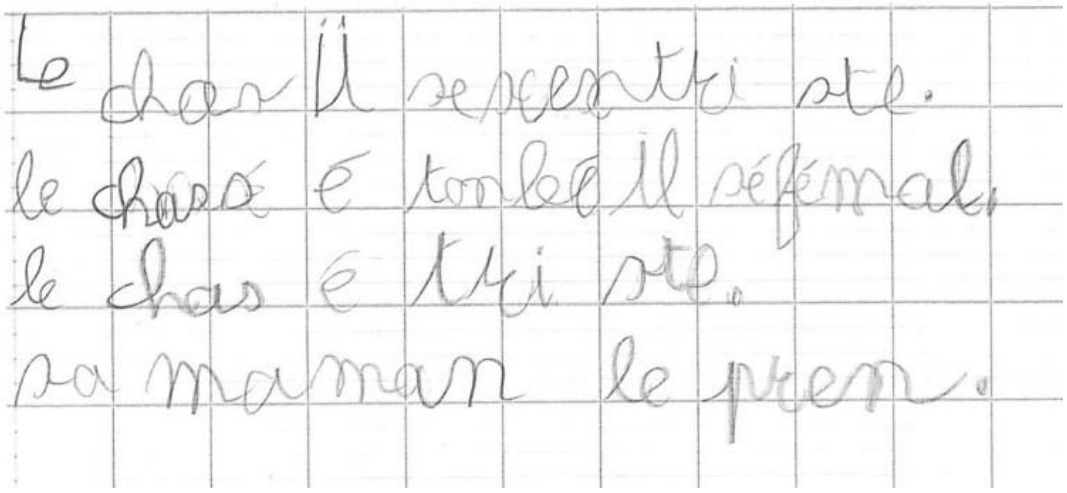


Figure 14 : Images présentées aux élèves lors de la production écrite en CP

[68] Pour plus de précisions méthodologiques, se référer à l'article de Y. Soulé, B. Kervyn, T.Geoffre, & J. C. Chabanne (2016).

La figure 15 illustre une production de texte recueillie en classe de CP selon cette consigne.



Transcription :
Le chas ii sescentri ste. / le chass<revision/> é tonbé il séfémal. / le chas é tri ste. // sa maman le pren.

Figure 15 : Production de texte en CP - élève 1154

2.2.1.2. Consigne de production de texte pour les classes de CE1 à CM2

À partir de l'année de CE1, quatre images figurant quatre personnages (Figure 16) étaient présentées aux élèves. Il était alors demandé aux élèves d'inventer une histoire (cf. Annexe 3) à partir de ces images. Les élèves entendaient la consigne suivante : « A*ujourd'hui, vous allez écrire chacun une histoire avec un ou deux personnages.* », puis, après que chaque personnage a été nommé par l'ensemble de la classe, le chercheur ou la chercheuse disait : *« Je vais vous lire la consigne qui est écrite sous les images, écoutez bien la consigne. Voici 4 personnages. Choisis un ou deux personnages et raconte une histoire. Entoure-le ou les personnages que tu as choisis.* » On leur précisait enfin : « *Avant de commencer à écrire, mettez bien dans votre tête l'histoire de ce personnage ou de ces deux personnages. Vous avez 20 minutes pour écrire cette histoire. Vous allez travailler seul ; personne ne vous aidera, par exemple à écrire un mot.* »

De la classe de CE2 à la classe de CM2, la consigne entendue par les élèves était sensiblement la même. Cependant, nous avons fait le choix d'augmenter le temps d'écriture en CM1 et en CM2. Les élèves ont alors disposé de 30 minutes pour écrire leur histoire. Dans la mesure du possible, les élèves ont été encouragés à ne produire que 25 minutes afin de disposer de 5 minutes de relecture. À partir du CE2, les grands carreaux ont également laissé la place aux petits carreaux, plus adaptés à ces classes.

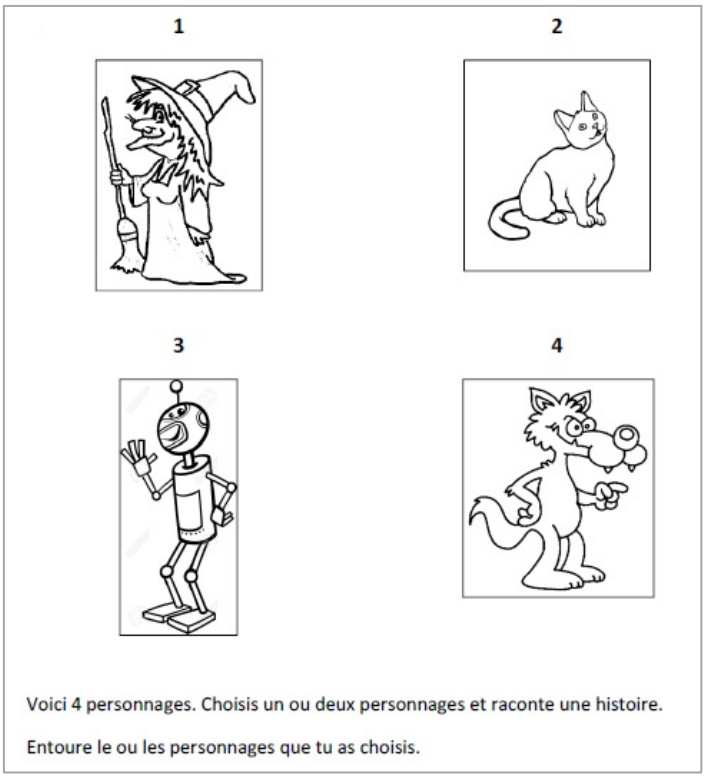
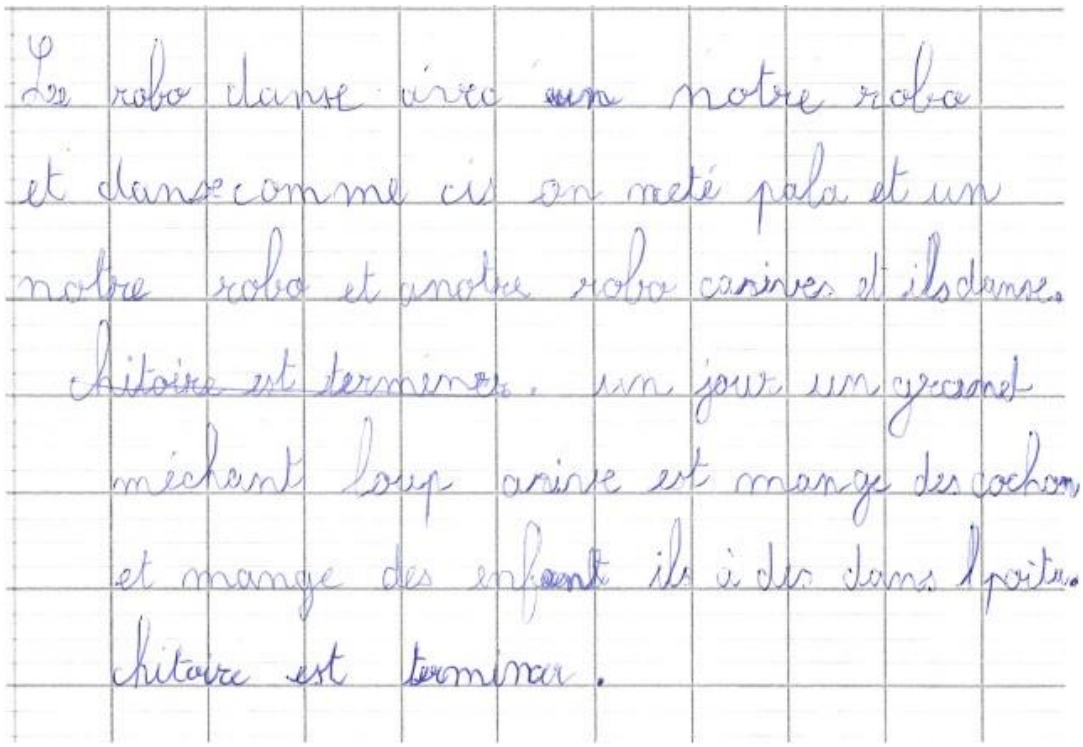


Figure 16 : Images présentées aux élèves lors de la production écrite en CE1, CE2, CM1 et CM2

La figure 15 et la figure 17 illustrent les productions de texte recueillies à partir de la classe de CE1 à l'aide de la consigne ci-dessus.


Le robo danse avec un notre robo
et danse comme ci on meté pala et un
notre robo et anotre robo casives et ils danse.
chitoire est terminer. un jour un grand
méchant loup arive est mange des cochon
et mange des enfant ils à des dans l'poitr.
chitoire est terminer.

Transcription :
Le robo danse avec <revision/>un notre robo / et danse comme cis on neté pala et un / notre robo et anotre robo carives et ils danse. / <revision/> un jour un grand / méchant loup arive est mange des cochon / et mange des enfant ils à des dans <revision/> poitu. / histoire est terminai.

Figure 17 : Production de texte en CE1 - élève 1 154

2.2.2. Dictées

En parallèle du recueil de productions de texte, une tâche de dictée a été demandée à une majorité des élèves concernés. Pour cette tâche, différents supports ont été utilisés, répétés ou non selon les années d'apprentissage.

2.2.2.1. Dictée du projet « Lire – Écrire au CP » (*DictéeIFÉ(1)* et *DictéeIFÉ(2)*)

Dans le cadre du projet « Lire – Écrire au CP », une dictée (*DictéeIFÉ(1)*) a été proposée à l'entrée du CP, au mois de septembre, contenant :

- les trois mots suivants : *lapin*, *rat* et *éléphant ;*
- la phrase : *Tom joue avec le rat*.

En fin d'année, la même dictée (*DictéeIFÉ(2)*) a été demandée aux élèves, mais, au regard de leurs apprentissages, la phrase suivante a été ajoutée :

- *Les lapins courent vite*.

Pour chacune de ces dictées, les enquêteurs étaient invités à lire les phrases lentement et non à les dicter mot à mot afin de ne pas induire la segmentation aux apprenants. Ils étaient également invités à faire preuve de bienveillance envers les productions des enfants et à les encourager à faire des essais même s'ils ne connaissaient pas l'orthographe d'un mot. Pour plus de détails, se référer aux consignes dans l'Annexe 3.

2.2.2.2. Dictées de mots et de phrases © DEPP (*DictéeDEPP(1)* et *DictéeDEPP(2)*)

Les items dictés en classe de CE1 sont issus des évaluations nationales de la Direction de l'Evaluation de la Prospective et de la Performance (DEPP). Ces épreuves ont été mises en place en 1989 à destination des élèves de CE1.

Les items retenus sont les suivants :

- Les six mots (*DictéeDEPP(1)*) : *patin*, *pâtisson*, *capuchon*, *récréation*, *charitable* et *magnifique* ;
- Les deux phrases (*DictéeDEPP(2)*) : *En été, les salades vertes poussent dans les jardins »* et *« Les jeunes canetons picorent le blé avec la poule noire*.

Après la dictée, les élèves sont invités à se relire ainsi qu'à vérifier le bon usage des majuscules et des points, tout comme la présence des accords.

La même dictée de mots et de phrases a été reprise en classe de CM1. En CM2, seules les phrases ont été reprises, notamment en raison de contraintes temporelles.

2.2.2.3. Texte *Le corbeau* (*DictéeCorbeau*)

À partir de la classe de CM1, le début d'un texte intitulé *Le corbeau* est ajouté à l'épreuve de dictée dans certaines classes[69]. Le texte dicté (*DictéeCorbeau*) est le suivant :

Le corbeau / perché sur l'antenne d'un bâtiment / tient dans son bec / une souris blessée. / Rendus furieux / par cet oiseau cruel, / des enfants lancent des cailloux / pour l'obliger à s'envoler.

Le symbole / marque les groupements de mots lus et relus lors de la dictée aux élèves.

Le tableau 6 résume les items choisis pour les épreuves de dictée.

Moment du recueil	***DictéeIFÉ(1)***	***DictéeIFÉ(2)***	***DictéeDEPP(1)***	***DictéeDEPP(2)***	***DictéeCorbeau***
Classe de CP – sept.	X	-	-	-	-
Classe de CP – juin	X	X	-	-	-
Classe de CE1	-	-	X	X	-
Classe de CE2	-	-	-	-	-
Classe de CM1	-	-	X	X	X
Classe de CM2	-	-	-	X	X

Tableau 6 : Supports de l'épreuve de dictée

2.3. Difficultés d'un recueil longitudinal en milieu scolaire

Le travail de recueil d'un corpus scolaire, qui plus est longitudinal, n'est pas exempt de difficultés et de facteurs de variation qui influent de fait sur la taille, la structure et le contenu du corpus. Ces difficultés sont de différentes natures : complexité des démarches administratives ; complexité de l'organisation du recueil ; difficultés du suivi longitudinal des élèves ; diversité des contextes de recueil.

[69] Le choix s'est porté sur ce texte car il est utilisé pour une autre étude en cours avec laquelle les bases d'un travail commun ont été posées (projet *Corpuscol*).

2.3.1. Complexité des démarches administratives

Pour les premières années du corpus, nous avons pu récupérer les productions issues d'un autre projet et donc bénéficier des démarches administratives réalisées pour ce projet (conventions avec les écoles, autorisations des parents d'élèves, etc.). Cependant, ce second projet n'ayant pas été prévu au moment de l'établissement des premières conventions, il nous a fallu recommencer ces démarches et établir de nouvelles conventions.

Pour commencer, il nous faut prévenir que ce processus administratif est un processus long et variable. En effet, les interlocuteurs à contacter sont différents selon les académies. Dans certaines académies, un contact avec l'IENA (Inspecteur de l'Éducation nationale, adjoint à l'inspecteur d'académie) est suffisant avant de contacter les écoles et les enseignants concernés. En revanche, dans d'autres académies, il nous a d'abord fallu contacter le CARDIE (Conseiller académique Recherche et Développement en Innovation et en Expérimentation), qui a lui-même contacté les DASEN (Directeur académique des services de l'Éducation nationale). Une fois leur accord obtenu, nous avons alors pu contacter les différents inspecteurs des circonscriptions concernées, avant de pouvoir contacter les directeurs concernés. Bien évidemment, ce processus est fortement chronophage et, bien que nous ayons commencé ces démarches en novembre, nous n'avons pu nous déplacer dans certaines écoles qu'à la fin du mois de juin. Nous avons donc dû nous résoudre pour l'année de CE2 à une seule épreuve, la production de texte, sans la dictée.

De plus, certaines écoles ont préféré ne pas répondre favorablement à notre demande. En effet, sur 50 écoles sollicitées, 40 ont accepté de poursuivre le projet. Beaucoup d'écoles n'ont pas répondu à notre demande ou seulement après plusieurs relances de notre part. Le nombre d'écoles ayant accepté notre demande est variable selon les académies et il semble que plusieurs facteurs rentrent en ligne de compte. Ainsi, la sollicitation importante pour de nombreux projets de recherche pour certaines écoles, ressentie notamment dans certaines académies et dans les écoles situées en ville semble freiner la participation des écoles. En revanche, l'habitude de certaines écoles ou académies à participer à des projets de recherche sur l'écriture ou la lecture a, semble-t-il, facilité le contact.

Les conventions établies au cours de ces démarches sont établies pour trois ans, le temps nécessaire au recueil des années de CE2 à CM2. Néanmoins il est nécessaire de refaire un travail de communication et d'argumentaire tous les ans, d'une part en raison du changement de direction dans certaines écoles ; d'autre part, parce que nous intervenons dans les classes d'enseignants différents chaque année. Chacun de ces acteurs peut s'avérer plus ou moins

favorable à notre recherche, voire hostile, ce qui peut expliquer qu'une ou l'autre école n'ait pu être suivie à un moment donné.

Après obtention des autorisations de l'éducation nationale et des écoles, nous avons demandé les autorisations des élèves et de leurs parents pour étudier et diffuser les productions recueillies.

Parallèlement à ces étapes, notre projet (deuxième phase de recueil) a été déposé auprès de la CNIL[70] afin d'obtenir la validation du processus complet, à savoir le recueil des données, leur traitement et leur conservation.

2.3.2. Complexité de l'organisation du recueil

Une autre des difficultés rencontrées est à la fois le cout financier, et surtout le temps nécessaire, tant pour l'organisation du recueil que pour le recueil lui-même. En effet, un recueil dans 40 écoles signifie près de 60 passations en classe qui nécessitent chacune entre 45 min à 1h30. Chaque école nécessite environ une demi-journée de disponibilité, si on considère que chaque école ouvre entre 8 et 9 demi-journées par semaine, notre recueil nécessite donc 5 semaines de déplacement au minimum. Un tel recueil est donc très chronophage.

L'objectif du projet est de recueillir les productions des élèves en fin d'année scolaire. Ce lourd recueil doit donc être réalisé dans un laps de temps assez court, de fin avril à fin juin et combiner les emplois du temps bien remplis des écoles avec ceux des chercheurs n'est pas toujours évident. Dans de rares cas, il nous a été nécessaire d'intervenir début juillet, ce qui a entrainé un fort taux d'absentéisme dans les classes.

En ce qui concerne le cout financier d'un tel recueil, il a été estimé, pour la phase de recueil *Scoledit*, entre 1 500€ et 2 000€ par an. Ce montant inclut à la fois les déplacements et les photocopies des feuilles de recueil. Il exclut le cout salarié des enquêteurs, car dans notre cas les enquêtes ont été réalisées par les quatre chercheurs de notre groupe. Heureusement, nous avons pu bénéficier d'un appui financier à trois reprises de la part du *WP3 Démarre SHS !* du *Data Institut* de l'université Grenoble Alpes.

Actuellement le discours institutionnel incite les chercheurs et chercheuses concernées à travailler en relation étroite avec le terrain, au sein des écoles notamment. On retrouve notamment cette volonté dans les appels à projet PIA3, déclinaisons territoriales du

[70] La commission nationale de l'informatique et des libertés (CNIL) est chargée de veiller à ce que l'usage de l'informatique ne porte pas atteinte à la vie privée ou aux libertés individuelles. Dans le cadre de la recherche, elle s'assure notamment que l'anonymat des participants soit respecté et que les autorisations nécessaires soient demandées.

Programme d'investissements d'avenir (PIA) qui vise à encourager les liens entre recherche et terrain, notamment au niveau des entreprises. Cette question est donc une question d'actualité pour la recherche. Cependant, comme nous l'avons exposé ici, ce travail de terrain a un cout et nécessite beaucoup de temps pour être réalisé dans de bonnes conditions.

2.3.3. Difficultés du suivi longitudinal

Il semble important également de mentionner la difficulté de suivre les élèves sur une durée aussi longue. En effet, certains changent d'école, d'autres redoublent ou changent de classe pour aller dans le niveau supérieur. Il est également possible que certains élèves soient absents le jour du recueil pour de multiples raisons : maladie, rendez-vous extrascolaires ou médicaux, départ en vacances, etc.

2.3.4. Diversité des contextes de recueil

Enfin, un des objectifs du projet *Scoledit* est l'élaboration d'un corpus recueilli dans des conditions uniformes, à l'aide d'une même consigne, mais il est nécessaire de prendre en compte la difficulté à respecter cette contrainte.

Ceci s'explique par la pluralité des conditions de passation. Nous avons ainsi eu le sentiment que la présence ou non de l'enseignant, d'un ou d'une AVS ou d'un autre adulte modifiait l'attention, la concentration des élèves, tout comme la salle de recueil, selon que c'était la salle de classe habituelle, la salle de classe des camarades, la salle informatique, la salle d'art ou la salle de science. Il arrivait également qu'un adulte présent donnait des consignes supplémentaires. De même, les affichages pouvant guider les élèves pouvaient varier selon les classes et les salles.

Outre les affichages, de nombreux facteurs peuvent faire varier le contenu des productions, comme les œuvres ou albums étudiés, les contenus linguistiques étudiés récemment, etc. Une attention particulière aux conditions de passation a été portée dans le cadre du projet « Lire – écrire au CP », notamment en termes d'affichage, notamment en raison de l'évolution des affichages dans les classes supérieures. Nous n'avons pas pris en compte ces aspects-là dans la suite du projet. Notons également que nous étions plusieurs à effectuer ce recueil et que nous pouvions donc malgré nous introduire de la variation dans les passations.

3. Caractérisation et structure du corpus de textes

3.1. Structure du corpus

L'ensemble des recueils a permis de rassembler près de 9 400 productions : environ 5 000 dictées et 4 300 productions de textes (Figure 18). Toutefois, pour la suite de notre travail, nous

ne conservons que les productions de texte des élèves présents du CP au CM1 (hachures, Figure 18). Ces productions constituent un sous-corpus appelé *corpus longitudinal*.

Ce choix s'explique pour plusieurs raisons :

- Le travail effectué pour cette thèse se concentre uniquement sur les épreuves de production de textes.
- En raison des contraintes du recueil et de la disponibilité des productions au moment de l'élaboration des traitements automatiques, les productions de CM2 n'ont pu être utilisées.
- Dans la perspective d'une étude longitudinale, seules les productions des apprenants présents à l'ensemble des phases de recueil sont retenues, soit 373 élèves.

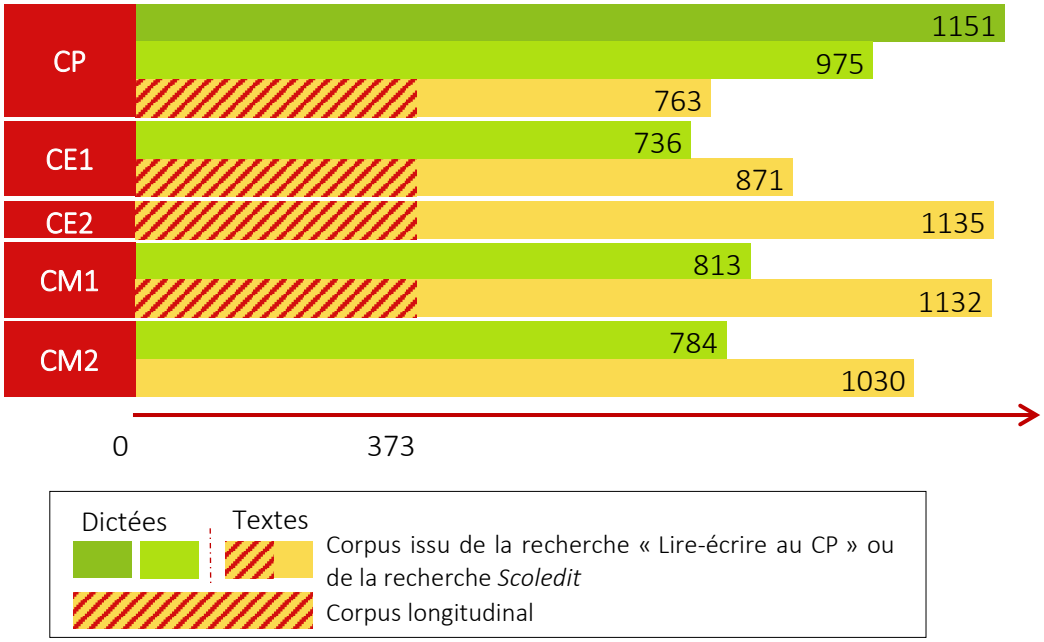


Figure 18 : Ensemble des données recueillies lors des deux phases de recueil

3.2. Caractérisation du corpus

Le tableau 7 recense différents indicateurs des productions de notre corpus, niveaux par niveaux. Ce tableau montre qu'au fil des années les productions s'allongent. Ce phénomène s'explique peut-être aussi en partie par l'allongement du temps de passation entre le CP, le CE1 et le CE2, mais cette explication ne suffit pas à justifier l'ensemble de l'accroissement constaté. Notons également que si le nombre de phrases augmente, le nombre moyen de mots par phrases reste stable. Cependant, cela ne signifie en rien que la structure des phrases reste inchangée.

	Nombre de productions	*Nombre de formes*[71]	*Nombre moyen de formes par production*	*Nombre minimal de formes*	*Nombre maximal de formes*
CP	373	**11 141**	**29,9**	3	93
CE1	373	**27 504**	**73,7**	7	234
CE2	373	**48 654**	**130,4**	6	363
CM1	373	**58 952**	**158,0**	19	466
CP-CM1	*1 492*	*146 251*	***98,0***	*3*	*466*

Tableau 7 : Caractéristiques du corpus longitudinal

4. Métadonnées

Le corpus *Scoledit* s'accompagne de différentes métadonnées qui portent sur le recueil, sur les écoles et les classes participant au projet et sur les élèves. Les métadonnées portant sur le recueil incluent principalement les consignes de recueil et les livrets sur lesquels les élèves ont produit les textes.

Pour chaque école, nous donnons l'académie de laquelle elle dépend et la composition sociale de ses élèves. Nous caractérisons également chaque école en termes de territoire (urbain / péri-urbain / rural). La plupart du temps, la composition de la classe est disponible, à savoir s'il s'agit d'un cours double ou d'un cours simple.

Enfin, nous incluons dans les métadonnées des caractéristiques propres aux élèves, notamment le mois et l'année de naissance, le sexe et si l'élève a redoublé le CP. La langue parlée à la maison et la catégorie socio-professionnelle des parents ont également été demandées.

Notons cependant que pour des raisons pratiques l'ensemble de ces données n'ont pu être demandées chaque année. Nous disposons d'un grand nombre d'entre elles, mais il est possible que certaines métadonnées soient manquantes dans la phase finale du corpus.

Bien que nous n'ayons pas prévu d'exploiter certaines de ces métadonnées dans le projet *Scoledit*, elles nous ont paru nécessaires pour que d'autres projets de recherche puissent émerger à partir du corpus *Scoledit*. Les graphiques ci-après présentent certaines de ces métadonnées (Figure 19, Figure 20, Figure 21 et Figure 22).

[71] Le nombre de formes est calculé à partir de la version normalisée des productions.

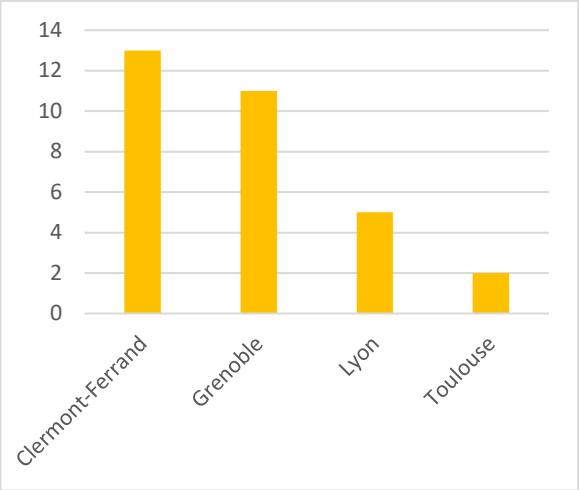


Figure 19 : Répartition des écoles au sein des académies

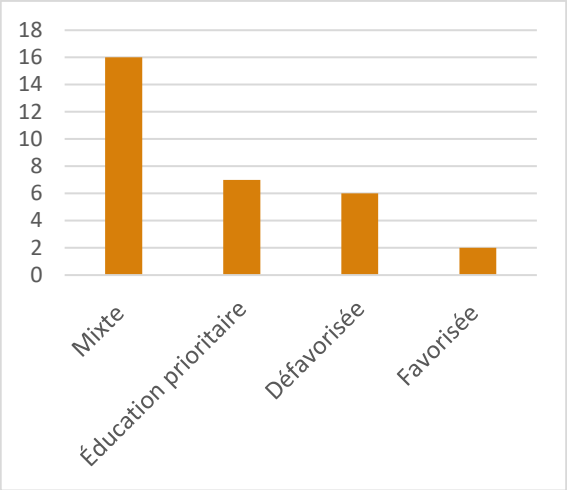


Figure 20 : Composition sociale des écoles

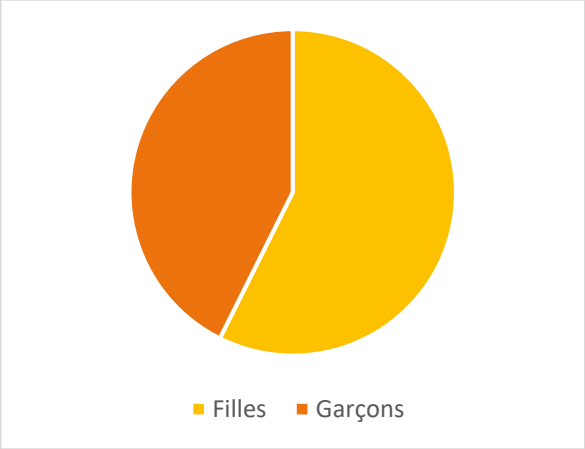


Figure 21 : Répartition genrée des élèves

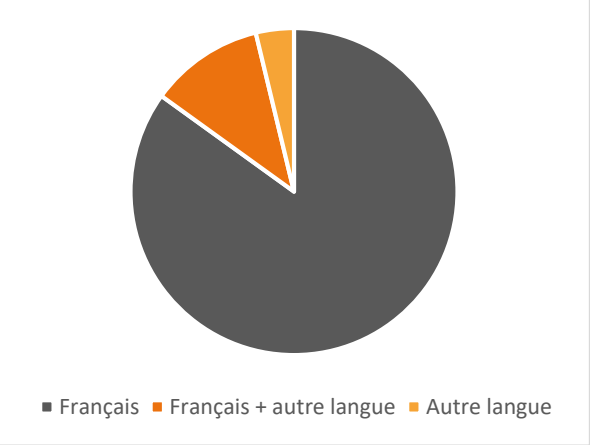


Figure 22 : Répartition des élèves par langue parlée à la maison

5. Numérisation du corpus

La numérisation du corpus se décline en deux étapes : le scan des productions et leur transcription. À l'issue de cette étape, nous obtenons donc deux nouvelles versions des productions, une reproduction photographique, consultable pour les lecteurs mais non exploitable informatiquement, et une reproduction dactylographique, manipulable à l'aide d'outils informatiques.

À terme, l'ensemble de ces versions sera disponible en ligne. Un travail soigneux d'anonymisation est donc nécessaire, de sorte que l'élève qui a produit le texte diffusé ne puisse être identifié. Pour ce faire, un identifiant a été attribué à chaque élève. Cet identifiant est l'unique repère utilisé par les membres du projet pour manipuler les productions.

Nous avons constaté que les élèves utilisaient peu leur propre prénom dans leur production et davantage les prénoms de leurs camarades de classe ou de leur famille. Nous avons donc jugé que les prénoms présents dans les productions ne permettaient pas l'identification des élèves et ils ont été conservés la plupart du temps. De manière générale, peu d'éléments permettaient l'identification des élèves scripteurs, ils ont donc été traités au cas par cas lors de la numérisation.

Comme nous l'avons vu précédemment, l'ensemble des caractéristiques visuelles d'une production manuscrite peuvent être reproduites numériquement et des choix doivent être réalisés par les transcripteurs. L'élaboration de conventions de transcription permet de rendre ce processus homogène sur l'ensemble du corpus, indépendamment, dans la mesure du possible, du transcripteur ou de la transcriptrice. Cette homogénéité est nécessaire à l'élaboration de traitements automatiques du corpus.

5.1. Choix de transcription pour le projet *Scoledit*

Dans notre cas, la transcription effectuée vise principalement à rendre visible et exploitable informatiquement le contenu textuel de la production finale de l'élève, indépendamment des aspects génétiques (révision du texte par ajout ou suppression par exemple) et visuels (sauts de ligne, changement de stylo, etc.) des productions. Néanmoins, dans un souci de cohérence avec les autres équipes de recherche qui travaillent à l'élaboration de corpus scolaires et afin de faciliter la transcription de notre corpus s'il venait à être employé dans d'autres perspectives de recherche ou par d'autres chercheurs, nous avons tout de même signalé certains éléments génétiques ou informations visuelles, sans les détailler. Enfin, la lecture des productions d'enfants n'étant pas une tâche triviale, il est parfois nécessaire d'indiquer les points de doute de la transcription.

Comme nous l'avons mentionné précédemment, un des objectifs du projet *Scoledit* est l'exploitation informatique et automatique des manuscrits d'élève. Les choix de transcription ont donc dû être faits en cohérence avec cet objectif. La transcription adoptée au sein du projet est donc une transcription linéaire, exploitable par des outils informatiques. Cependant, nous conservons certaines données concernant la mise en page des productions.

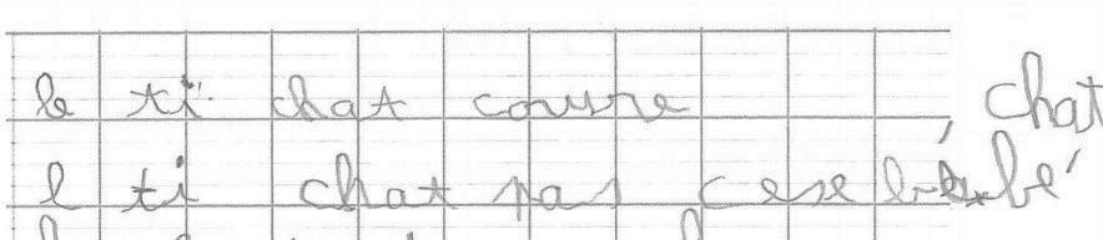
De plus, en accord avec l'objectif de description linguistique du projet, la transcription effectuée vise principalement à rendre visible le contenu textuel de la production finale de l'élève. Les aspects génétiques (révision du texte par ajout ou suppression par exemple) et visuels (sauts de ligne, changement de stylo, etc.) des productions y sont donc peu détaillés.

À partir de ces perspectives d'exploitation, nous avons mis au point un protocole d'annotation (disponible en annexe, cf. Annexe 5). Les choix effectués au cours de l'étape de transcription sont détaillés ici.

5.1.1. Eléments génétiques, visuels ou méta-textuels

Les aspects génétiques, visuels ou méta-textuels nous intéressent assez peu mais ils peuvent intéresser d'autres équipes de recherche, c'est pourquoi nous nous contenterons de relever leur présence.

Les traces de révisions, comme les réécritures, les ratures et les traces de gomme, ont été signalées par la balise **<revision/>**, tout comme les ajouts, de quelque nature que ce soit, ont été signalés par la balise **<ajout>***ElementAjouté***</ajout>**.



Transcription :
le ti chat coure // l ti chat pas c ese b<revision/>ébé <ajout>chat</ajout>

Figure 23 : Scan et transcription de la production de texte de CP de l'élève 1 363

Certains éléments non textuels ou non inclus dans les productions mais visibles sur les scans des productions ont également été reportés dans la transcription afin de permettre une plus grande cohérence visuelle entre le scan d'une production et sa transcription. La balise **<dessin/>** permet ainsi de marquer la présence d'un dessin dans le texte ou à la suite du texte.

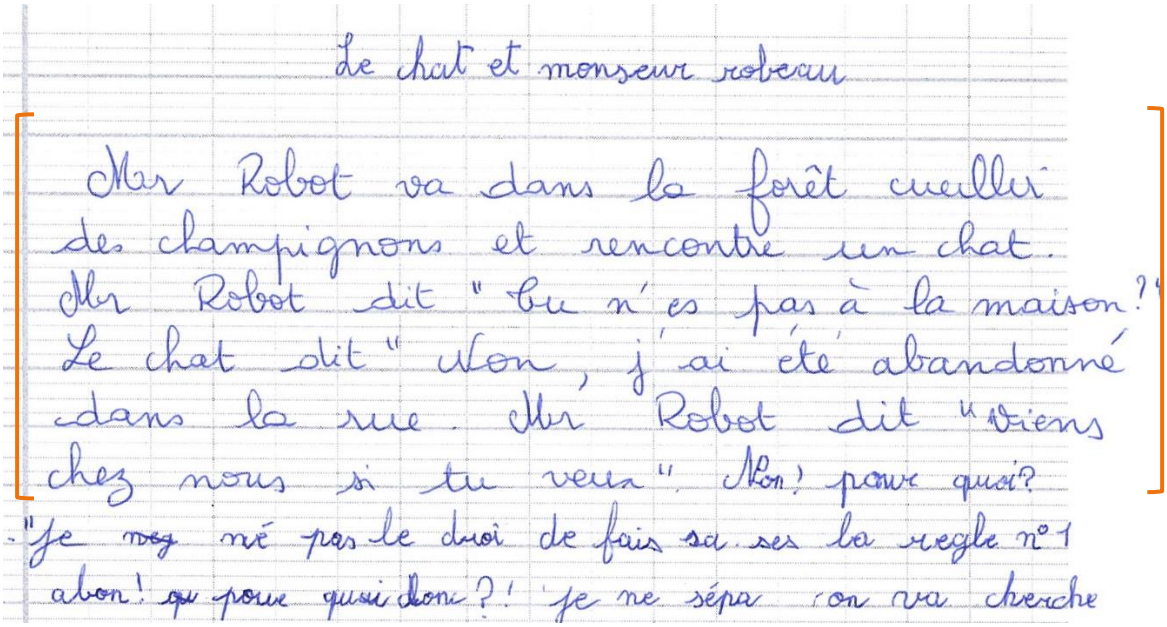
Transcription :
[…] mère la mai avec les otre est / lanbrase. <dessin/>

Figure 24 : Scan et transcription de la production de texte de CP de l'élève 2 413

La balise **<meta>***ElementConcerné***</meta>** a une fonction quelque peu similaire puisqu'elle permet de signaler un élément visible mais qui ne sera pas pris en compte dans le traitement

du contenu textuel. Elle permet également de rendre visibles des pans de production qui ne seront pas pris en compte dans notre analyse, mais qui peuvent l'être pour d'autres analyses.

La figure 25 illustre un exemple dans lequel une part importante de la production a été rédigée en dictée à l'adulte. Cette partie de transcription a été transcrite, mais par soucis d'homogénéité avec les autres productions nous ne l'inclurons pas dans nos analyses. De plus, cette partie pourrait fausser certaines de nos analyses, comme les analyses orthographiques.



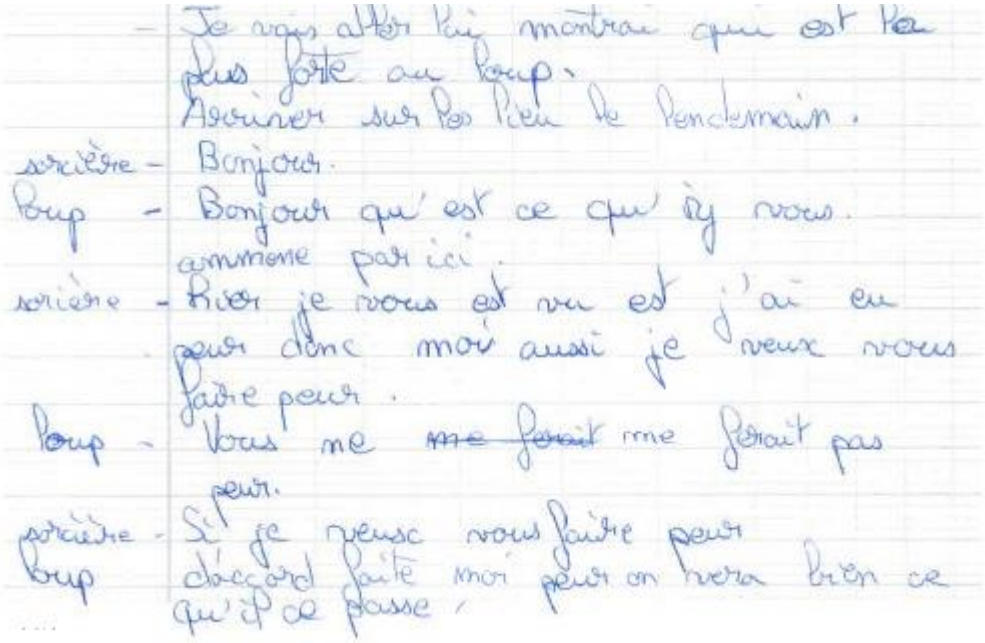
Transcription :

<titre>Le chat et monseur robeau</titre> <meta>Mr Robot va dans la forêt cueillir / des champignons et rencontre un chat. / Mr Robot dit « Tu n'es pas à la maison ? » / Le chat dit « Non, j'ai été abandonné / dans la rue. Mr Robot dit « Viens / chez nous si tu veux ».</meta> Non! pour quoi ? […]

Figure 25 : Scan et transcription de la production de texte de CE2 de l'élève 1 101

La figure 26 illustre un exemple où l'élève a précisé le locuteur en début de ligne à chaque nouveau tour de parole, à l'instar de ce qui se fait pour le théâtre. Bien que ce soit un procédé

intéressant, ces mentions risquent de perturber certaines analyses, comme les analyses syntaxiques, nous choisissons donc de ne pas les inclure dans l'analyse.

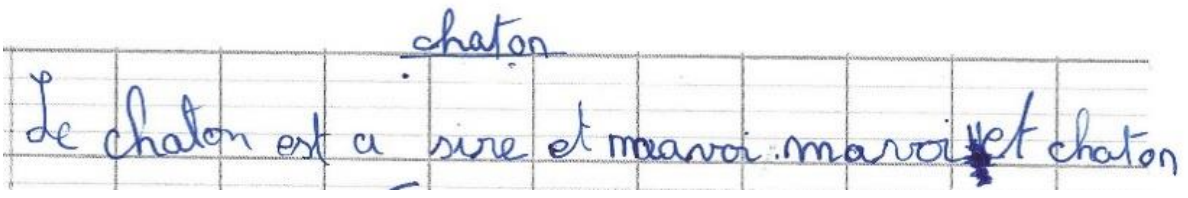


Transcription :

[…] - Je vais aller lui montrai qui est l<revision/>a / plus forte au loup. / Arriver sur les lieu le lendemain. / <meta>sorcière</meta> - Bonjour. / <meta>loup </meta> - Bonjour qu'est ce qu'<unsure>iy</unsure> vous / ammene par ici. / <meta>sorcière</meta> - hier je vous est vu est j'ai eu / peur donc moi aussi je veux vous / faire peur. / <meta>loup</meta> - Vous ne <revision/> me ferait pas / peur / <meta>sorcière</meta> - Si je veux vous faire peur / <meta>loup</meta> - d<ajout>'</ajout>accord faite moi peur on vera bien ce / qu'il ce passe. /# […]

Figure 26 : Scan et transcription de la production de texte de CM2 de l'élève 103

Toujours dans un souci d'approcher le travail réalisé par les autres équipes de recherche, nous avons fait le choix d'ajouter une balise **<titre>***TitreDeLaProduction***</titre>** qui, comme son nom l'indique, permet d'identifier le titre de la production de texte. Cette balise représente déjà un premier élément d'interprétation des productions.



Transcription :

<titre>chaton</titre> Le chaton est a sire et mavoi mavoi <revision/>et chaton

Figure 27 : Scan et transcription de la production de texte de CE1 de l'élève 1 283

Nous appelons productions vides, les productions pour lesquelles l'élève n'a rien produit malgré sa présence le jour de la passation. Ces cas se retrouvent essentiellement en CP et restent relativement marginaux. Pour distinguer ces productions des productions non transcrites et des cas d'absence des élèves, nous avons créé la balise **<empty/>** qui met en lumière ces productions vides.

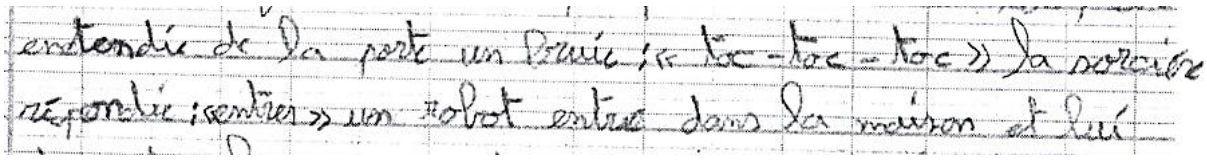
5.1.2. Eléments textuels

Puisque ce sont les éléments textuels qui nous intéressent en premier lieu, présentons maintenant les choix de transcription effectués à leur égard. Tout d'abord, reprécisons que la tâche de transcription permet de rendre manipulable informatiquement la production de l'élève. La transcription est donc une représentation numérique de cette production. À ce titre, la transcription reprend les choix textuels effectués par l'apprenant ou l'apprenante, tant en termes d'orthographe, que de syntaxe ou encore de lexique.

La transcription reproduit également certains choix graphiques effectués par l'élève, tels que le choix des majuscules accentuées ou non. De même, le choix de l'usage des guillemets *français* « et » ou *anglais* " et " est reproduit.

Transcription :
[...] mieux / elle atrappait un enfant et "hop" elle le mettez / [...]

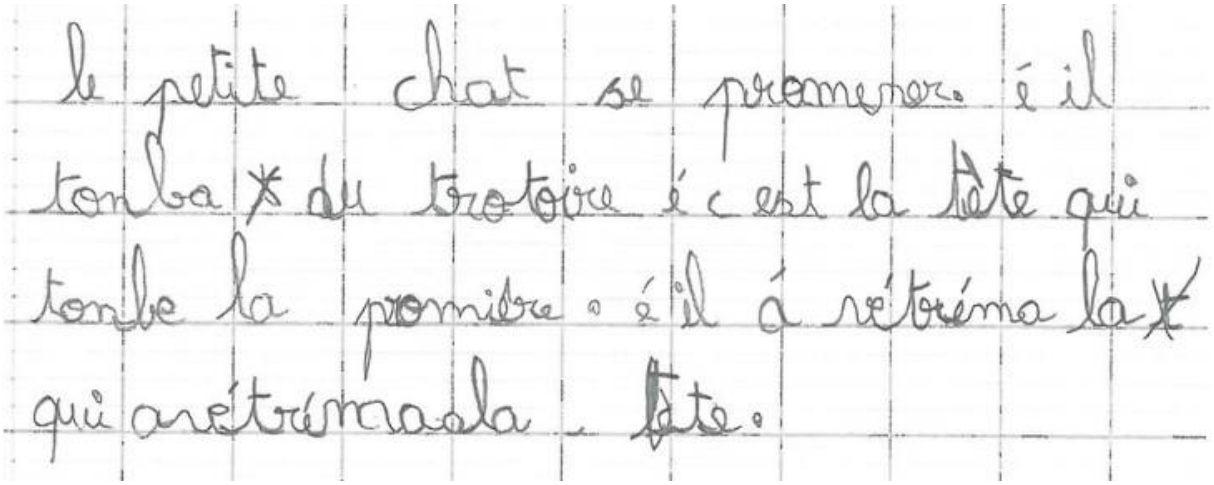
Figure 28 : Scan et transcription de la production de texte de CE1 de l'élève 1 283



Transcription :
[...] entendie de la porte un bruie : « toc-toc-toc » la sorciére / répondie : « entrer » un robot entre dans la maison et lui [...]

Figure 29 : Extrait du scan et de la transcription de la production de texte de CE2 de l'élève 3 003

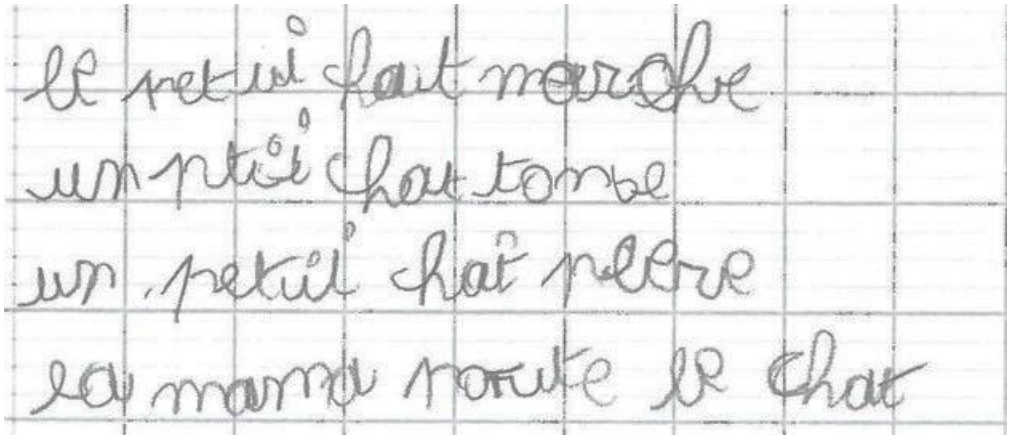
En classe de CP principalement, il peut arriver que la lettre *a* soit mal accentuée et que l'apprenant ou l'apprenante ait utilisé un accent aigu au lieu d'un accent grave, écrivant la lettre *á* au lieu de la lettre *à*. Ce phénomène est reproduit dans la transcription.



Transcription :
le petite chat se promener. é il / tonba <revision/> du trotoire é c est la tète qui / tonbe la promière. é il á vétréma la <revision/> / qui avétrémaala <revision/>tète.

Figure 30 : Scan et transcription de la production de texte de CP de l'élève 1 573

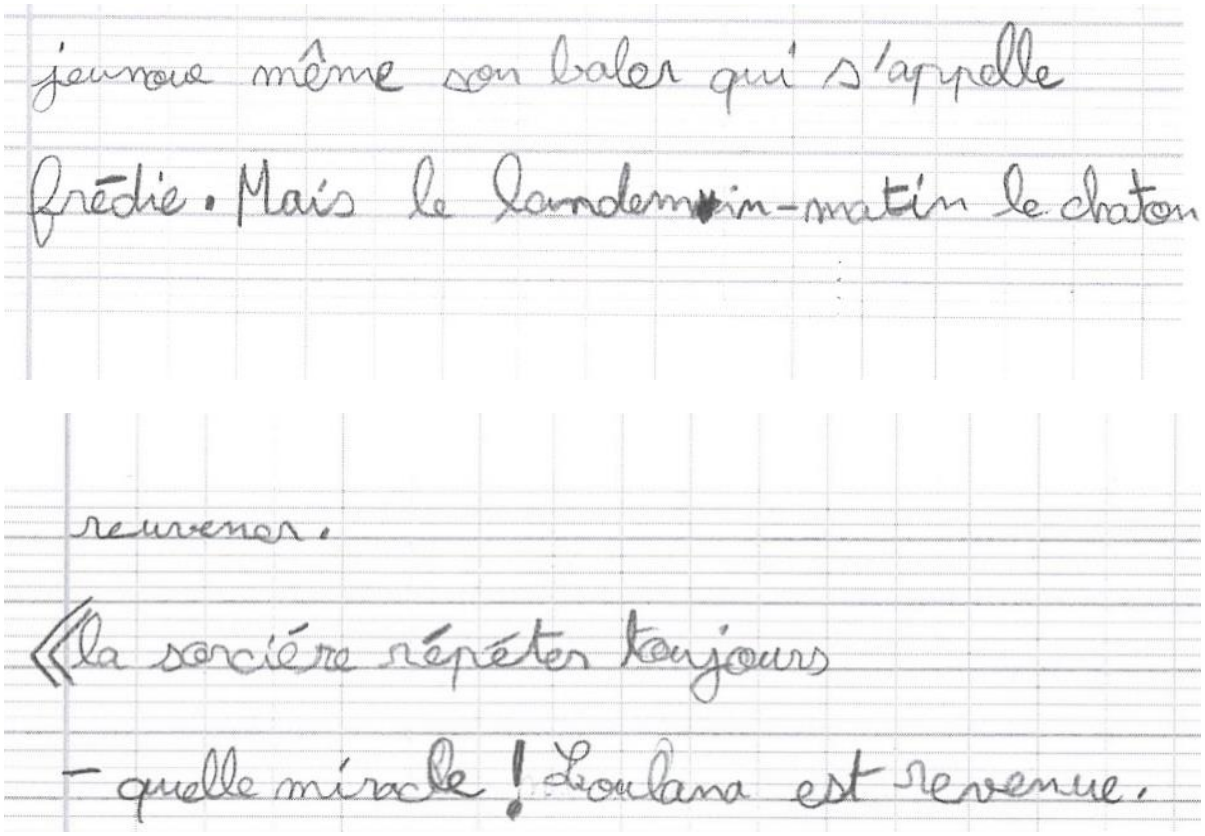
S'il ne nous a pas paru important de conserver de nombreux éléments de mise en page tels que l'alignement à la marge et les sauts de ligne, ce n'est pas le cas des retours à la ligne. Cette décision s'explique pour plusieurs raisons. En premier lieu, elle est née d'un souci pratique de visualisation. En effet, il nous a paru important, en particulier pour les utilisateurs peu habitués à l'usage des corpus de pouvoir afficher une transcription alignée sur le scan afin d'en faciliter la lecture. En second lieu, et c'est là une raison bien plus déterminante, nous estimons qu'un retour à la ligne peut avoir une signifiance. Chez des scripteurs experts, un retour à la ligne peut indiquer un nouveau paragraphe. Chez des scripteurs débutants, il peut également indiquer une nouvelle phrase ou une nouvelle unité syntaxique, comme dans l'exemple suivant. En classe de CP, il peut également correspondre à la description d'une nouvelle image.



Transcription :
le petui chat m<revision/>arche // un ptii chat tonbe // un petii chat plere // la mama porite le chat

Figure 31 : Scan et transcription de la production de texte de CP de l'élève 218

Le caractère *retour à la ligne* étant un caractère qui a plusieurs usages en informatique, il est ambigu. Il n'était donc pas possible de simplement indiquer un retour à la ligne dans la production par la touche *retour à la ligne* du clavier. Un symbole a donc été choisi pour le représenter. Le second avantage de cette solution est qu'elle permet de distinguer les retours à la ligne en fin de support d'écriture, qui peuvent ne pas avoir de signification textuelle, des retours à la ligne avant la fin du support qui, nous le supposons, peuvent marquer une fin de phrase ou de paragraphe. Les premiers sont identifiés par le symbole *slash* **/** tandis que les seconds sont identifiés par le symbole *double slash* **//**. Dans un souci d'affichage, nous avons également signalé le changement de page par l'ajout du symbole *dièse* **#** aux symboles précédents.



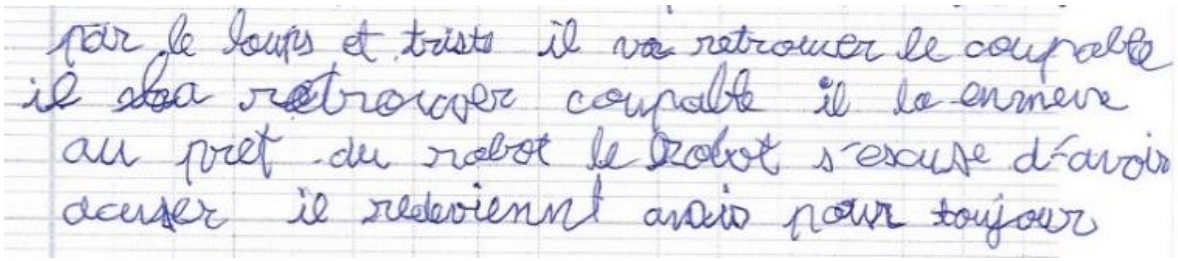
Transcription :
[...] jeunoue même son baler qui s'appelle / frédie. Mais le landemin-matin le chaton /# reuvener. //
« la sorciére répéter toujours // - quelle miracle ! Loulana est revenue.

Figure 32 : Extrait du scan et de la transcription de la production de texte de CE2 de l'élève 1 339

5.1.3. Indications de l'incertitude du transcripteur

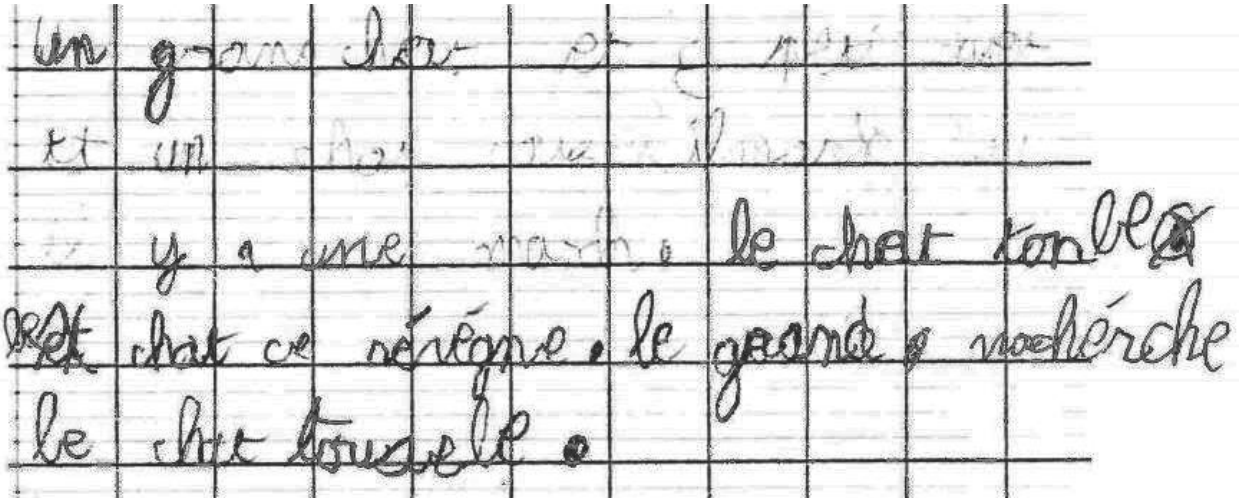
Enfin, comme nous l'avons mentionné en introduction, la lecture de ces productions de texte peut parfois s'avérer ardue et il n'est pas toujours possible de savoir avec certitude ce que l'élève a écrit ou a voulu écrire. Pour marquer ces cas et ainsi en informer l'utilisateur ou

l'utilisatrice, deux balises sont employées. La balise **<unsure>***ElémentPeuSûr***<unsure>** qui permet de signaler les doutes dans la transcription et la balise **<illisible/>** pour les cas où l'identification des lettres même approximativement n'est pas possible.



Transcription :
[...] il <revision/>la retrouver coupable il la <unsure>enmene</unsure> / / [...]

Figure 33 : Extrait du scan et de la transcription de la production de texte de CE2 de l'élève 239



Transcription :
un grand chat et <illisible/> petit <illisible/> / et un chat <illisible/> il <illisible/> / y a une march. le chat tonbl<revision/> / et chat ce révégne. le grande vachérche / le chat tousele.

Figure 34 : Scan et transcription de la production de texte de CP de l'élève 207

Les choix exposés ici sont les choix finaux ou arrêtés pour le moment dans l'étape de transcription. Mais il ne faudrait pas croire que ces choix ont été une évidence dès le départ. Ils n'ont cessé d'évoluer tout au long de l'élaboration du corpus et sans doute pourrait-on distinguer deux principaux facteurs de cette évolution.

Au début d'un projet, il y a toujours une part d'inconnu. En l'occurrence, il nous était difficile de connaitre à l'avance le contenu des productions. Nous avons donc commencé par annoter un grand nombre d'éléments plus spécifiques au CP, comme la présence de lettres mal formées, du prénom de l'élève, d'une hésitation entre deux lettres ou encore d'une séquence de lettres détachées. Au fur et à mesure de l'avancement du projet, du recueil et du traitement

des autres niveaux d'apprentissage, certains phénomènes qui nous ont paru au départ assez conséquents ont paru finalement assez marginaux, comme la présence d'une séquence de lettres détachées pour ne citer qu'un exemple. De même, au cours de la mise en place des outils de traitement, certains phénomènes nous ont paru inutilisables en termes de traitement et donc peu utiles à relever. C'est le cas de l'indécision entre deux lettres, par exemple. Suite à ces constats, nous avons fait le choix d'éliminer un certain nombre d'annotations de nos transcriptions.

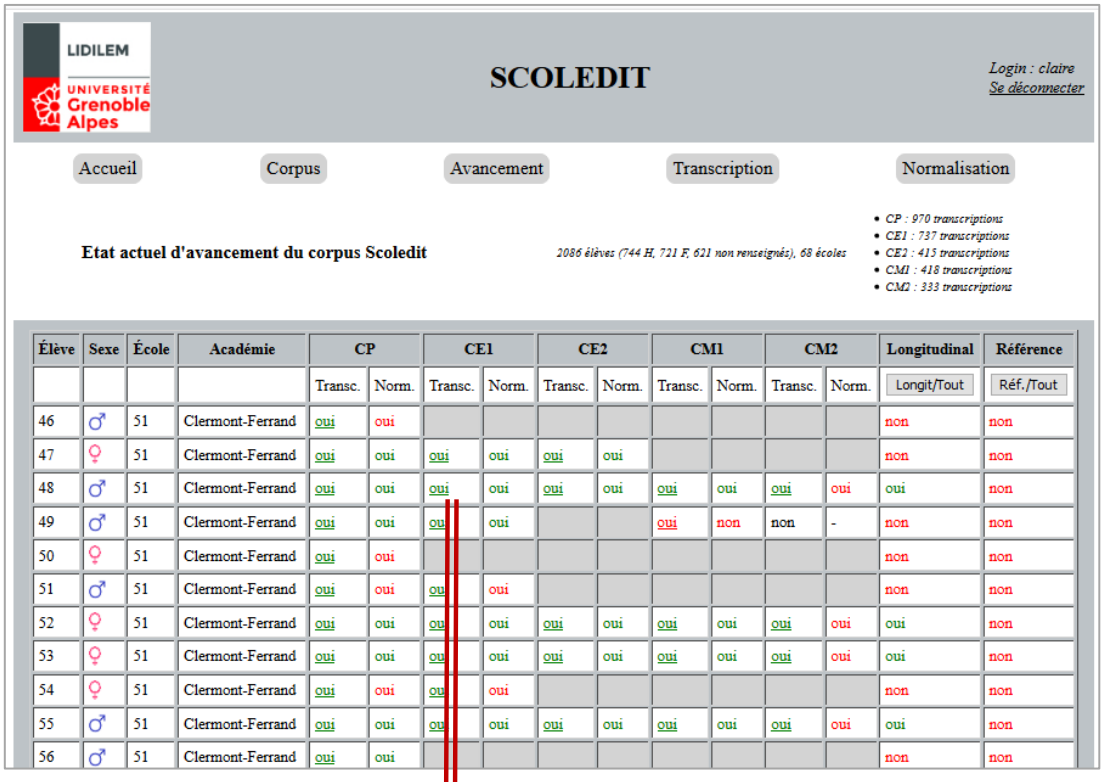
À l'inverse, tout comme certains phénomènes nous ont paru diminuer d'importance dans la suite du recueil, d'autres phénomènes sont apparus ou en plus grand nombre. Certaines annotations, comme la balise **<titre>***TitreDeLaProduction***</titre>**, ont donc dû être ajoutées par la suite. En effet, en classe de CP la présence d'un titre est extrêmement rare. Les titres commencent véritablement à apparaitre en classe de CE1 et leur présence se renforce en classe de CE2.

5.2. Outil de transcription

Au fur et à mesure du projet, nous nous sommes également dotés d'un outil de transcription, le site internet http://scoledit.org/scoledition/[72]. Ce site, à destination des étudiants et enseignants-chercheurs qui participent à la transcription et à la normalisation du corpus, a pour but de faciliter ces étapes (Figure 35) et participe ainsi à l'amélioration de la qualité des données. Il permet aussi de faciliter leur sauvegarde dans la base de données qui contient le corpus Scoledit. Ce site est également un outil de visualisation pour les membres du projet. Il permet en effet à la fois la visualisation des différentes versions des productions (scan, transcription et normalisation) (Figure 36) et l'état d'avancement de la numérisation et de l'enrichissement du corpus (Figure 37).

[72] Conception et réalisation : Claude Ponton, membre du projet *Scoledit*.

Figure 35 : Interface de transcription



Élève	Sexe	École	Académie	CP		CE1		CE2		CM1		CM2		Longitudinal	Référence
				Transc.	Norm.	Transc.	Norm.	Transc.	Norm.	Transc.	Norm.	Transc.	Norm.	Longit/Tout	Réf./Tout
46	♂	51	Clermont-Ferrand	oui	oui									non	non
47	♀	51	Clermont-Ferrand	oui	oui	oui	oui	oui	oui					non	non
48	♂	51	Clermont-Ferrand	oui	oui	oui	oui	oui	oui	oui	oui	oui	oui	oui	non
49	♂	51	Clermont-Ferrand	oui	oui	ou	oui			oui	non	non	-	non	non
50	♀	51	Clermont-Ferrand	oui	oui									non	non
51	♂	51	Clermont-Ferrand	oui	oui	ou	oui							non	non
52	♀	51	Clermont-Ferrand	oui	oui	ou	oui	oui	oui	oui	oui	oui	oui	oui	non
53	♀	51	Clermont-Ferrand	oui	oui	ou	oui	oui	oui	oui	oui	oui	oui	oui	non
54	♀	51	Clermont-Ferrand	oui	oui	ou	oui							non	non
55	♂	51	Clermont-Ferrand	oui	oui	ou	oui	oui	oui	oui	oui	oui	oui	oui	non
56	♂	51	Clermont-Ferrand	oui	oui									non	non

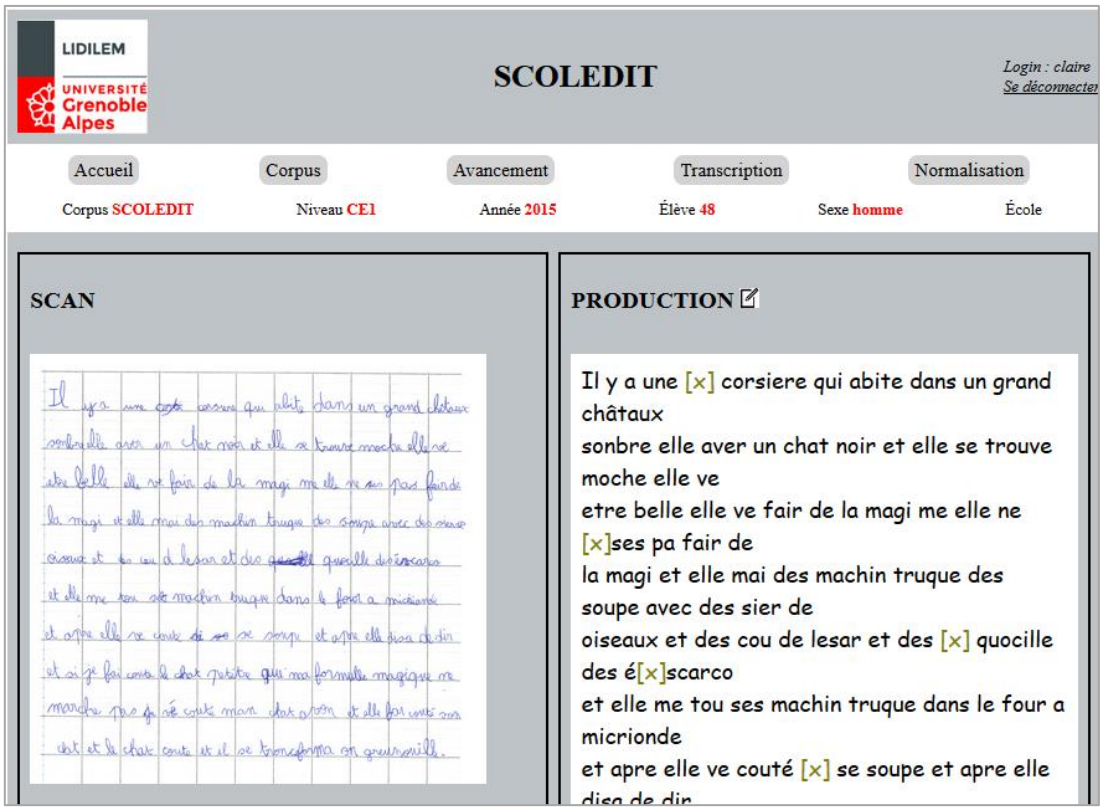
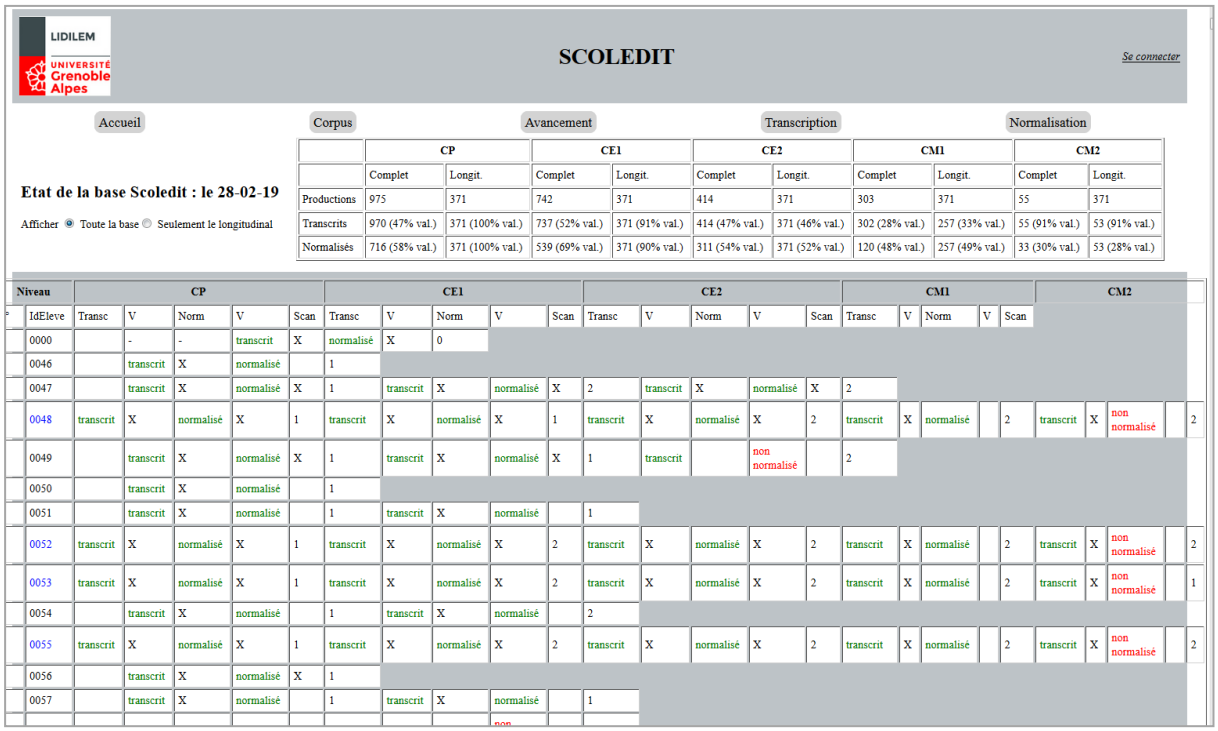


Figure 36 : Interface de visualisation du corpus (ensemble du corpus + production particulière)



SCOLEDIT

Se connecter

Accueil | Corpus | Avancement | Transcription | Normalisation

Etat de la base Scoledit : le 28-02-19

Afficher ◉ Toute la base ○ Seulement le longitudinal

	CP		CE1		CE2		CM1		CM2	
	Complet	Longit.	Complet	Longit.	Complet	Longit.	Complet	Longit.	Complet	Longit.
Productions	975	371	742	371	414	371	303	371	55	371
Transcrits	970 (47% val.)	371 (100% val.)	737 (52% val.)	371 (91% val.)	414 (47% val.)	371 (46% val.)	302 (28% val.)	257 (33% val.)	55 (91% val.)	53 (91% val.)
Normalisés	716 (58% val.)	371 (100% val.)	539 (69% val.)	371 (90% val.)	311 (54% val.)	371 (52% val.)	120 (48% val.)	257 (49% val.)	33 (30% val.)	53 (28% val.)

Niveau	CP					CE1					CE2					CM1					CM2				
IdEleve	Transc	V	Norm	V	Scan	Transc	V	Norm	V	Scan	Transc	V	Norm	V	Scan	Transc	V	Norm	V	Scan					
0000		-	-	transcrit	X	normalisé	X	0																	
0046		transcrit	X	normalisé		1																			
0047		transcrit	X	normalisé	X	1	transcrit	X	normalisé	X	2	transcrit	X	normalisé	X	2									
0048	transcrit	X	normalisé	X	1	transcrit	X	normalisé	X	1	transcrit	X	normalisé	X	2	transcrit	X	normalisé		2	transcrit	X	non normalisé		2
0049		transcrit	X	normalisé	X	1	transcrit	X	normalisé	X	1	transcrit		non normalisé		2									
0050		transcrit	X	normalisé		1																			
0051		transcrit	X	normalisé		1	transcrit	X	normalisé		1														
0052	transcrit	X	normalisé	X	1	transcrit	X	normalisé	X	2	transcrit	X	normalisé	X	2	transcrit	X	normalisé		2	transcrit	X	non normalisé		2
0053	transcrit	X	normalisé	X	1	transcrit	X	normalisé	X	2	transcrit	X	normalisé	X	2	transcrit	X	normalisé		2	transcrit	X	non normalisé		1
0054		transcrit	X	normalisé		1	transcrit	X	normalisé		2														
0055	transcrit	X	normalisé	X	1	transcrit	X	normalisé	X	2	transcrit	X	normalisé	X	2	transcrit	X	normalisé		2	transcrit	X	non normalisé		2
0056		transcrit	X	normalisé	X	1																			
0057		transcrit	X	normalisé		1	transcrit	X	normalisé		1														

Figure 37 : Interface de visualisation de l'état d'avancement

5.3. Processus de transcription

Cette tâche a nécessité l'intervention de nombreuses personnes. Au fur et à mesure du projet, plusieurs étudiants ont été recrutés chaque année pour réaliser des transcriptions. La plupart de ces étudiants ont été recrutés au sein du master *Industries de la langue* de l'UGA et sont donc sensibilisés aux enjeux de numérisation. Assez rapidement, il nous a paru important de les faire travailler par deux. La communication sur leurs doutes réciproques semble en effet augmenter la qualité des transcriptions.

Chaque nouveau duo d'étudiants a été formé selon un protocole identique. Tout d'abord, chacun d'eux a été invité à réaliser quelques transcriptions de manière indépendante à l'aide du guide de transcription. Puis, leurs transcriptions ont été mises en commun et vérifiées par l'un des membres du projet. Une fois le guide de transcription compris, les transcriptions à réaliser ont été réparties entre les étudiants. Ils étaient cependant invités à communiquer entre eux, voire à se relire l'un l'autre.

L'ensemble des transcriptions a été ou est en cours de vérification par un des membres du projet. In fine, chaque transcription est travaillée par au moins deux personnes.

Malgré ce processus de transcription et de vérification, il est nécessaire de garder à l'esprit que les erreurs subsistent et que le corpus est amené à être révisé sans cesse.

Il aurait été intéressant de procéder à l'évaluation du guide de transcription et ainsi de la fiabilité de nos données par un calcul de l'accord inter-annotateurs. Cependant, ce travail n'a pas pu être réalisé par manque de temps. L'annotation réalisée dans cette étape de transcription est peu fine et laisse peu de place à l'interprétation. Les transcripteurs ont donc exprimé assez peu de désaccords, nous n'avons pas priorisé l'évaluation de cette étape et avons préféré garder le temps dédié à l'évaluation pour d'autres processus.

6. Diffusion du corpus

Dès le début du projet, un des objectifs définis était l'élaboration d'une ressource linguistique accessible par les communautés enseignantes et scientifiques. Pour ce faire, en parallèle du processus d'élaboration du corpus et des outils d'exploitation, un site internet de diffusion du corpus a été mis en place, disponible à l'adresse http://scoledit.org/scoledit/. Celui-ci doit permettre la mise à disposition du corpus (Figure 38), des outils d'exploitation (Figure 39) et des résultats de recherches obtenus. Il vise principalement les membres de la communauté enseignante : enseignants et enseignantes, formateurs et formatrices, conseillers et conseillères pédagogiques, et les membres de la communauté scientifique : linguistes, didacticiens et didacticiennes, spécialistes du TAL.

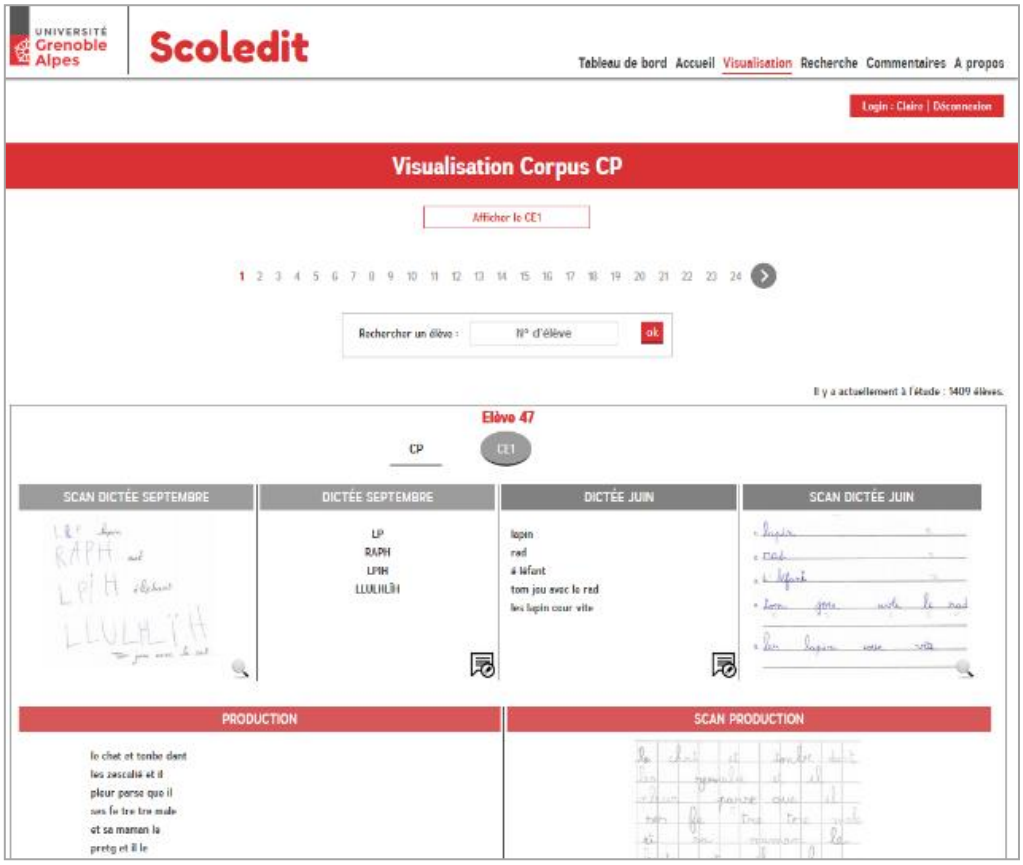


Figure 38 : Interface de visualisation du corpus

La figure 38 permet de se représenter la visualisation des productions sur le site *Scoledit*. Nous avons fait le choix de représenter dans une même fenêtre l'ensemble des productions de l'élève pour une année donnée.

Figure 39 : Exemple d'outil d'exploration du corpus

À l'origine, ce site devait également servir d'interface entre l'équipe du projet *Scoledit* et les utilisateurs et utilisatrices du corpus, qu'ils et elles soient enseignants ou chercheurs. Néanmoins, cet aspect interactif n'a pas porté ses fruits. Il était prévu que les utilisateurs puissent poster des commentaires pour indiquer la présence d'une erreur ou d'une approximation ou pour suggérer une modification dans la transcription (Figure 40). Cette fonctionnalité devait donc nous permettre d'améliorer la qualité de nos données au fur et à mesure de l'usage de notre corpus. Il était également prévu un espace de commentaires pour que les utilisateurs et utilisatrices puissent apporter un avis, proposer une modification ou encore énoncer un besoin d'outil ou de fonctionnalité pour utiliser le corpus.

Figure 40 : Commentaires laissés dans l'espace de commentaires à propos de la transcription

Malheureusement, cette interaction n'a pas eu lieu et seuls quelques commentaires ont été postés sur le site. Plusieurs raisons peuvent être suggérées pour expliquer cela, principalement la lenteur de la numérisation du corpus, de l'élaboration de l'outil de visualisation et des outils de traitement. En effet, alors que nous recueillions les textes de CE2, et que par conséquent nous informions principalement les enseignants de CE2 de l'existence de ce site, seules les productions de CP étaient visibles. Un décalage de deux à trois ans s'est ainsi créé entre le niveau enseigné par les utilisateurs et le niveau des productions disponibles. Ceci s'explique par le temps nécessaire à la transcription, à la normalisation et à la vérification des productions mais également par le temps nécessaire à l'élaboration du site, qui doit être adapté chaque année. Pour mieux comprendre, précisons que pour chaque étape de développement, différentes étapes préalables sont nécessaires : 1/ la recherche du financement nécessaire, 2/ la recherche de la ou les personnes compétentes, 3/ le développement de la ressource ou de l'outil.

De même, même si ce sont les membres permanents de l'équipe qui les développent, l'élaboration des outils de traitement prend du temps et leur application à l'ensemble du corpus également. Il y a donc peut-être eu une frustration des utilisateurs et utilisatrices à ne pouvoir que consulter le corpus et non à le manipuler.

Toutefois, avec plus de 300 utilisateurs (Figure 41), le site est actif. Il permet essentiellement aux enseignants de réaliser des visualisations simples des productions et a permis certains contacts avec des chercheurs.

role	count(id)
admin	4
Autre	59
Chercheur	32
Enseignant	248
Parent	1

Figure 41 : Utilisateurs du site Scoledit

Pour le moment, ce site est loin d'être achevé et seules les productions des deux premières années d'apprentissage, à savoir le CP et le CE1, sont visibles, mais il est prévu d'y faire apparaitre l'ensemble des productions. À terme, et ce travail est déjà entamé, il est prévu que le corpus puisse être parcouru en synchronie (toutes les productions d'une même année), en diachronie (toutes les productions d'un ou d'une élève), mais aussi, grâce aux outils de traitement, selon des critères linguistiques, tels qu'un mot ou une séquence de mots, une séquence de caractères, une catégorie grammaticale.

Afin de pallier le manque de réactivité de ce premier site, un deuxième site a été mis en place. Celui-ci est plus sommaire en terme d'interface mais l'ensemble des productions peut y être visualisé (http://scoledit.org/scoledition[73]).

7. Conclusion

L'ambition du projet *Scoledit* est de constituer un corpus longitudinal, largement accessible. Ce processus est un processus lent qui englobe de nombreuses étapes. La première de ces étapes, est le recueil du corpus durant cinq ans, afin de permettre le suivi des élèves. Puis, pour qu'un corpus puisse être exploité et diffusé, il est nécessaire de le numériser, c'est-à-dire le scanner et le transcrire. Cette étape de transcription n'est pas triviale et de nombreux choix ont été nécessaires, de même que la conception d'un outil spécifique. Enfin, pour que ce corpus puisse être diffusé, différentes plateformes ont dû être élaborées et sont toujours en cours de développement.

[73] Conception : Claude Ponton.

Chapitre 6 - La normalisation, une annotation déportée

1. Type de d'enrichissement choisi 121
2. Enjeux de la normalisation 122
3. Principes de normalisation 123
4. Conventions de normalisation 125
5. Processus de normalisation 131
6. Conclusion 132

1. Type de d'enrichissement choisi

Après un premier essai d'intégration des données de correction aux différentes formes erronées du corpus (Wolfarth, 2015), il a été choisi de produire une normalisation déportée, indépendante de la transcription, à l'instar de celle pratiquée dans les projets de C. Fairon *et al*. (2006). Ce choix se justifie notamment par le trop grand nombre de variations à la norme dans les écrits scolaires, notamment au cours des premières années d'apprentissage. Nous appellerons ces versions corrigées des productions des *normalisations*. Ce terme a été choisi pour plusieurs raisons. 1/ Utiliser le terme *correction* sous-tendrait qu'il y aurait une bonne façon d'écrire à partir de laquelle tous les écrits peuvent être corrigés ; utiliser le terme *normalisation* sous-tend que les productions vont être comparées à une norme qu'il reste à définir. 2/ Nous nous plaçons dans une perspective similaire à celle de C. Bonnet (1994), C. Fabre-Cols (2000) puis M.-L. Elalouf (2005) et C. Doquet et ses collègues (2017). Les productions des apprenants ne sont pas analysées uniquement par le prisme de leurs erreurs mais sont vues comme un registre de textes particulier, présentant des potentialités et des phénomènes linguistiques propres à ces écrits. À l'instar des projets visant la description des caractéristiques des écrits SMS et de leurs riches variations, la normalisation a pour but d'accéder au contenu linguistique des productions à l'aide d'outils automatiques. 3/ Enfin, il s'agit également de traduire l'importance de cette version. Dans certains projets, comme ce fut le cas dans un premier temps dans le projet coordonné par M.-L. Elalouf (2005), des versions orthographiées selon la norme sont conçues uniquement dans le but de faciliter la lecture. Plus tard, elles ont été utilisées à des fins d'analyse. Dans le cadre du projet *Scoledit*, les normalisations sont parties intégrantes de l'analyse.

2. Enjeux de la normalisation

Comme nous l'avons mentionné à plusieurs reprises, la plupart des outils de traitement sont conçus pour des écrits *standards*. Ils échouent à traiter efficacement les productions très éloignées de la norme comme les écrits scolaires. En l'état actuel des outils de TAL, il n'est donc pas possible de leur donner à analyser les transcriptions du corpus *Scoledit*. La production des normalisations, plus proches des écrits normés pour lesquels les outils de TAL ont été conçus, représente donc une étape cruciale pour l'exploitation du corpus.

Une fois les versions normalisées établies, il est possible de leur appliquer des outils tels que des étiqueteurs morphosyntaxiques, des analyseurs syntaxiques, etc. qui permettent d'annoter et d'analyser le corpus. Parallèlement, un lien peut être établi entre transcription et normalisation. La normalisation sert donc d'interface entre la transcription de la production et les outils d'exploitation.

L'enjeu est alors de déterminer les phénomènes linguistiques à normaliser et la façon dont on souhaite les normaliser. De ces choix dépendront les phénomènes qui pourront être analysés. Prenons quelques exemples.

(1) le chatpleur (78, CP)
[normalisation] le chat pleure

(2) la maman chat grdre set 3 petite chat un chat vaver la maman chat (670, CP)
[normalisation] La maman chat garde ses 3 petits chats <segmentation/> un chat va vers la maman chat

Dans l'exemple (1), seule la forme *le* pourrait être identifiée par un outil de traitement automatique basé sur un dictionnaire de formes, comme le sont la plupart des étiqueteurs morphosyntaxiques. Une normalisation de la segmentation en mots et de l'orthographe permettrait à la fois de reconnaitre les formes *chat* et *pleure* et d'étudier leurs variations orthographiques.

Dans l'exemple (2), aucun indice dans la production de l'élève ne laisse apparaitre de différenciation syntaxique entre les fragments *la maman chat garde ses 3 petits chats* et *un chat va vers la maman chat*. Un analyseur syntaxique analysera donc cet ensemble comme un seul fragment syntaxique. Pour contrer ce phénomène, l'ajout d'un élément – marque de ponctuation, balise, etc. – marquant la séparation entre ces deux fragments textuels est

nécessaire. Cet élément (dans l'exemple, la balise <segmentation/>) permet également de recenser le nombre de ponctuants manquants.

(3) [...] li plere miaou miaou, et quel-quin vin le / **réconcolier**. (500, CP)
*[normalisation] [...] il pleure miaou miaou, et quelqu'un vient le **reconsoler**.*

Il est parfois nécessaire de prioriser les phénomènes à normaliser. Dans l'exemple (3), l'élève a produit le segment « réconcolier », proche de la forme *réconcilier*. Néanmoins, dans ce contexte les formes *reconsoler* ou *réconforter* semble plus en adéquation avec l'intention de l'élève. Il est donc nécessaire de choisir entre une observation orthographique, entre les formes « réconcolier » et *réconcilier*, et une observation d'ordre lexical, entre les formes « réconcolier » et *reconsoler* ou *réconforter*.

(4) il **tombe** et il **pleurer**. (586, CP)
*[normalisation] Il **tombe** et il **pleurait**.*

Un choix un peu similaire se présente dans l'exemple (4). Il est nécessaire de choisir entre normaliser la forme *pleurer* par la forme *pleure* ou la forme *pleurait*. Choisir la première possibilité permet de révéler dans l'analyse les problèmes de cohérence des temps qui peuvent être rencontrés dans ces productions et de correspondre davantage à la norme scolaire. Choisir la deuxième possibilité permet de mettre l'accent sur les difficultés en morphologie verbale que rencontrent les apprenants. De plus, cette possibilité parait correspondre davantage à l'intention de l'élève et met ainsi en exergue les spécificités de ce type de production.

3. Principes de normalisation

Comme l'illustrent ces différents exemples, élaborer une convention de normalisation pour l'ensemble des productions et des niveaux d'apprentissage est un processus long et qui doit pouvoir faire l'objet de remaniements aussi fréquents que nécessaires. En effet, les choix de normalisation ont évolué au gré des discussions et des découvertes de nouveaux phénomènes linguistiques, niveau après niveau. Notons également que diverses interactions avec une équipe de recherche italienne menée par Lilia Teruggi (Università degli Studi di Milano-Bicocca) a encore permis d'affiner cette convention de normalisation. Pour finir, la normalisation retenue est régie par trois grands principes qui parfois se complètent, parfois s'opposent :

a. Normaliser les productions au plus près de la production initiale de l'apprenant ;
b. Normaliser en considérant les phénomènes que l'on souhaite étudier ;
c. Normaliser en faisant appel le moins possible à l'interprétation (en cas de doute la primauté est donnée à l'oral).

Pour mieux comprendre ce dernier principe, prenons l'exemple (5). Dans cet exemple, le segment « peré » peut être normalisé *a pleuré*, plus proche graphiquement parlant de la production de l'élève, ou *pleurait*, plus proche phonologiquement parlant. En vertu du principe 3), l'oral est privilégié et la deuxième possibilité est appliquée.

(5) […] il sai fai mal il peré et la maman chat […] (1152, CP)
[normalisation] […] il s'est fait mal <s/> il pleurait et la maman chat […]

De ce dernier principe émerge la nécessité pour certains phénomènes d'utiliser des balises génériques, englobant plusieurs possibilités, afin de laisser transparaitre le moins possible l'interprétation ou le style de l'annotateur.

(6) sa mamant il dormer le petit chat il a marcher (70, CP)
[normalisation] Sa maman elle dormait <segmentation/> le petit chat il a marché.

(7) [...] il di silete plé tu pe me prande ouit alé on rantre a la méson (1313, CP)
[normalisation] [...] il dit <dialogue> s'il te plait <s/> tu peux me prendre ? </dialogue> <dialogue> oui <s/> allez <s/> on rentre à la maison </dialogue>.

Tout comme dans l'exemple 2 donné précédemment, l'exemple 6 ne présente aucune marque de différenciation entre les fragments textuels « sa mamant il dormer » et « le petit chat il a marcher ». Pour les raisons évoquées précédemment (permettre des analyses syntaxiques notamment), il nous parait important de marquer cette différenciation. Cependant les possibilités sont multiples, prenons quelques exemples :

- Sa maman elle dormait**,** le petit chat il a marché.
- Sa maman elle dormait**.** Le petit chat il a marché.
- Sa maman elle dormait **et** le petit chat il a marché.

Il n'y a donc ce cas-là aucune norme unique attendue mais plusieurs conventions possibles. Afin de ne pas avoir à choisir entre ces différentes conventions, qui dépendent davantage du style du transcripteur que des choix du scripteur, nous proposons d'utiliser des balises génériques, comme la balise <segmentation/>.

Dans le domaine de la ponctuation et de la segmentation textuelle (segmentation en fragments textuelles ou en propositions), nous utilisons principalement trois balises génériques[74] :

- La balise <segmentation/> qui marque une rupture forte (changement de thématique, d'épisode ou de situation) entre deux fragments textuels (exemple (6)) ;
- La balise <s/> lorsqu'un signe de ponctuation est nécessaire mais qu'il n'y a pas de rupture forte entre les deux propositions qu'il relie (exemple (7)) ;
- La balise <dialogue></dialogue> qui marque le début et la fin d'un tour de parole dans un dialogue. En cas de manque, cette balise peut remplacer un signe de ponctuation introductif ou conclusif d'un dialogue (exemple 6).

Ces balises ne sont utilisées qu'en cas de manque ou de mauvaise utilisation d'un ponctuant ou d'un séparateur.

(8) [...] sa révéill la maman vien laidai. (562, CP)
[normalisation] [...] ça réveille la maman ***<omission type="pronom"/>*** *vient l'aider.*

Une balise générique est également utilisée pour marquer l'omission de certains mots (verbe, nom, pronom, adjectif (qualificatif), adverbe, préposition, déterminant). La transcription donnée en exemple 7, pourrait être normalisée de différentes façons :

- [...] ça réveille la maman qui vient l'aider.
- [...] ça réveille la maman, elle vient l'aider.

Là encore, pour ne pas avoir à faire un choix, une balise générique est déterminée, la balise <omission type="CATEGORIE"/>.

4. Conventions de normalisation

Comme pour l'étape de transcription, l'ensemble des conventions de normalisation ont été consignées dans un guide de normalisation. Nous en récapitulons ici les grandes lignes. Le détail de ces conventions figure dans le guide en annexe (Annexe 6).

Tout d'abord, précisons que ne sont retenus dans la normalisation que les éléments textuels produits par l'enfant : les mots, ou groupes de lettres, et la ponctuation. Les balises introduites lors de la phase de transcription ne sont pour la plupart pas conservées, à l'exception de

[74] Pour une explication plus détaillée de la réflexion qui a mené à choisir l'usage de balises génériques, se référer à l'article « Transcrire et normer un corpus scolaire, pour quelles analyses ? » (Wolfarth, Brissaud, et al., 2018). Une partie des choix effectués dans le traitement de la segmentation en fragments textuels y est explicitée.

quelques balises ou marqueurs, comme le retour à la ligne, les marques d'incertitude (<unsure></unsure> et <illisible/>) et les balises de titre.

4.1. À l'échelle lexicale et micro-syntaxique

Comme pour la plupart des projets de constitution de corpus scolaires proposant une version normée (projet Ecriscol, Doquet *et al.*, 2017 ; Elalouf, 2005 ; Roubaud, 2017), nous avons fait le choix de rétablir :

- la segmentation en mots attendue (« chatpleur », *chat pleure* dans l'exemple (1)) ;
- une orthographe standard[75] (« pleur », *pleure* dans l'exemple (1) et « cha », *chat* dans l'exemple (9)) ;
- la morphologie verbale attendue (« prena », *prit* dans l'exemple (9)).

(9) elle **prena** le petit **cha** (584, CP)
*[normalisation] elle **prit** le petit **chat***

Ces premières corrections sont essentielles pour l'emploi des outils d'annotation et de traitement automatique, comme les étiqueteurs morphosyntaxiques et les outils d'analyse textométrique, comme *TXM* augmenté de *TreeTagger*, souvent basés sur une analyse des formes du corpus. Ces corrections, accompagnées d'un alignement entre transcription et normalisation (cf. partie 3), rendent également possible l'élaboration de listes de variantes orthographiques employées par les élèves pour une forme normée donnée.

Dans un certain nombre de productions, on constate l'absence de certaines formes (la marque de négation dans l'exemple 124(10) et la reprise pronominale dans l'exemple (11) qui rend la structure syntaxique non standard à l'écrit et qui peut donc faire échouer les outils d'analyse. Lorsque l'élément absent est identifiable de manière non ambigüe (*n'* dans l'exemple (10)), il est introduit dans la normalisation. Si plusieurs possibilités sont disponibles (reprise pronominale sous forme de pronom relatif *qui* ou de pronom personnel *il* dans l'exemple (11)), un marqueur est introduit dans la normalisation sous la forme de la balise <omission type="CATEGORIE"/> où CATEGORIE correspond à la catégorie du mot attendu (verbe, nom, pronom, adjectif (qualificatif), adverbe, préposition, déterminant).

(10) mai il a pas vus le trautoir (562, CP)
*[normalisation] Mais il **n'a** pas vu le trottoir.*

[75] En cas de mot concerné par les rectifications orthographiques de 1990, le choix de l'élève est respecté.

(11) il été t'une foi un petit chat se promené acoté de sa litière (1055, CP)
[normalisation] Il était une fois un petit chat **<omission type="pronom"/>** *se promenait à côté de sa litière.*

En cas de choix lexicaux non conventionnels, tant en terme de vocabulaire (*souricette* dans l'exemple (12)) qu'en terme de registre de langue (exemple (13)), en vertu du principe de conservation des choix de l'élève, ils sont conservés dans la normalisation. En effet, il ne s'agit pas d'erreurs à proprement parler mais plutôt de variantes qui peuvent ne pas correspondre à la norme scolaire ou à la norme standard, comme c'est le cas du terme *souricette,* souvent inconnu des dictionnaires mais issu d'une construction morphologique plausible à partir de formes existantes.

À l'inverse, le segment « réconcolier » (exemple (3)) n'est pas issu d'une construction morphologique à partir d'une forme existante. Le segment « réconcolier » n'est donc pas conservé dans l normalisation.

(12) il demander à c'est amis qui / l'aider toujour à chasser des petites **sourissette** / ai des oiseaux. (586, CE1)
[normalisation] il demandait à ses amis qui l'aidaient toujours à chasser des petites ***souricettes*** *et des oiseaux.*

(13) il se ro **refou** en bébé (114,CP)
[normalisation] il se ***refout*** *en bébé.*

Afin de pouvoir analyser au mieux les productions, les accords en genre, en nombre et en personne sont rétablis dans la normalisation (exemple (14)).

(14) et les **chaton** aussi (1143, CP)
[normalisation] et les ***chatons*** *aussi*

4.2. À l'échelle syntaxique

Toujours en accord avec le principe d'approcher au plus près la production et l'intention de l'élève, nous reportons dans la normalisation les constructions syntaxiques choisies par lui ou elle, même lorsqu'il s'agit de constructions issues de l'oral (exemple (15)).

(15) Le cha il mache (61, CP)
[normalisation] Le chat il marche

De même, les temps choisis par les élèves sont respectés, parfois au détriment du principe de cohérence des temps (exemple (4)). Lorsqu'il y a ambigüité sur le temps du verbe choisi en raison d'une homophonie, l'oral prime sur l'écrit.

Un des aspects caractéristiques des productions d'apprenants, particulièrement au cours des premières années d'apprentissage, est l'absence d'un grand nombre de marques de séparation (ponctuation ou connecteurs) entre les différentes propositions ou fragments textuels du texte. Or, contrairement à l'orthographe pour laquelle il existe une norme *forte*, selon les mots de J. Popin et C. A. Thomasset (1998, p. 13), qui laisse peu de place à la variation, la ponctuation « échappe au concept de norme » et « entre dans le domaine du standard ». D. Bessonnat (1991) distingue également une ponctuation prescriptive, régie par une norme, d'une ponctuation facultative, appartenant au domaine de la stylistique. La ponctuation est donc sujette à variation selon les scripteurs. Des conventions simples de normalisations sont donc difficiles à établir et peuvent ne pas faire l'unanimité même parmi des scripteurs experts.

Néanmoins, la ponctuation et les connecteurs représentent un réel enjeu pour les possibilités d'analyse ultérieures puisqu'ils représentent des marqueurs sur lesquels se fondent certains outils d'analyse syntaxique. Les conventions de normalisation à ce sujet ont donc été établies en tension entre ces deux contraintes : le refus d'imposer un choix de ponctuation contestable par un scripteur expert ; la nécessité d'insérer des marqueurs qui délimitent les différents fragments textuels lorsqu'ils sont absents.

En cas de ponctuation non problématique, celle-ci a bien évidemment été reprise dans la normalisation. En cas de ponctuation surnuméraire (comme la virgule dans exemple (16)), celle-ci a été supprimée. En revanche, il semble important d'expliciter les choix réalisés en cas d'absence de marques de segmentation.

(16) [...] Il était un fois**,** une sorcière avec un chat noir. [...] (3001, CE1)
[normalisation] [...] Il était une fois une sorcière avec un chat noir. [...]

Lorsque cela était possible de manière suffisamment consensuelle, des signes de ponctuation ont été insérés dans la normalisation (virgules, points et points d'interrogation notamment). Dans les autres cas, nous avons des balises génériques pour spécifier les marqueurs de segmentation absents.

En cas de présence d'un point, non suivi d'une majuscule, la majuscule est rétablie (exemple (17)). De même, en cas de présence d'une majuscule mais d'absence d'une ponctuation forte ou inversement, situé à un endroit où une ponctuation pourrait être insérée, le point (exemple (18)) ou le point d'interrogation, en cas d'interrogation directe, sont rétablis (exemple (19)).

(17) le cha et tonbe sur le tapi**.** **le** cha pler (48, CP)
[normalisation] le chat est tombé sur le tapis. ***Le*** *chat pleure*

(18) Le petit chat sest fé male Le peti chat a male (1531, CP)
[normalisation] Le petit chat s'est fait mal. Le petit chat a mal

(19) - bon jour la sorsiere commens allé vous. (1927, CE2)
[normalisation] <dialogue> - Bonjour la sorcière <s/> comment allez-vous ? </dialogue>

Outre le point et le point d'interrogation dans les cas que nous venons de spécifier, la virgule a également été rétablie dans un cas très particulier : en cas d'énumération d'au moins trois éléments.

En cas d'énumération sont introduits dans la normalisation à la fois le signe virgule entre chacun des éléments de l'énumération, sauf les deux derniers, et la conjonction *et* entre les deux derniers éléments (exemple (20)). En cas d'énumération de deux éléments, seule la conjonction *et* est rétablie.

La notion d'énumération comprend les énumérations de noms, d'adjectifs ou de verbes. On considère également un enchainement de propositions comme énumération, à condition que le sujet grammatical reste le même (exemple (20)).

(20) Le peti cha senva il tond il révége / sa maman sa maman le souoige. (97-CP)
[normalisation] Le petit chat s'en va, il tombe et il réveille sa maman <segmentation/> sa maman le soigne.

Dans cet exemple, les trois premières propositions ont le même sujet, *le petit chat*, on considère donc qu'il s'agit d'une énumération. En revanche, la dernière proposition introduit un nouveau sujet et n'est donc plus incluse dans cette énumération.

Ces différentes configurations n'épuisent pas les cas où un marqueur de segmentation est nécessaire sans être présent dans la production de l'enfant. Dans les cas de changement de thématique/épisode/situation, repérables notamment dans le changement de détermination (de l'indétermination à la détermination), du sujet grammatical, des temps verbaux, on considèrera qu'il y a changement de « fragment textuel ». Ces changements sont indiqués par la balise <segmentation/> (exemple (21)).

(21) la maman chat grdre set 3 petite chat un chat vaver la maman chat (670, CP)
[normalisation] La maman chat garde ses 3 petits chats <segmentation/> un chat va vers la maman chat

Dans cet exemple, on constate un changement de temps verbal, passant d'une action longue à une action courte située dans le temps.

Dans les autres cas où un signe de ponctuation semble nécessaire, sans que s'impose une ponctuation forte comme le point, on utilise la balise **<s/>** (exemple (22)).

> (22) Il pler sa maman vien le / chré (1125 CP)
> *[normalisation] Il pleure* **<s/>** *sa maman vient le chercher.*

Les deux fragments textuels sont liés ici par un lien de cause à effet qui nécessite une ponctuation moins forte.

Nous avons également choisi, afin de faciliter l'exploitation des normalisations, de marquer les épisodes de dialogues qui utilisent généralement des structures syntaxiques spécifiques. Au sein de ces épisodes de dialogue, chaque nouveau tour de parole est signalé à l'aide des balises **<dialogue>** (au début) **</dialogue>** (à la fin) (exemple (23)).

> (23) le chat dit : Bonjour méchante et orible sorciere j'ai apris que tu allais dominer le monde. Oui c'est sa et comment t'appelle tu je m'appelle jean - pomme de terre. Sa te dit con fase équipe ensemble [...] (1336, CM1)
> *[normalisation] le chat dit : <dialogue> Bonjour méchante et horrible sorcière <s/> j'ai appris que tu allais dominer le monde. </dialogue> <dialogue> Oui c'est ça et comment t'appelles-tu ? </dialogue> <dialogue> Je m'appelle Jean-Pomme de terre. </dialogue> <dialogue> Ça te dit qu'on fasse équipe ensemble ? </dialogue> [...]*

4.3. Marques de structuration

Quelques balises permettent également de structurer la production de l'élève, notamment la balise <p/> (exemple (24)) qui reprend les symboles // de la transcription et qui permet de marquer un retour à la ligne intentionnel, que ce soit pour marquer un nouveau paragraphe ou un nouveau fragment textuel. Notons qu'en classe de CP, ce procédé marque bien souvent la description d'une nouvelle image (cf. Chapitre 4 - 1 pour la description des consignes de recueil).

> (24) Le chat eon ilavan // Le chateon et blécé // Le chateon i pler // [...] (58, CP)
> *[normalisation] Le chaton il avance.* **<p/>** *Le chaton est blessé.* **<p/>** *Le chaton il pleure.* **<p/>** *[...]*

Les balises <titre> (début) et </titre> (fin), présentes dans la transcription et dans la normalisation, permettent également de relever la présence d'un titre (exemple (25)) qui génère souvent une syntaxe quelque peu particulière.

(25) Le chat Il était une fois un chat (1845, CE2)
[normalisation] ***<titre>*** *Le chat* ***</titre>*** *Il était une fois un chat*

4.4. Marques d'incertitudes

Tout comme lors de l'étape de transcription, il peut y avoir des segments non lisibles ou non interprétables et des segments où la normalisation est peu sure. Les balises <incomprehensible/> (exemple (26)) et <unsure> etc. </unsure> (exemple (27)) permettent d'identifier ces zones d'indécision.

(26) l ti chat pas c ese bébe cha (1363, CP)
[normalisation] le petit chat ***<incomprehensible/>*** *bébé chat*

(27) la maman elle ve **la deté** (1183, CP)
[normalisation] la maman elle veut ***<unsure>****l'adopter****</unsure>****.*

5. Processus de normalisation

La version des conventions de normalisation présentée ici est une version relativement stabilisée mais ce processus d'élaboration a été très long (de mars 2015 à décembre 2018) et a beaucoup évolué, notamment sur les questions de ponctuation. On retrouve ces mêmes questionnements dans différents projets de corpus scolaires (Boré & Elalouf, 2017, par exemple). Le protocole de normalisation s'est notamment enrichi grâce à des échanges avec les équipes de recherche des différents projets français[76], mais également avec l'équipe italienne de Lilia Teruggi, comme signalé précédemment.

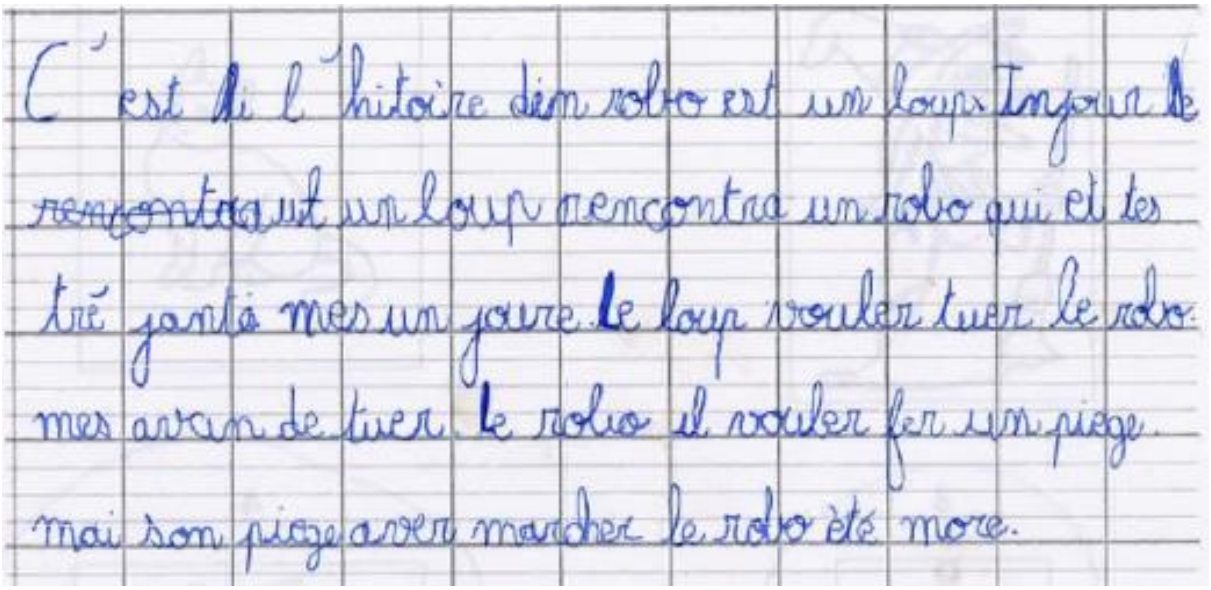
Tout comme pour l'étape de transcription, différentes personnes, chercheurs et étudiants, ont contribué à ce travail de normalisation du corpus, permettant d'éprouver le protocole établi et de le faire évoluer au fur et à mesure des phénomènes rencontrés. Comme pour l'étape précédente, les personnes chargées de la normalisation ont principalement travaillé par équipe de deux, afin de faciliter les échanges pour une meilleure qualité des données.

Actuellement, une phase de validation de la normalisation de chaque production est en cours de réalisation par les membres chercheurs du projet. Néanmoins, le protocole de normalisation

[76] De ces échanges qui ont eu lieu au sein du projet *E-CALM* est en train d'émerger un protocole de normalisation commun. Nous avons veillé à l'interopérabilité des deux formats.

ayant beaucoup évolué, il n'est pas facile d'obtenir une normalisation homogène de l'ensemble de ces données, ni de faire une évaluation de ce processus.

À l'issue de cette première étape d'enrichissement, on obtient donc pour chaque production, trois versions (Figure 42) : un scan, une transcription et une version normalisée.



Transcription :

C'est <revision/> l'histoire din robo est un loup. Injour <revision/> / <revision/> un loup rencontra un robo qui et tes / tré jant<revision/>i mes un joure. Le loup vouler tuer le robo. / mes avan de tuer. Le robo il vouler fer un piege. / mai son piege aver marcher le robo èté more.

Normalisation :

C'est l'histoire d'un robot et un loup. Un jour un loup rencontra un robot qui était très gentil mais un jour le loup voulait tuer le robot. Mais avant de tuer le robot il voulait faire un piège. Mais son piège avait marché <s/> le robot était mort.

Figure 42 : Exemple de trois versions d'une production (200, CE1)

Comme nous l'avons dit précédemment, la normalisation ne doit être qu'une interface entre les outils d'exploration automatique et les transcriptions des productions. Il est donc nécessaire de construire un lien entre ces transcriptions et leurs normalisations. Ce lien sera établi par un alignement automatique qui fait l'objet des prochains chapitres.

6. Conclusion

L'étape d'enrichissement linguistique d'un corpus est une étape particulièrement stratégique et délicate puisque c'est d'elle que découleront les analyses linguistiques qui pourront être effectuées sur le corpus. À rebours des approches majoritaires, nous avons choisi de procéder

à un enrichissement déporté et non embarqué, comme le sont souvent les annotations. Cette approche nous contraint à élaborer une version normalisée pour chaque production. Tout comme lors de l'annotation d'un corpus, la phase de normalisation nécessite d'opérer des choix quant aux phénomènes à considérer ou non.

Partie 3 - Application de l'approche par comparaison : aligner transcriptions et normalisations

Chapitre 7 - Algorithmes d'alignement, revue des méthodes 137
Chapitre 8 - *AliScol*, un système d'alignement pour les écrits scolaires 157
Chapitre 9 - Évaluation de l'aligneur *AliScol* 181

Chapitre 7 - Algorithmes d'alignement, revue des méthodes

1. Introduction 137
2. État de l'art des méthodes d'alignement 143
3. Conclusion 156

1. Introduction

Un grand nombre d'outils de traitement automatique des langues et d'exploration linguistique prennent pour unité linguistique le mot, la forme ou le token, trois unités proches que nous discutons ci-après. Il est donc nécessaire de disposer d'un corpus dont les formes sont identifiables pour ce type d'outils. Cependant, comme nous l'avons spécifié précédemment, le grand nombre d'écarts à la norme présents dans les corpus scolaires empêche la reconnaissance des formes. De ce fait, on choisit souvent d'apposer une couche de correction ou de normalisation au corpus. L'association entre les formes produites et les formes normalisées est alors nécessaire (David & Doquet, 2016).

Lorsque la correction d'une forme ou d'un groupe de formes se fait à l'aide de balises (cf. Chapitre 4) cette association peut sembler relativement évidente. Au sein du projet *Scoledit*, en revanche, nous avons choisi d'adopter une approche par comparaison (cf. Chapitre 4 - 3). Cette approche nécessite de produire une normalisation de l'ensemble de la production et un alignement de cette normalisation avec la transcription des productions des élèves. Cet alignement forme à forme permet alors d'identifier les formes du corpus, préambule aux comparaison mais également à l'emploi d'outils TAL plus génériques.

1.1. Définition des notions

Commençons par définir les termes et les notions utilisés, les définitions et appellations pouvant différer d'un domaine de la linguistique à l'autre. Nous n'utilisons pas la notion de *mot* qui est une notion trop floue et dont les définitions varient régulièrement. Nous lui préférons les notions de *formes* et de *segments*, que nous allons définir.

Dans le domaine du TAL, les unités utilisées sont les *tokens*. Nous sommes donc amenée à utiliser ce terme lorsque nous parlons d'un outil TAL comme *TreeTagger*. B. Sagot et P. Boullier

(2008) donnent la définition suivante : « le terme *token* pour dénoter une séquence de caractères présente dans le corpus et séparée de ses voisins par des espaces ou par certaines autres marques typographiques (ponctuation, etc.) ». À l'image de celle-ci, beaucoup de définitions du terme *token* ne prennent en considération que les « mots », délaissant les autres caractères tels que les signes de ponctuation. Or, dans les faits, les outils TAL prennent en compte à la fois les mots et certains signes de ponctuation, nous préférons donc utiliser la définition du *Oxford Handbook of computational linguistics* (Mitkov, 2003) qui décrit la tokenisation comme l'action de segmenter en « linguistic units such as words, punctuation, numbers, alphanumerics, etc. » (p. 201). Dans ce travail, un *token* est donc une unité linguistique composée d'une séquence de caractères qui peut être isolée par des blancs (espaces, retours à la ligne), les caractères alphanumériques et les signes de ponctuation étant considérés comme des tokens distincts. Par exemple, l'extrait de production normalisée suivant contient 28 tokens :

[...] Il était une fois une sorcière et un chat, la sorcière qui se nomme Pustula, elle détestait ce chat qui lui rendait la vie impossible... [...] (élève 98 - CM2)

Il - était - une - fois - une - sorcière - et - un - chat - , - la - sorcière - qui - se - nomme - Pustula - , - elle - détestait - ce - chat - qui - lui - rendait - la - vie - impossible - ... : 28 tokens

Nous verrons dans la suite que la tokenisation diffère selon les outils TAL utilisés. Cette différence s'explique par de nombreuses difficultés d'identification des tokens et doit être prise en compte.

B. Sagot et P. Boullier (2008), par exemple, mettent en avant les difficultés liées à certains caractères typographiques tels que le tiret et l'apostrophe. En effet, ces signes peuvent être tour à tour inclus dans un token (exemple *aujourd'hui*) ou séparateurs de deux tokens (exemple *m'a*). Dans ce dernier cas, nous considérons qu'il s'agit de deux tokens distincts. Nous verrons par la suite que cette ambigüité peut générer de nombreuses erreurs dans les systèmes de traitement automatique.

Une autre différence notable entre les outils de tokenisation porte sur la prise en compte des unités polylexicales. Par exemple, le groupement *pomme de terre* peut être considéré comme un seul token, c'est-à-dire une unité polylexicale, ou comme trois tokens, c'est-à-dire trois formes. Dans le cadre du projet *Scoledit*, notre attention porte davantage sur la segmentation en mots que sur la dimension sémantique, il y a donc peu d'enjeu à reconnaitre les unités polylexicales. Le groupement *pomme de terre* y est donc reconnu comme trois tokens distincts.

Dans la suite de notre propos, le terme *token* est réservé au contexte du TAL et des outils de traitement. Pour la description et l'analyse linguistique, nous utilisons les termes *segment* et *forme (graphique)*.

Le *segment* représente une unité graphique, c'est-à-dire une suite de caractères séparés par des blancs ou par des signes de ponctuation, tandis que le terme *forme graphique*, raccourci en *forme*, renvoie aux formes attestées en français, y compris les formes fléchies. Les *segments* correspondent donc à ce qui est attesté en corpus. Les *formes graphiques* correspondent à l'attendu. Dans l'exemple « pleur » (*pleure*, 78 - CP), « pler » est désigné comme un segment et *pleure* comme une forme.

Selon les cas, les découpages en segments et en formes coïncident ou non. Lorsque segments et formes correspondent, on admet que la segmentation est normée. Lorsque plusieurs segments transcrits correspondent à une unique forme normée, on parle de phénomène d'*hypersegmentation* (exemple : « en semble » pour *ensemble*). À l'inverse, lorsqu'un même segment transcrit correspond à plusieurs formes normées, on parle d'*hyposegmentation* (exemple : « ilfut » pour *il fut*).

1.2. Principe de l'approche par comparaison appliquée aux formes

L'architecture générale de l'approche par comparaison a été présentée précédemment (cf. Chapitre 4 - 3), il s'agit de comparer des éléments transcrits et des éléments normalisés pour en extraire des éléments d'analyse. Cette approche nécessite une phase d'alignement entre les éléments transcrits et les éléments normalisés que l'on souhaite comparer.

Le niveau lexical étant un niveau d'analyse prépondérant en linguistique de corpus, et particulièrement en linguistique de corpus outillée (utilisant des outils de traitement automatique des langues), c'est le premier niveau que nous choisissons pour appliquer l'approche par comparaison.

Pour ce faire, différentes étapes, illustrées dans la figure 43, sont nécessaires.

1. La phase de prétraitement :
    a. **L'étiquetage morphosyntaxique** des normalisations (ou POS tagging) : pour chaque forme du corpus, la catégorie grammaticale et le lemme sont calculés à l'aide d'un outil d'étiquetage morphosyntaxique.
    b. **Le calcul des représentations phonologiques** des productions (abrégé en phonologisation) : les transcriptions et normalisations de chaque production sont converties en représentations phonologiques.

2. **L'alignement** des segments transcrits avec les formes normalisées : au cours de cette étape les segments transcrits sont mis en correspondance avec les formes normalisées à l'aide d'un algorithme d'alignement fondé sur des indices graphiques et phonologiques (cf. Chapitre 8).
3. **Les comparaisons et analyses** : après alignement des segments et des formes, de nombreuses applications sont possibles, comme la production d'un index des formes du corpus et des différents segments, correctement orthographiés ou non, associés à ces formes. Il est également possible de comparer segments transcrits et formes normées pour en dégager les différentes erreurs et les réussites orthographiques ou morphologiques. Au cours de cette étape, d'autres outils de comparaison peuvent être développés pour analyser d'autres niveaux linguistiques, comme les désinences verbales ou les graphèmes.

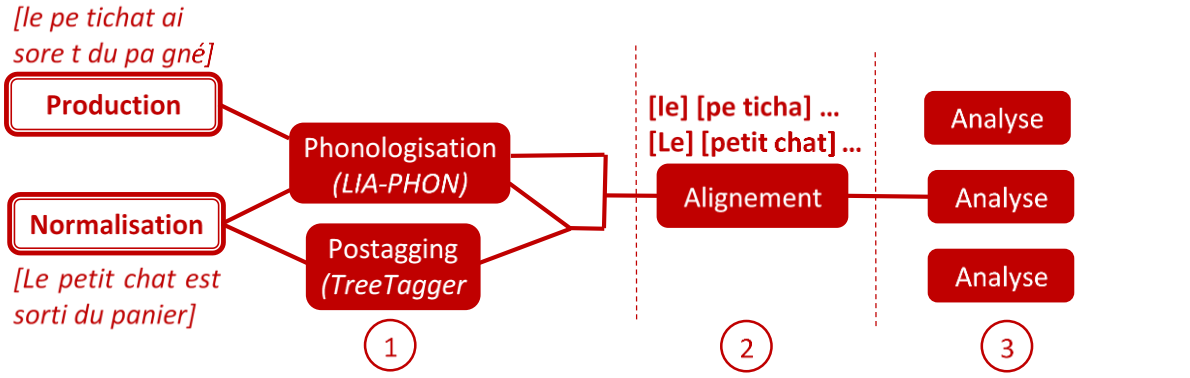


Figure 43 : Approche par comparaison appliquée à l'identification des formes du corpus

L'approche que nous utilisons nécessite plusieurs outils de traitement automatique des langues : (1) un analyseur morphosyntaxique (ou POS tagger), (2) un outil permettant de convertir des formes graphiques en représentations phonologiques et (3) un aligneur des segments et des formes.

(1) L'outil *TreeTagger*[77] est utilisé pour l'étiquetage morphosyntaxique, il présente l'avantage d'être gratuit et d'être applicable à un nombre de langues relativement grand au regard des autres outils, ce qui permettrait une adaptation ultérieure de notre système. Après

[77] Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees, Intl. In *Conference on New Methods in Language Processing. Manchester*. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ [consulté le 23/09/2019]

tokenisation, cet outil permet de fournir, pour chaque token, le lemme et la catégorie grammaticale associée à celui-ci.

(2) Pour réaliser le calcul des représentations phonologiques, l'outil *LIA-PHON*[78] est utilisé. Outre sa gratuité, l'avantage de cet outil est son adaptabilité aux corpus très éloignés de la norme, comme le sont les corpus scolaires. En effet, la plupart des outils de phonologisation se basent sur des listes de représentations phonologiques des formes fléchies du français. Ces outils échouent donc à produire une représentations phonologique d'un mot inconnu, c'est-à-dire absent de ces listes, comme le sont souvent les mots comportant des erreurs orthographiques. L'outil *LIA-PHON* en revanche se base sur un système de règles de phonologisation du français qui détermine la valeur phonologique de chaque caractère ou groupe de caractères selon le contexte dans lequel il est situé. Ces règles explicitent, par exemple, la valeur phonologique de la lettre *c* selon que celle-ci soit placée devant un autre *c* (/k/ ou muet la plupart du temps), devant les voyelles *e*, *é*, *è*, *ê*, *i*, *î* ou *y* (/s/ en général) ou devant un autre caractère (/k/ le plus souvent).

(3) En revanche, il n'existe pas d'outil pertinent et accessible librement à l'heure actuelle pour réaliser l'alignement nécessaire entre les transcriptions et les normalisations. La tâche d'élaboration de cet algorithme d'alignement fait l'objet de la suite de ce chapitre.

1.3. Algorithme d'alignement, quelques explications

Selon R. Beaufort (2008, p. 156) un alignement est « l'établissement de la correspondance maximale entre deux séquences, en déterminant leurs sous-séquences identiques ou similaires, ainsi que leurs sous-séquences propres. » Dans certains cas, par exemple dans le cas de la commande *diff* qui permet de comparer deux fichiers textuels, on parle aussi de synchronisation. Un algorithme d'alignement cherche donc à faire correspondre deux séquences de caractères, de formes ou de segments, en maximisant le nombre ou la longueur des séquences identiques.

Prenons un exemple qui pourrait être extrait d'un corpus de SMS : l'emploi de la séquence de caractères « a tt », signifiant « à toute ». La figure 44 et la figure 45 illustrent différents alignements possibles. On attribue la valeur 1 aux caractères qui correspondent, 0 aux autres. Dans cet exemple, l'alignement le plus simple revient à aligner le caractère *n* d'une séquence avec le caractère *n* de l'autre séquence. Mais, ce n'est pas l'alignement le plus pertinent et il importe d'aller rechercher une correspondance maximum entre les caractères (Figure 45). Pour

[78] Béchet, F. (2001). *LIA-PHON*: Un système complet de phonétisation de textes. *TAL. Traitement automatique des langues*, 42(1), 47-67. http://pageperso.lif.univ-mrs.fr/~frederic.bechet/download.html [consulté le 23/09/2019]

ce faire, différentes méthodes ont été développées. Nous proposons une revue de ces méthodes dans la suite de cette section.

	1	2	3	4	5	6
Segment produit :	a	t	t			
Correspondant standard :	à	t	o	u	t	e
Valeur de la comparaison :	1[79]	1	0	0	0	0

Figure 44 : Alignement simple

	1	2	3	4	5	6
Segment produit :	a	t			t	
Correspondant standard :	à	t	o	u	t	e
Valeur de la comparaison :	1	1	0	0	1	0

Figure 45 : Alignement maximisé

1.4. Domaines d'application des algorithmes d'alignement

Les tâches d'alignement sont utilisées pour de nombreuses applications, les principales applications étant la traduction et la construction de lexiques multilingues. L'alignement a alors pour rôle de faire correspondre deux ou plusieurs langues. Les algorithmes utilisés pour ce type d'applications sont très différents selon qu'ils concernent :

- des langues non apparentées ou apparentées de manière lointaine, comme l'anglais et l'hindi (Singh et al., 2010), le vietnamien et le français (Bigi & Le, 2008), etc.
- des langues dites apparentées comme le portugais et l'espagnol (Mann & Yarowsky, 2001), les différents dialectes alsaciens (Bernhard & Steiblé, 2015), ou encore les différents dialectes germanophones de Suisse (Scherrer, 2007a).

On retrouve également un emploi important des algorithmes d'alignement dans des tâches de normalisation. Ces tâches peuvent être de natures très diverses :

- La normalisation d'un standard orthographique vers un autre standard orthographique. On citera pour exemple les travaux de J. Porta, J.-L.Sancho et J. Gómez (2013) dont l'objectif est de faire le lien entre la norme orthographique des années 1930 et la norme orthographique des années 1970, ou encore les travaux de M. Kestemont, W.Daelemans et G. De Pauw (2010) qui posent le problème de la lemmatisation des textes en néerlandais moyen à partir de manuscrits antérieurs à 1300.

[79] On considèrera ici que « a » est un équivalent acceptable de « à »

- La normalisation de SMS en un écrit plus standard, analysable à l'aide d'outils automatiques (Beaufort, 2008).
- La normalisation de textes produits par des apprenants lors d'exercices de dictées ou de production de texte en vue de leur apporter un feedback sur leurs productions (Véronis, 1988 ; Berkling, 2001).

Les algorithmes d'alignement sont également très présents en traitement automatique de la parole, notamment dans les applications de synthèse vocale (Ristad & Yianilos, 1998 ; Beaufort, 2008) qui nécessitent un alignement du texte et de la représentation graphique des sons (alignement graphème – phonème le plus souvent).

Enfin, il existe d'autres domaines dans lesquels les algorithmes d'alignement sont utilisés dans différents travaux de recherche, à l'exemple de certains systèmes de lemmatisation ou de la construction de ressources bilingues, telles que des lexiques. Dans le premier cas, il s'agit de faire correspondre des formes avec un lemme ; dans le second, il s'agit d'associer des formes de langues différentes.

Il est intéressant d'observer les domaines pour lesquels ont été développés ces algorithmes d'alignement, mais ce qui va nous intéresser plus encore, ce sont les méthodes d'alignement utilisées.

2. État de l'art des méthodes d'alignement

De très nombreuses méthodes d'alignement ont été développées dans le domaine de la traduction. Certaines de ces méthodes, développées pour des langues non apparentées, utilisent des algorithmes complexes basés sur des indices syntaxiques ou sémantiques qui nécessitent de nombreuses données préalables, telles que des lexiques, des modèles syntaxiques, des bases de données sémantiques, etc. Or, les études sur corpus scolaires de taille suffisamment conséquentes n'en étant qu'à leur début, ces données n'existent pas encore. De plus, les normalisations des écrits scolaires étant produites à partir de ces écrits, les transcriptions et les normalisations présentent des différences non comparables à celles qu'on peut observer entre deux langues différentes. S'appuyer sur des méthodes développées pour la traduction automatique de langues non apparentées représenterait donc une complexité peu pertinente et non adaptée. Nous excluons donc ces méthodes de notre état de l'art.

La plupart des méthodes d'alignement développées en traduction concernent souvent des langues apparentées de manière plus ou moins lointaines, comme l'anglais et le français. Elles sont souvent surfaciques et se basent sur deux types d'indices : l'observation des longueurs de phrases et la recherche de ressemblances lexicales (Lamraoui & Langlais, 2013). Le plus

souvent, ces méthodes sont combinées à des modèles statistiques permettant d'induire les variations de longueurs de phrase (Brown, Della-Pietra, Della-Pietra, & Mercer, 1993 ; Gale & Church, 1991, cités par Lamraoui & Langlais, 2013) ou les correspondances lexicales (Church, 1992, cité par Kraif, 2001 ; Simard, Foster et Isabelle, 1992, cités par Lamraoui & Langlais, 2013). Les modèles les plus fréquemment utilisés sont les modèles IBM (décrits par Brown *et al.*, 1993) et implémentés dans l'outil *Giza++* (Och & Ney, 2000). Comme le pointe O. Kraif (2001, p.227), « le grand mérite de ces premières recherches est de montrer qu'il est possible d'aligner sans passer par le sens, en se basant sur des propriétés purement formelles. »

De même, le point commun de la plupart des méthodes utilisées dans les champs d'applications cités précédemment (traduction de langues apparentées, normalisation, synthèse vocale, etc.) est de se fonder non pas sur des indices sémantiques ou syntaxiques, comme c'est souvent le cas dans le champ des applications plurilingues, mais bien plus sur des indices graphiques ou phonologiques. Elles reposent donc souvent sur des mesures de similarité (graphiques ou phonologiques) et sur des règles de transformation. Elles peuvent parfois être facilitées par l'usage de pivots ou de cognats, c'est-à-dire des paires de mots facilement identifiables car très proches ou parce qu'elles sont connues au préalable. Ces méthodes sont rendues possibles par la grande proximité entre les séquences à aligner.

Ainsi, nous commençons par faire la même hypothèse que Y. Scherrer (2007, p. 55), « phonetic (or rather graphemic, as we use written data) similarity measures are the most appropriate because they require less sophisticated training data than semantic or syntactic similarity models[80] », et nous ajoutons que selon les modèles, il peut même n'être nécessaire d'entrainer aucune donnée.

2.1. Distance d'édition ou distance de Levenshtein

Un certain nombre de méthodes d'alignement s'appuient sur des mesures de similarité et particulièrement l'une d'elle, appelée distance d'édition, ou encore distance de V. Levenshtein (1966). Avant de faire une revue des méthodes d'alignement, une explication de cette distance d'édition semble nécessaire.

2.1.1. Principe général

Cette mesure permet de calculer la distance minimale entre deux chaines de caractères à partir de trois opérations d'édition : l'insertion, l'omission et la substitution, dans le cas de la distance

[80] Les mesures de similarité phonétiques (ou plutôt graphémiques, comme nous utilisons des données écrites) sont les plus appropriées parce qu'elles requièrent des données d'entrainement moins sophistiquées que les modèles de similarité sémantiques ou syntaxiques.

de Levenshtein. La distance de Damereau-Levenshtein (Damerau, 1964) prend également en compte l'opération d'inversion.

Prenons un exemple issu du corpus *Scoledit*, soit le segment transcrit *pran* et la forme normée *prend* (1189, CP). Considérons, le calcul de la distance de Levenshtein entre ces deux formes. Chaque opération d'identité vaut 0 (c'est-à-dire lorsque les deux caractères considérés sont identiques). Les opérations de substitution, de suppression ou d'ajout valent 1. La distance d'édition est la somme de la valeur de ces opérations. Classiquement, ce calcul est représenté dans une matrice (Figure 46). Dans cette matrice, chaque case vaut la valeur de comparaison (0 ou 1), additionnée du chemin le plus court, contenu dans la cellule, à gauche, au-dessus ou dans la diagonale haute gauche.

		1	2	3	4
		p	r	a	n
1	p	0	1	2	3
2	r	1	0	1	2
3	e	2	1	1	2
4	n	3	2	2	1
5	d	4	3	3	2

Figure 46 : Matrice d'édition entre les formes pran et prend

Dans cet exemple, il faut lire :

- en [1,1] « le caractère p de *pran* est identique au caractère p de *prend*, la comparaison vaut donc 0 ;
- en [1,2] « le caractère p de *pran* est différent du caractère r de *prend*, la comparaison vaut donc 1 ; on additionne 1 au 0 en [1,1] ;
- en [2,1] « le caractère r de *pran* est différent du caractère p de *prend*, la comparaison vaut donc 1 ; on additionne 1 au 0 en [1,1] ;
- en [2,2] « le caractère r de *pran* est identique au caractère r de *prend*, la comparaison vaut donc 0 ; la distance minimale entre [1,1], [1,2] et [2,1] est [1,1] puisqu'elle vaut 0, on additionne donc 0 à 0 en [1,1] ;
- etc.
- en 4,5 « le caractères *n* de *pran* est différent du caractère *p* de *prend*, la comparaison vaut donc 1 ; la distance minimale entre [3,4], [4,4] et [3,5] est [4,4] puisqu'elle vaut 1, on additionne donc 1 à 1, soit 2, en [4,5] ;

On conserve le chemin le plus court (surligné dans l'exemple), soit une distance de 2 incluant une substitution de *a* par *e* en [3,3] et une insertion de *d* en [5,4], soit l'alignement : p – p ; r – r ; a – e ; n – n ; d – d ; – s.

En 1974, R. A. Wagner et M. J. Fischer propose un algorithme qui reprend et adapte l'algorithme d'A. J. Viterbi (1967) afin de calculer cette distance d'édition à l'aide de la programmation dynamique.

2.1.2. Distance d'édition avec poids

La distance d'édition de Levenshtein se base uniquement sur des indices graphiques et de nombreux travaux d'alignement se base sur cette mesure de similarité (Véronis, 1988 ; Mann & Yarowsky, 2001 ; Scherrer, 2007 ; Dasigi & Diab, 2011). En 2007, par exemple, D. Lyras, K. Sgarbas et N. Fakotakis, utilisent la distance de Levenshtein pour une tâche de lemmatisation en grec moderne et en anglais. Dans cet exemple, calculer la distance d'édition entre une forme et l'ensemble des lemmes contenus dans une liste de 30 000 lemmes doit permettre de retrouver le lemme associé.

Mais, cette mesure est aveugle de toutes caractéristiques linguistiques. De multiples variations ont pu être proposées pour correspondre au mieux aux données linguistiques concernées. Par exemple, dans le cadre de travaux sur des orthographes diachroniques, il peut apparaitre certaines substitutions récurrentes d'une lettre ou d'un groupe de lettres par une autre lettre ou groupe de lettres (Pilz, Ernst-Gerlach, Kempken, Rayson, & Archer, 2008). De même, dans le cadre de travaux sur la production de dictées, le remplacement de lettres ou groupes de lettres de même son est fréquent, de même que la suppression de lettres non prononcées (Véronis, 1988).

Pour prendre en compte ces spécificités, la plupart des travaux utilisant des distances d'édition s'appliquent à faire varier la valeur des couts pour donner une importance différenciée aux différents types d'opérations en fonction d'indices de différentes natures.

Dans le calcul de distance original, chaque opération d'édition différente de l'identité vaut 1. De nombreux auteurs proposent de faire varier ces couts en donnant différentes valeurs comprises entre 0 et 1 (Mann & Yarowsky, 2001, par exemple), voire même en attribuant un cout 0, similaire à une opération d'identité à des opérations qui n'en sont pas (Véronis, 1988).

À l'instar de G. Kondrak et T. Sherif (2006), on distingue principalement :

- les **variations de couts établies manuellement**, basées sur des intuitions linguistiques, des mesures statistiques ou des règles de transformation ;

- les **variations de couts établies par apprentissage automatique**, généralement construites à partir de données d'entrainement.

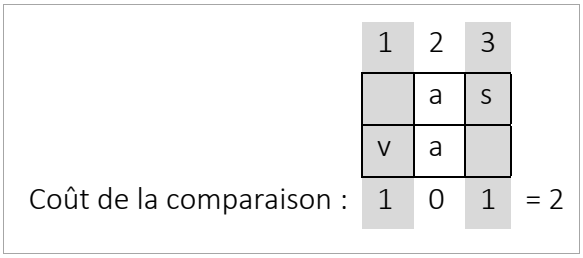
2.1.3. Unités de mesure

Dans le cas le plus fréquent, la distance d'édition de Levenshtein permet de calculer la distance entre deux séquences de caractères en prenant en compte le nombre de caractères qui diffèrent entre ces deux séquences. L'unité la plus souvent utilisée dans ce cas-là est l'unigramme, c'est-à-dire un caractère. Il existe des variantes (Inkpen et al., 2005) de cette distance qui prennent pour unité de mesure des n-grammes (plusieurs caractères), souvent des bigrammes (deux caractères) ou des trigrammes (trois caractères).

2.1.4. Distance d'édition normalisée

Proposée par A.Marzal et E. Vidal (1993), la distance d'édition normalisée (Normalized Edit Distance) est une variante de la distance de Levenshtein. Dans cette variante, la somme des couts d'édition est divisée par la longueur du plus long segment (Marzal & Vidal, 1993 ; Inkpen *et al.*, 2005) ou par la longueur du plus long alignement (Heeringa et al., 2006).

Dans le cas d'un calcul de distance classique, la distance entre *as* et *va* est moins élevée qu'entre *mangeait* et *manger* (Figure 47). Entre autres apports, la distance d'édition normalisée permet de neutraliser les couts élevés liés à la comparaison de segments longs.



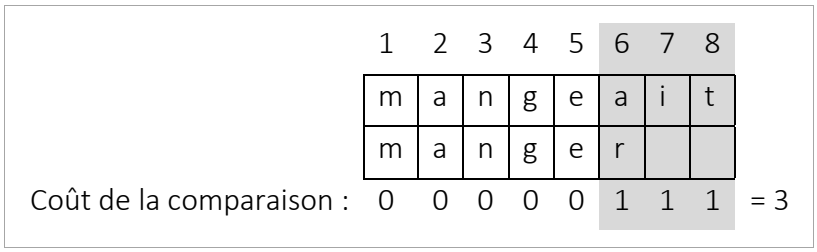


Figure 47 : Calcul de distance entre deux séquences de caractères

2.2. Critères et indices d'alignement

L'aspect qui nous intéresse le plus dans la revue de ces différents travaux est la nature des indices utilisés pour réaliser ces alignements. Nous avons précédemment écarté les indices de nature syntaxique et sémantique en raison des données importantes qu'ils nécessitent. Nous

verrons que ces travaux peuvent également utiliser des indices graphiques, phonologiques ou encore stochastiques.

2.2.1. Méthodes basées sur des indices graphiques

2.2.1.1. Mesures de distance et de similarité

Les mesures de similarité permettent de calculer la proximité entre deux séquences de caractères. À l'inverse, les mesures de distance calculent l'éloignement entre deux séquences de caractères. Il existe un grand nombre de mesures différentes, développées dans de multiples domaines comme le traitement automatique des langues, les télécommunications, le séquençage du génome, etc.

En traitement automatique des langues, la mesure la plus utilisée est la distance d'édition (Levenshtein, 1966), décrite précédemment (2.1.1). Dans la plupart des travaux présentés dans la suite de ce chapitre, les mesures de distance représentent un point de départ à partir duquel sont développés de nouveaux algorithme par variation de différents paramètres.

2.2.1.2. N-grammes

Comme nous l'avons vu dans l'explication de la distance d'édition (2.1.3.), les mesures de distances peuvent également se baser sur des n-grammes (plusieurs caractères ensemble), au lieu des caractères seuls (unigrammes). Ainsi, D. Inkpen *et al*. (2005), dans leur étude pour identifier les paires de cognats (mots apparentés) et de faux-amis (mots semblables mais non apparentés), reprennent la mesure DICE (Adamson & Boreham, 1974), qui mesure le nombre de bigrammes communs entre deux segments par rapport au nombre total de bigrammes (deux caractères). Ils réalisent également la même mesure à partir de trigrammes (trois caractères). De même, S. Kempken,W. Luther et T. Pilz (2006) dans leurs travaux sur les variations orthographiques historiques ont testé différentes méthodes, dont une comparaison de bigrammes, pour générer une liste de variantes orthographiques.

Dans leur étude sur les cognats et les faux amis, D. Inkpen *et al*. (2005) ont également essayé de combiner l'usage de bigrammes et des calculs de distance de Levenshtein, ce qui accroit leurs résultats. Dans une étude sur la modélisation de distance entre dialectes, W. Heeringa *et al*. (2006) appliquent le même procédé en l'appliquant aux représentations phonologiques de formes des dialectes étudiés. Leur objectif est d'utiliser une méthode sensible au contexte phonologique, modélisé par les phonèmes adjacents.

2.2.1.3. Séquences communes

Ces méthodes ne sont pas les seules à s'appuyer sur des indices graphiques. En effet, certaines méthodes se basent sur la recherche des plus longues séquences communes (séquences de

lettres identiques et dans le même ordre, pas nécessairement consécutives) pour apparier des formes ou des cognats (Debili & Sammouda, 1992 ; Melamed, 1995 ; Kraif, 2001).

D'autres méthodes (Simard, Foster, & Isabelle, 1993 ; Inkpen *et al*., 2005) se concentrent plus spécifiquement sur les préfixes ou les premières lettres de chaque forme et identifient les formes apparentées en se basant sur la comparaison de ces segments.

2.2.1.4. Règles orthographiques

Guidée par la volonté d'accompagner l'apprentissage de l'orthographe du français, C. Santiago Oriola (1998) a mis en place le système *DICTOR*. Afin d'identifier les erreurs commises par les apprenants, ce système intègre un algorithme appelé *Veritext*[81] qui permet l'alignement du texte dicté avec la production de l'apprenant. Cet algorithme est basé sur un ensemble de règles de réécriture et s'appuie sur un modèle stochastique, c'est-à-dire qui utilise des probabilités. Il se décompose en deux parties : une première partie qui permet la réécriture des erreurs orthographiques et une deuxième partie qui permet de réécrire les erreurs typographiques (erreurs dues à l'usage d'un clavier). Un des points intéressants de ce travail est qu'il ne propose pas un alignement caractère par caractère mais un alignement en groupements de caractères, appelés *groupes posant des problèmes orthographiques*, correspondant approximativement à des graphèmes.

En s'inspirant des travaux de M. Mohri (2003), R. Beaufort (Beaufort, Roekhaut, & Fairon, 2008 ; Beaufort, 2008) propose une alternative à l'implémentation de la distance d'édition au moyen de la programmation dynamique, telle que proposée par R. A. Wagner et M. J. Fisher (1974), en s'appuyant sur des transducteurs, ou transducteurs à états finis. Un transducteur est un automate qui permet de passer d'un état A à un état B grâce à des transitions qui définissent les réécritures possibles, c'est-à-dire les opérations d'édition. Généralement, les transitions sont appliquées à l'échelle du caractère et plusieurs transitions sont applicables à chaque caractère. La distance d'édition minimale correspond alors au nombre minimal de transitions pour obtenir l'état final à partir de l'état initial. Dans leur application à la normalisation de SMS, R. Beaufort *et al*. (2008) définissent un certain nombre de règles de réécriture afin de limiter le nombre de transitions possibles. L'avantage de leur méthode est qu'elle permet de prendre en compte les phénomènes d'agglutination (ou hyposegmentation), en prenant pour unité le caractère et en alignant du même coup l'ensemble du SMS.

[81] Pécatte, J. M. (1992). *Tolérance aux fautes dans les interfaces homme-machine. Traitement des chaînes phonétiques, des chaînes orthographiques et des structures syntaxiques.* (Thèse de doctorat).

Parallèlement, Y. Scherrer (2007) propose également un transducteur à base de règles, dans l'objectif d'induire un lexique bilingue entre un dialecte suisse allemand et l'allemand standard. Le transducteur utilisé s'appuie à la fois sur un ensemble de 50 règles de transposition contextuelles et sur les opérations d'édition classiques établies pour la distance de Levenshtein, non contextuelles. La prévalence est donnée aux règles de transposition grâce à un système de poids attribué à chaque transition possible (1 pour les règles, 2 pour l'identité, 3 pour les autres opérations d'édition).

En 2011, M. Hulden, I. Alegria, I. Etxeberria et M. Maritxalar (2011) s'appuient sur un outil élaboré pour permettre la réécriture de textes en portugais selon différentes conventions orthographiques appelé *lexdiff* (Almeida et al., 2010) pour apprendre les transformations morphologiques d'une forme dialectale vers la langue basque standard. Pour cela, les règles de transformation les plus fréquentes, générées à l'aide de l'outil *lexdiff* sont implémentées dans un transducteur.

2.2.1.5. Génération de candidats par règles

De nombreux travaux ne reposent pas sur des mesures de proximité mais s'appuient sur la génération de variantes potentielles et la comparaison de ces variantes potentielles à des candidats possibles. Selon les tâches, ces candidats possibles peuvent être les formes d'un lexique (tâche de conception d'un lexique bilingue, tâche de correction orthographique à partir d'un dictionnaire, etc.) ou les formes contenues dans la version normalisée du texte travaillé (tâche d'alignement avec une production normalisée, par exemple). C'est ce que proposent A. Ernst-Gerlach et N. Fuhr (2006) et T. Pilz et ses collègues (2008) dans leurs travaux sur les textes médiévaux en allemand et en anglais. À partir de la comparaison de variantes anciennes et de variantes contemporaines, ils génèrent des règles de transposition spécifiques (le caractère transposé, son contexte gauche, son contexte droit), puis en déduisent des règles plus générales, pour finalement conserver, à l'aide d'une approximation de l'algorithme *Prism* (Cendrowska, 1987), les règles les plus utiles.

2.2.2. Méthodes basées sur des indices phonologiques

2.2.2.1. Distance d'édition avec variation de poids

En 1988, J. Véronis, travaillant à une tâche de correction de dictées en français, fait le constat que le calcul de la distance d'édition permet de corriger certaines erreurs d'orthographe, comme les erreurs typographiques (dues à l'action d'écrire au clavier), mais échoue à corriger certaines autres, comme certaines erreurs phonographiques. Par exemple, la forme *ippeauttainnuze* est graphiquement très éloignée de la forme normée *hypoténuse*. Pourtant, il y a peu de doute pour un locuteur du français qu'il s'agit de la forme voulue. J. Véronis propose

une adaptation de la distance d'édition qui prend en compte la substitution d'un graphème par un autre de même valeur phonique, par exemple *eau*, *au* et *o*. Mentionnons que cette variation nécessite de modifier l'algorithme afin qu'il puisse prendre en compte plusieurs caractères successifs.

En 2001, G. S. Mann et D. Yarowsky développent une méthode de traduction qui se base sur la recherche de cognats, c'est-à-dire de formes lexicales apparentées, dans une langue proche ou plusieurs langues proches. Pour une tâche de traduction entre anglais et portugais, par exemple, des cognats sont recherchés dans des lexiques du français et de l'espagnol. Pour ce faire, les auteurs utilisent plusieurs mesures de similarité, dont plusieurs variations de la distance d'édition de Levenshtein. Ils proposent ainsi de diminuer le cout d'une substitution d'une voyelle par une autre. De la même façon, W. Heeringa *et al*. (2006) proposent d'apposer un cout infini aux substitution d'une voyelle par une consonne, et inversement, afin d'exclure ces possibilités.

Dasigi et Diab (2011) dans leurs travaux sur la recherche de variantes orthographiques dans les dialectes arabes utilisent, outre la mesure directe de la distance d'édition de Levenshtein, une variante appelée *Biased Edit Distance Metric* dans laquelle les substitutions entre deux lettres qui peuvent avoir le même son valent 0.

2.2.2.2. Méthodes à partir de clés phonétiques

Afin de comparer phonétiquement des séquences de caractères, différents algorithmes existent tels que l'algorithme *Soundex* [82] , l'algorithme *Phonix* [83] et l'algorithme *Double Metaphone* [84] qui permettent de convertir une séquence de caractères en clé phonétique. *Soundex*, par exemple, réalise cette conversion en conservant la première lettre de la séquence de caractères et en substituant les lettres suivantes par des chiffres. Un même chiffre est donné aux lettres de valeur phonique proche ou ayant des traits articulatoires communs (ainsi les lettres *b*, *p*, *f* et *v* sont remplacées par un même chiffre).

Les clés associées à chaque forme peuvent alors être comparées directement. Différents travaux de recherche (Zobel & Dart, 1996 ; Inkpen *et al*., 2005 ; Kempken *et al*., 2006) proposent également de combiner de telles mesures avec des algorithmes de distance d'édition. Alors qu'ils devaient résoudre un problème dans le champ de la recherche d'information, nécessitant parfois de retrouver des noms identiques mais orthographiés différemment par différents

[82] Odell, M., & Russell, R. (1918). The soundex coding system. *US Patents*, *1261167*.
[83] Gadd, T. N. (1990). PHONIX: The algorithm. *Program*, 24(4), 363-366.
[84] Phillips, L. (2000). The double metaphone search algorithm.

employés de centre d'appel, J. Zobel et P. Dart (1996) ont ainsi développé un algorithme appelé *Editex* qui procède à un calcul de distance d'édition sur des clés phonétiques similaires à celles calculées à partir de Soundex et de Phonix. De façon similaire, D. Inkpen *et al*. (2005) ont développé un algorithme qui retourne la distance d'édition entre des clés phonétiques calculées par *Soundex*.

Sans se baser sur une distance d'édition, dans leurs travaux sur les variantes orthographiques de l'Alsacien, D. Bernhard et L. Steiblé (2015) considèrent que différentes formes graphiques correspondent à la même unité lexicale si elles remplissent trois critères : avoir la même traduction en français, appartenir à la même partie du discours et avoir une clé phonétique en commun. Ces clés phonétiques sont calculées à l'aide de l'algorithme *Double Metaphone*.

2.2.2.3. Méthodes par conversion phonétique

Plutôt que de s'appuyer sur les lettres pouvant avoir la même prononciation, certains algorithmes s'appuient directement sur des transcriptions phonétiques des segments ou séquences de segments à comparer.

Ainsi, la méthode développée par J. Zobel et P. Dart (1996), appelée *méthode phonométrique*, est constituée de deux étapes. Une première étape permet de convertir des séquences de lettres en séquences de phonèmes. La distance d'édition entre ces dernières est calculée lors de la deuxième étape. Des poids différenciés sont attribués aux phonèmes comparés selon le nombre de traits phonétiques distincts et leur nature.

De façon assez similaire, à partir de séquences de phonèmes, l'algorithme *Aline* (Kondrak, 2000) propose un calcul de distance phonétique basé sur la comparaison de 12 traits articulatoires et phonétiques accompagnés de poids différents.

Adoptant une méthode tout à fait différente pour faire correspondre formes anciennes et formes nouvelles en espagnol, J. Porta et ses collègues (2013), tout en s'appuyant sur une conversion phonétique des formes, utilisent une combinaison de transducteurs. Le processus adopté est le suivant : les formes étudiées sont d'abord converties en représentations phonétiques à l'aide de règles de transposition ; puis ces dernières sont retranscrites en graphèmes ; enfin un ensemble de règles de variation graphique permettent de sélectionner les formes de surface optimales.

2.2.2.4. Méthodes par apprentissage automatique

L'algorithme *Cordi* (Kondrak, 2002) produit un modèle de traduction en calculant par apprentissage automatique les correspondances phonétiques récurrentes dans une liste

bilingue de mots. Ces correspondances sont ensuite utilisées pour calculer un score de similarité entre deux mots.

M. Hulden *et al*. (2011) produisent un algorithme similaire en utilisant la programmation logique inductive[85] (Muggleton & De Raedt, 1994). Pour ce faire, les mots sont alignés à l'aide d'une distance d'édition simple puis des règles de réécritures phonétiques sont apprises automatiquement. La recherche de contre-exemples permet ensuite d'identifier le contexte minimal (généralement les lettres qui précèdent ou qui suivent) autorisant l'application de la règle de réécriture.

2.2.3. Méthodes basées sur des indices stochastiques

L. Bahl et F. Jelinek (1975) sont les premiers à proposer une interprétation stochastique de la distance d'édition. Plutôt que de modifier l'algorithme de programmation dynamique développé par R. A. Wagner et M. J. Fischer (1974), ils proposent une nouvelle méthode d'implémentation de la distance d'édition, les transducteurs stochastiques à états finis (cf. 41.3). Par ailleurs, P. Hall et G. Dowling (1980) développent eux aussi un calcul de distance d'édition basé sur des calculs stochastiques en proposant une variante proche de l'algorithme de programmation dynamique.

En 1998, E. Ristad et P. Yianilos proposent d'introduire des méthodes d'apprentissage automatique dans les calculs de distance d'édition. Pour ce faire, ils reprennent à L. Bahl et F. Jelinek (1975) l'idée de s'appuyer sur un transducteur stochastique. À partir d'un corpus d'exemples de paires de mots correctes, ils entrainent ce transducteur qui estime de manière probabiliste le poids d'une opération d'édition (omission, insertion, substitution). Cet entrainement est réalisé de manière non-supervisée, c'est-à-dire sans données linguistiques étiquetées au préalable, à l'aide de l'algorithme de maximisation de l'espérance (EM). Leur méthode est appliquée à une tâche d'apprentissage de la prononciation de mots en conversation.

Quelques années plus tard, G. S. Mann et D. Yarowsky (2001) se penchent sur des lexiques de traduction d'une langue source vers une langue cible, en s'appuyant sur des lexiques en langues proches. Pour ce faire, ils utilisent un algorithme de distance de Levenshtein dont ils font varier le poids des opérations d'édition en reprenant le transducteur stochastique d'E. Ristad et

[85] La programmation logique inductive est un type d'apprentissage automatique qui s'appuie sur des exemples positifs et des exemples négatifs pour générer un programme hypothétique, un ensemble de règles dans notre cas.

P. Yianilos (1998). L'apprentissage des poids est réalisé de deux façons : à partir d'un couple de langues proches ou à partir de l'ensemble de la famille de langues.

La méthode proposée par E. Ristad et P. Yianilos (1998) a été adaptée à de nombreux travaux d'alignement de langues proches (Kondrak & Sherif, 2006 ; Kempken *et al*., 2006 ; Scherrer, 2007) et dans le domaine de la parole (Jansche, 2003).

L'algorithme de maximisation de l'espérance est utilisé à plusieurs reprises, par exemple dans les travaux de E. Ristad et P. Yianilos (1998), mais aussi dans les travaux de R. I. Damper, Y. Marchand, J.-D. S. Marsters et A. I. Bazin (1997) portant sur l'alignement de lettres et de phonèmes à partir d'un algorithme de programmation dynamique. Mais, d'autres modèles stochastiques sont utilisés pour produire des mesures de similarité, comme les réseaux bayésiens dynamiques (Filali & Bilmes, 2005) et les *Pair Hidden Markov Model*, une variante des modèles de Markov cachés (Mann & Yarowsky, 2001 ; Mackay & Kondrak, 2005). Selon les travaux, ces mesures stochastiques ont été implémentées à l'aide de transducteurs à états finis ou par réécriture de l'algorithme de Viterbi (programmation dynamique). L'objectif de ces méthodes est de définir les séquences de probabilité maximale pour définir le meilleur alignement.

De façon assez similaire et suite aux précédents travaux (cf. 2.2.1.4) réalisés dans le domaine de la normalisation de SMS (Beaufort *et al*., 2008), de la reconnaissance d'unités en synthèse vocale et de la correction orthographique (Beaufort, 2008), R. Beaufort (Beaufort, 2010 ; Cougnon & Beaufort, 2009 ; Beaufort & Roekhaut, 2011) propose une variante au transducteur proposé précédemment. Dans cette nouvelle approche, les règles de réécriture définies par observation du corpus font place à des opérations d'édition pondérées par des poids, définis de manière stochastique. Dans le cadre de la normalisation de SMS, L.-A. Cougnon et R. Beaufort (2009) définissent les poids de manière non supervisée à l'aide d'un algorithme itératif. Plus une transition, c'est-à-dire une opération d'édition, est utilisée, plus son poids est conséquent ; un réajustement est opéré à chaque itération. En 2011, cette méthode est appliquée dans le domaine de l'Apprentissage et de l'Enseignement des Langues Assistés par Ordinateur (ALAO et ELAO) pour l'automatisation d'exercices de dictée (Beaufort & Roekhaut, 2011). Le transducteur doit alors permettre de corriger et analyser les erreurs. Il est associé à une analyse morphosyntaxique automatique afin de générer un diagnostic. L'avantage de cette méthode est de permettre d'identifier les erreurs constituées d'une substitution d'un ou plusieurs caractères par un nombre différent de caractères.

2.3. Méthodes d'alignement

Des différentes études mentionnées ci-dessus se dégagent différentes mesures de similarité ou de distance pour aligner ou faire correspondre des formes, à savoir les mesures de distance de Levenshtein, implémentées de manière dynamique ou à l'aide de graphes tels que les transducteurs à états finis, la recherche de n-grammes ou de séquences communes, ou encore la recherche par clés, comme le sont les clés phonétiques.

Dans bien des cas, la recherche de la forme à apparier se fait directement dans la liste des formes d'un lexique, pour les tâches de constitution de lexique bilingue ou encore de lemmatisation, ou des formes / segments de la version normalisée / standard / phonétisée, dans le cas de tâches de normalisation ou de correspondance texte-son. Cependant, dans certains cas, il est nécessaire de générer au préalable des formes dérivées, souvent à l'aide de règles, qui seront elles-mêmes comparées à ces listes de formes.

2.4. Unité d'alignement

Contrairement aux algorithmes d'alignement basés sur des indices syntaxiques ou sémantiques, la plupart des algorithmes basés sur des indices graphiques ou phonétiques se basent sur des comparaisons au niveau des caractères et très peu au niveau des formes ou des segments. Les méthodes qui permettent ensuite de faire correspondre les formes identifiées sont assez peu décrites, ou alors la production ou le texte étudié est vu comme une séquence unique de caractères et le niveau des formes disparait de l'analyse, comme c'est le cas dans les travaux de l'équipe du CENTAL (Beaufort *et al*., 2008 ; Beaufort, 2010) sur la normalisation de SMS.

Précisons également que, généralement, pour ces algorithmes d'alignement, les comparaisons se font entre un caractère et un autre caractère ou une forme et une autre forme. Comme nous l'avons vu, ce phénomène peut poser des problèmes (phénomènes d'agglutination dans les SMS par exemple, ou encore correspondance d'un son avec deux lettres en tâche de synthèse vocale). Certaines méthodes (Véronis, 1988 ; Jansche, 2001 ; Scherrer, 2008) essaient de pallier ce problème en élargissant à plusieurs unités le contexte d'étude, ce que Jansche (2001) appellent des *fenêtres* et Scherrer (2008), à sa suite, des *fenêtres glissantes*.

2.5. Adaptabilité des méthodes

Il semble important également de faire la distinction entre les méthodes qui nécessitent un apprentissage et donc des données préalablement étiquetées ou annotées (en cas d'apprentissage supervisé), des méthodes qui ne nécessitent pas d'apprentissage ou un apprentissage non supervisé, sans données préalables. En effet, s'agissant des corpus scolaires,

il n'existe pas encore de large projet qui permette de disposer d'un ensemble de données suffisant pour entrainer un système. Les méthodes nécessitant un apprentissage ne peuvent donc être adaptées à notre corpus.

Enfin, G. S. Mann et D. Yarowsky (2001) distinguent les méthodes dites adaptatives des méthodes statiques. Les méthodes adaptatives sont celles qui nécessitent d'être adaptées à la langue ou au contexte, soit parce qu'elles utilisent des règles développées spécifiquement, soit parce qu'elles ont dû être entrainées sur des données. À l'inverse, les méthodes statiques ne dépendent pas de la langue ou du contexte et peuvent être appliquées à d'autres données sans adaptation.

3. Conclusion

La profusion des méthodes et des algorithmes utilisés pour des tâches d'alignement et d'appariement traduit la diversité des données et des corpus manipulés. En effet, bien que certaines méthodes n'aient besoin d'aucune adaptation pour être appliquées à d'autres données, elles ne sont pas forcément les plus pertinentes pour les données considérées. Il est donc souvent nécessaire d'adapter un algorithme ou d'en développer un nouveau selon les données traitées et la tâche à effectuer.

Il en va de même pour l'alignement réalisé dans le cadre de la constitution du corpus *Scoledit*, pour lequel nous avons fait le choix de développer un algorithme original, appelé *AliScol*. Les caractéristiques et le fonctionnement de cet algorithme, sont présentés dans le chapitre suivant.

Chapitre 8 - *AliScol*, un système d'alignement pour les écrits scolaires

1. Composition du corpus de développement 158
2. Unité et critères d'alignement 159
3. Prétraitements nécessaires 161
4. Algorithme d'alignement 166
5. Conclusion 179

Les algorithmes d'alignement ne sont pas nouveaux et nous avons pu en présenter un échantillon au chapitre précédent. Cependant, aucun d'entre eux n'a été développé ni ne convient pour le traitement des corpus scolaires francophones. Dans le cadre du projet *Scoledit*, il a donc été nécessaire de développer, à partir des algorithmes existants, un nouvel algorithme, que nous appelons *AliScol*[86]. Cet algorithme a été spécialement conçu pour prendre en considération les caractéristiques des corpus d'apprenants en début d'apprentissage de l'écriture.

Notre étude s'inscrivant dans un projet visant à l'étude des productions d'enfants, il est important de maitriser et comprendre les indices de comparaison utilisés pour l'alignement entre production transcrite et production normalisée. Ce choix s'explique par la volonté de comprendre les erreurs produites par les enfants et de proposer un premier diagnostic. Nous excluons donc de notre algorithme les indices de types stochastiques. En revanche, dans la mesure où bon nombre d'erreurs produites sont orthographiques, il est intéressant de se baser sur une comparaison de type graphique afin de conserver des informations graphiques. Dans les cas, fréquents chez les jeunes apprenants, où la graphie est trop éloignée de la graphie attendue, un recours aux indices phonologiques semble pertinent. L'algorithme prend donc en compte des indices de type phonétique et graphique. Cet aspect de la méthode est développé dans la section 2.

La distance d'édition de Levenshtein est une mesure maintes fois éprouvée, et les travaux reportés au chapitre précédent en sont la preuve. Nous nous sommes donc appuyés sur cette mesure et en avons proposé une variante, plus adaptée à notre corpus (section 4).

[86] Le code de cet algorithme, rédigé en langage Perl, est disponible à l'adresse https://claire.wolfarth.cf/.

Un des usages de notre algorithme d'alignement étant de constituer un lexique et d'analyser l'usage des formes au sein de notre corpus, l'alignement que nous proposerons a pour unité la forme (pour la version normalisée) ou le segment (pour la version transcrite). Cependant, un recours à une comparaison par caractère est souvent nécessaire. Afin de traiter les phénomènes d'hypersegmentation et d'hyposegmentation, nous avons également recours à la notion de fenêtre glissante. Le détail de ces traitements est donné à la section 4.2.1.

À partir des contraintes et besoins exposés précédemment, nous proposons la méthode suivante (Figure 48) :

1. Une phase de prétraitement permet la sélection des productions, la mise en forme de celles-ci et l'ajout d'informations linguistiques (lemme, catégorie syntaxique, phonétisation) pour correspondre aux besoins de l'algorithme.
2. L'algorithme d'alignement basé sur des indices graphiques, phonologiques et archiphonologiques (contraintes phonologiques relâchées) permet de faire correspondre les segments issus de la version transcrite et les formes issues de la version normalisée.
3. Une phase de post-traitement permet de mettre en forme les données pour faciliter leur affichage et leur sauvegarde.

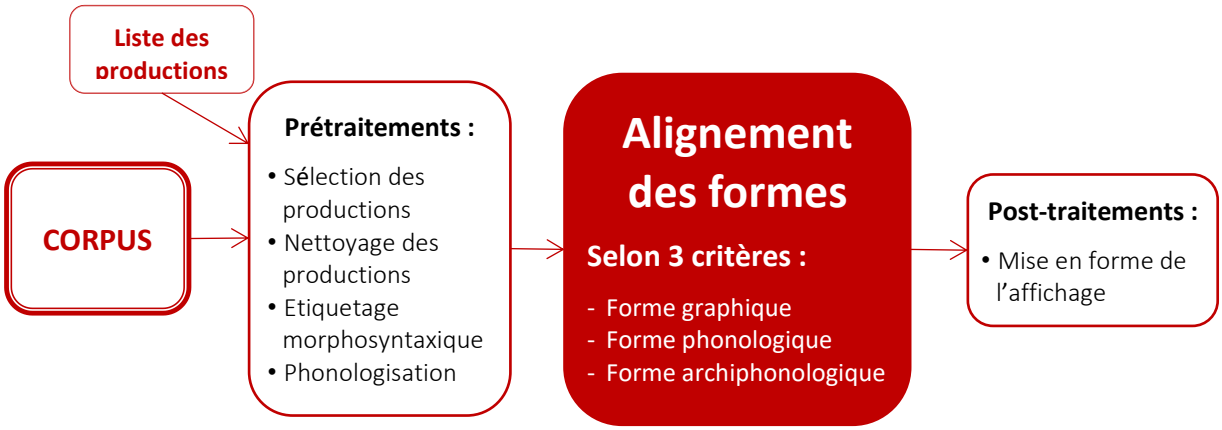


Figure 48 : Schéma général de l'algorithme d'alignement

1. Composition du corpus de développement

Le corpus de développement (cf. Glossaire) est un sous-corpus à partir duquel sont réalisées des observations en vue d'élaborer le programme et sur lequel l'aligneur est testé tout au long de son élaboration. Il est important que ce corpus contienne le plus grand nombre de phénomènes présents dans le corpus total afin d'envisager la majorité des cas de figure au moment de l'élaboration de l'outil. Par moment, lorsque nous avions connaissance de certains

phénomènes non présents dans ce corpus de développement, il nous est arrivé de nous référer à d'autres productions.

Afin de développer cet algorithme, nous nous sommes appuyés sur un corpus de développement composé de 53 productions, réparties du CP au CE2 (cf. Annexe 1). En raison du recueil longitudinal du corpus, les productions de CM1 et de CM2 n'étaient pas encore disponibles au moment de la phase de développement de l'algorithme.

Le corpus de développement élaboré rassemble des productions réparties de manière à ménager un équilibre entre les trois niveaux, du CP au CE2, à la fois en termes de nombre de productions et de nombre de tokens.

2. Unité et critères d'alignement

L'objectif de l'outil présenté ici est de mettre en correspondance les segments écrits par l'élève et les formes normalisées par le biais d'un alignement entre productions transcrites et productions normées, selon le principe développé dans notre approche par comparaison. Cet alignement permet ensuite d'inférer une catégorisation des formes présentes en terme de distance à la norme.

Cependant, l'unité d'alignement utilisée est le segment. Elle présente l'avantage de se fonder directement sur la production de l'élève. Choisir le segment[87] pour unité permet de gérer les erreurs de segmentation telles que les erreurs d'*hyposegmentation* (exemple : « quelle » pour *qu'elle*) et les erreurs d'*hypersegmentation* (exemple : « en semble » pour *ensemble*) dont les corpus scolaires présentent de nombreuses occurrences.

Lors d'un précédent travail (Wolfarth, 2015) effectué sur une portion restreinte des productions de CP, nous avons montré que pour près de 75 % des formes erronées, la phonologie de la forme normée est la même (exemple : « cha » pour *chat*). Par conséquent et à l'instar de K. Berkling et ses collègues (2011), nous choisissons de ne pas nous appuyer uniquement sur des indices graphiques pour l'alignement mais également sur des indices phonologiques et archiphonologiques.

2.1. Comparaison à l'aide d'indices graphiques

La comparaison à l'aide d'indices graphiques repose sur les formes graphiques des segments transcrits et des segments normés. Elle permet d'identifier les segments correctement

[87] Est entendu comme segment une séquence de caractères encadrée de deux séparateurs, c'est-à-dire des espaces, des signes de ponctuations, des retours à la ligne, etc.

orthographiés ou à la graphie proche de la forme normée (exemple : « réflichi » pour *réfléchit*, 1560 - CM1, cf. la notion de *comparaisons relatives*, section 4.1).

2.2. Comparaison à l'aide d'indices phonologiques

Afin de pouvoir baser la comparaison sur des indices phonologiques, les segments transcrits et normés sont convertis en représentation phonologique à l'aide d'un outil de phonologisation, en l'occurrence *LIA-PHON*[88]. Cette comparaison permet d'identifier les segments comportant des erreurs ne modifiant pas ou peu la phonologie (exemple : « fesait » pour *faisaient*).

2.3. Comparaison à l'aide d'indices archiphonologiques

Dans de nombreux cas, des segments dont on pourrait penser qu'ils transcrivent la même valeur phonologique ne sont pas considérés comme tels par l'outil de phonologisation *LIA-PHON* pour des raisons d'ambigüités de prononciation. Ces ambigüités peuvent être notamment dues à des différences contextuelles ou à des variantes régionales ou sociales. Par exemple, les graphèmes *o* et *au* peuvent se prononcer /o/ ou /ɔ/ selon le contexte. Lorsque la forme *saute* et le segment « sote » sont fournies à l'outil *LIA-PHON*, celui-ci génère deux représentations phonologiques différentes. De même, selon ses habitudes linguistiques, il n'est pas toujours aisé de discriminer les valeurs des différentes finales verbales en /e/ ou /ɛ/, à l'exemple d'*allait*, *aller* ou *allé*, ou encore les formes *es* et *est*. L'outil *LIA-PHON* marque cependant ces différences.

Pour toutes ces raisons, une représentation archiphonologique des transcriptions et des normalisations est produite. C'est-à-dire que certaines oppositions phonologiques sont neutralisées, en accord avec la notion d'archiphonème de N. Catach. Un archiphonème est un « représentant de l'ensemble des traits phoniques pertinents communs à deux ou plusieurs phonèmes, qui sont par rapport aux autres dans un rapport exclusif » (Catach, 1980, p. 16).

En accord avec cette théorie, nous neutralisons les oppositions suivantes : /ø/, /œ/ et /ə/ ; /e/ et /ɛ/ ; /o/ et /ɔ/ ; /ɛ̃/ et /œ̃/. Cette neutralisation répond à plusieurs contraintes :

- Les différences régionales et d'idiolecte : les distinctions entre ces phonèmes sont plus ou moins marquées entre les personnes, les générations et les régions ;
- La réalité des productions des jeunes apprenants : à la lecture des productions des apprenants, on remarque que ces oppositions ne sont souvent pas encore maitrisées. Tenir compte de cette difficulté doit donc permettre d'améliorer l'efficience de notre système ;

[88] Béchet, F. (2001). *LIA-PHON*: Un système complet de phonétisation de textes. TAL. Traitement automatique des langues, 42(1), 47-67.

- L'ambigüité des graphèmes : un même graphème peut, selon le contexte dans lequel il est placé, encoder l'un ou l'autre des phonèmes des oppositions précités (*o* et *au* pour /o/ et /ɔ/). Une modification des graphèmes connexes peut donc entrainer une modification de la représentation phonologique de ces graphèmes. Citons l'exemple du couple *sot* /so/ et *note* /nɔt/. Dans le premier cas, la suite de lettre *ot* se prononce /o/, tandis que dans le second cas, cette même suite de lettres se prononce /ɔt/, en raison de l'ajout du *e* final. Certains graphèmes, qui ne se prononcent pas, peuvent tout de même modifier la prononciation.

Par convention, les archiphonèmes sont représentés en majuscules.

Phonème 1	*Phonème 2*	*Archiphonème*
/e/	/ɛ/	/E/
/o/	/ɔ/	/O/
/ø/	/œ/	/Œ/
/œ̃/	/ɛ̃/	/Œ̃/

Tableau 8 : Correspondance entre phonèmes et archiphonèmes

Pour résumer, pour la forme *faire* et le segment « fér » nous disposons de trois représentations différentes (Tableau 9) qui sont autant d'indices de comparaisons différents.

	Représentation graphique	*Représentation phonologique*	*Représentation archiphonologique*
Exemple 1	faire	/fɛR/	/fER/
Exemple 2	fér	/feR/	/fER/

Tableau 9 : Les différentes représentations d'une forme, les exemples de faire et « fér »

3. Prétraitements nécessaires

L'approche par comparaison, comme n'importe quelle autre approche de traitement automatique, peut nécessiter une série de prétraitements. Afin de reprendre l'approche proposée par O. Kraif et C. Ponton (2007) et de permettre une souplesse des outils existants utilisés, nous avons fait le choix de rendre ces prétraitements modulaires.

a) Fonction *select_prod* : ce premier prétraitement permet de sélectionner les productions à traiter. Ces productions peuvent être sélectionnées selon des critères donnés par l'utilisateur : liste d'identifiants et de niveaux ; mais ce prétraitement permet également d'écarter les productions non complètes (non transcrites ou non normées). Cette sélection est particulièrement intéressante dans le cadre d'un corpus longitudinal

puisqu'elle permet à l'utilisateur d'étudier le ou les niveaux qui l'intéressent. Mais, puisque le corpus est longitudinal et qu'il demeure donc longtemps en cours de construction, cette sélection permet aussi de ne conserver que les productions dont la numérisation est achevée.

b) Fonction *effaceBal* : ce prétraitement permet d'effacer les balises incluses dans la transcription qui ne sont pas nécessaires à l'alignement. En effet, ces balises ne sont que des indications spatiales ou d'éléments spécifiques. Par exemple, la balise <revision>, présente dans de nombreuses transcriptions, est une balise qui marque la présence d'une rature. En revanche, les balises de la normalisation sont conservées car utiles. Par exemple, les balises <segmentation/> et <s/>, marques de ponctuations absentes, participent à la construction de la syntaxe et du sens.

c) Fonction *etiquetage* : l'objectif de ce traitement est d'étiqueter les productions à l'aide d'un étiqueteur morphosyntaxique, *TreeTagger* dans notre cas. Après tokenisation, le processus d'étiquetage morphosyntaxique consiste à attribuer à chaque forme un lemme et une catégorie grammaticale, éventuellement aussi certains traits flexionnels pour certaines catégories comme les verbes, les pronoms et les déterminants. Le tableau 10 illustre le résultat fourni par *TreeTagger* (cf. Annexe 7 pour le liste des étiquettes utilisées par cet outil) pour la séquence *La sorcière voulait manger le chat.* Ce processus implique donc une étape de tokenisation. Ce traitement est bien connu en TAL et représente donc une étape relativement fiable.

Forme	*Catégorie*	*Lemme*
La	DET:ART	le
sorcière	NOM	sorcier
voulait	VER:impf	vouloir
manger	VER:infi	manger
le	DET:ART	le
chat	NOM	chat
.	SENT	.

Tableau 10 : Étiquetage morphosyntaxique à l'aide de l'outil TreeTagger

d) Fonction *phonetisation* : cette étape permet de convertir en représentation phonologique les deux versions du corpus, la transcription et la normalisation. Pour ce faire, le programme fait appel à l'outil *LIA-PHON* qui, à une liste de tokens donnée (colonne 1), fournit les représentations phonologiques associées (colonnes 2 - 4, cf. Annexe 8 pour la table des correspondances des formats de transcription phonologique), d'un étiquetage morphosyntaxique (colonne 5) et d'une syllabation (colonne 6). Un exemple est donné dans le tableau 11.

1 *Forme*	*2* *Format LIA*	*3* *Format SAMPA*	*4* *Format API HTML*	*5* *Catégorie*	*6* *Syllabes*
la	llaa	l/a	l/a	[DETFS]	la
sorcière	ssoorrssyyairr	s/O/R/s/j/E/R	s/ɔ/r/s/j/ɛ/r	[NFS]	sor | ciè | re
voulait	vvoullai	v/u/l/E	v/u/l/ɛ	[V3S]	vou | lait
manger	mmanjjei	m/a~/Z/e	m/ɑ̃/ʒ/e	[VINF]	man | ger
le	lleu	l/@	l/ə	[DETMS]	le
chat	chaa	S/a	ʃ/a	[NMS]	chat

Tableau 11 : Représentation phonologique produite par l'outil LIA-PHON[89]

Les représentations phonologiques sont données dans trois formats différents :

- Le **format API** (colonne 4). Ce format est couramment utilisé en linguistique. Dans ce format, un phonème est représenté par un caractère. Il nécessite donc un grand nombre de caractères différents et fait donc appel à de nombreux caractères spéciaux qui peuvent poser des problèmes de format.
- Le **format SAMPA** (colonne 3). Ce format vise une simplification des caractères spéciaux utilisés pour éviter les problèmes d'encodage. Néanmoins, un phonème peut être représenté par plusieurs caractères (par exemple /a~/) ce qui rend la division en phonèmes plus compliquée.
- Le **format LIA** (colonne 2). Ce format a été développé pour l'usage de l'outil *LIA-PHON* particulièrement adapté à l'orthographe française. Au sein de ce format, tous les phonèmes sont représentés par une combinaison de deux caractères. Aucun caractère spécial n'est utilisé, il n'y a donc pas de problème d'encodage et le découpage est facilité par le nombre fixe de caractères utilisés par phonèmes. C'est ce format que nous avons choisi d'utiliser dans notre travail.

Le processus de phonologisation implique une tokenisation de la séquence de segments ou de formes à phonologiser. Cependant, cette tokenisation n'est pas exploitable en l'état et une seconde étape est nécessaire. En effet, l'objectif de cet outil étant de préparer une séquence à sa phonologisation, certains signes typographiques inutiles dans cette optique sont omis, d'autres sont convertis pour correspondre à une valeur sonore. Par exemple, les majuscules n'apparaissent plus et certains signes de ponctuation sont indiqués comme des « PAUSE »

[89] Les balises </s> et <FIN> sont des balises ajoutées par *LIA-PHON* qui marquent respectivement la fin d'une phrase et la fin d'un texte.

(exemple en tableau 12 de la séquence *Sa maman le prend*). Une étape de rétablissement des segments d'origine est nécessaire.

Forme	*Représentation phonologique*	*Catégorie grammaticale*	*Syllabes*	
sa	ssaa	[MOTINC->propername_5]	sa	
maman	mmaamman	[NFS]	ma	man
le	llee	[DETMS]	le	
prend	pprran	[V3S]	prend	
pause	##			
</s>	#??#	[ZTRM->EXCEPTION]	</s>	
<FIN>	????	[]	<FIN>	

Tableau 12 : Représentation phonologique produite par l'outil LIA-PHON

e) Fonction *archiphonétisation* : lors de cette étape, chaque représentation phonologique est augmentée d'une représentation archiphonologique qui neutralise les oppositions évoquées précédemment. Cette étape neutralise également le *schwa*, que l'outil *LIA-PHON* choisit de transcrire ou non selon les segments phonétisés (exemple dans les tableau 13 et tableau 14).

Forme	*Représentation phonologique*	*Représentation archiphonologique*
la	llaa	llaa
sorcière	ssoorrssyyairr	ssaurrssyyeirr
voulait	vvoullai	vvoullei
manger	mmanjjei	mmanjjei
le	lleu	llee
chat	chaa	chaa

Tableau 13 : Représentation archiphonologique produite lors de la phase de prétraitement (exemple 1)

Forme	*Représentation phonologique*	*Représentation archiphonologique*
il	iill	iill
se	ssee	ssee
sent	ssan	ssan
triste	ttrriissttee	ttrriisstt

Tableau 14 : Représentation archiphonologique produite lors de la phase de prétraitement (exemple 2)

f) Fonction *fusionPhonMs* : comme nous l'avons évoqué, l'emploi des outils *TreeTagger* et *LIA-PHON* implique deux processus de tokenisation. Toutefois, ces tokenisations ne se basent pas sur les mêmes règles et les résultats de ces tokenisations sont donc différents (exemple dans le tableau 15). Une étape permettant de fusionner les sorties de ces deux outils est donc nécessaire.

Tokenisation de l'outil LIA-PHON	*Tokenisation de l'outil TreeTagger*
<s>	Il
il	était
était	une
une_fois	fois
une	une
sorcière	sorcière
qui	qui
s'	s'
appelait	appelait
Meg	Meg
pause	.
</s>	Un
<s>	jour
un	ils
jour	allèrent
ils	visiter
allèrent	un
visiter	zoo
un	,
zoo	
pause	

Tableau 15 : Exemple de tokenisations des outils LIA-PHON et TreeTagger

À la fin de ces prétraitements, nous disposons donc de deux listes de tokens (segments), avec les informations suivantes :

Token transcrit				
Segment transcrit	Représentation phonologique	Représentation archiphonologique		
Token normé				
Segment normé	Représentation phonologique	Représentation archiphonologique	Catégorie grammaticale	Lemme

Tableau 16 : Données accessibles à la fin de la phase de prétraitements

4. Algorithme d'alignement

Comme nous l'avons dit précédemment, notre objectif était de développer un algorithme d'alignement qui prenne en compte les spécificités des corpus scolaires, et particulièrement les spécificités du corpus *Scoledit*. Pour rappel, les spécificités majeures sont :

- Un fort taux d'erreurs ne modifiant pas la phonologie ou générant des variantes phonologiques proches, par exemple « révéier » pour *réveillé*.
- La **présence d'erreurs de segmentation**, par exemple « ilfut » pour *il fut* (phénomène d'hyposegmentation) ou « la pin » pour *lapin* (phénomène d'hypersegmentation).

Cette première spécificité est prise en compte dans le choix des indices de comparaison, à savoir une comparaison à l'aide d'indices graphiques, phonologiques et archiphonologiques. La prise en compte de la deuxième spécificité se fait dans le parcours de l'algorithme, qui détermine le ou les segments comparés entre eux.

Pour ce faire, nous avons principalement développé et testé deux algorithmes, (A) un algorithme plus séquentiel qui s'inspire du concept des fenêtres glissantes proposées par Jansche (2001) et (B) un algorithme issu de la programmation dynamique et qui reprend l'algorithme de calcul de la distance de Levenshtein proposé par R. A. Wagner et M. J. Fischer (1974).

Dans les deux algorithmes, les mêmes mesures de comparaison sont utilisées. À la fois, une comparaison stricte entre deux chaines de caractères et une comparaison relative à l'aide d'un calcul de la distance de Levenshtein. Chacune de ces comparaisons est réitérée sur les trois versions de chaque segment : la forme graphique, la représentation phonologique, la représentation archiphonologique.

4.1. Mesure de comparaison

Pour chaque paire de segments comparés, nous pouvons donc appliquer six comparaisons.

1. Trois **comparaisons strictes** : il s'agit de comparaison au caractère près entre deux chaines de caractères.
    a. **Comparaison graphique stricte** : il s'agit d'une comparaison orthographique uniquement. Si les deux segments, normés et transcrits sont identiques modulo les majuscules, les deux segments sont appariés ; par exemple, « il » || *il.*

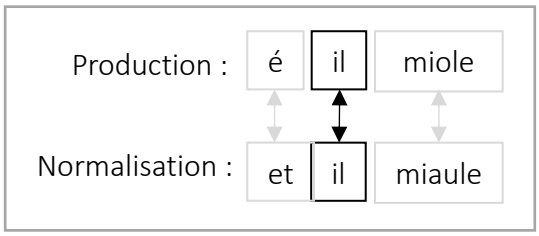


Figure 49 : Résultat de la comparaison graphique stricte

    b. **Comparaison phonologique stricte** : il s'agit d'une comparaison stricte entre les représentations phonologiques des segments transcrits et normés ; par exemple, « ce » || se.

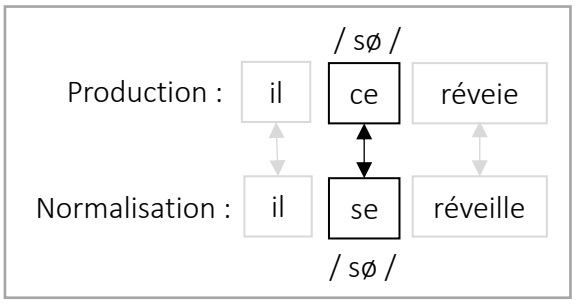


Figure 50 : Résultat de la comparaison phonologique stricte

    a. **Comparaison archiphonologique stricte** : il s'agit d'une comparaison stricte entre les représentations archiphonologiques des segments transcrits et normés ; par exemple, « miole » || miaule.

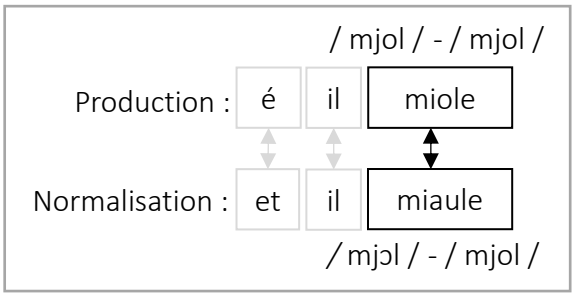


Figure 51 : Résultat de la comparaison phonologique avec archiphonologique stricte

2. Trois **comparaisons relatives** : pour réaliser cette comparaison, un calcul de la distance de Levenshtein est réalisé entre les deux chaines concernées. Si la distance est

inférieure à la moitié de la longueur de la plus courte chaine de caractère[90], on considère que la forme et le segment considéré sont comparables et peuvent être alignés.

a. **Comparaison de la distance graphique** : comparaison à partir des formes graphiques, à nouveau la casse n'est pas prise en compte ; par exemple : « réveie » || *réveille*. La proximité graphique de ces deux formes (6 lettres communes sur les 6 lettres de la chaine la plus courte) permet à l'algorithme d'aligner ces deux formes.
b. **Comparaison de la distance phonologique** : comparaison à partir des représentations phonologiques ; par exemple : « ce » - /sø/ || *que* - /kø/. La proximité phonologique (moitié des phonèmes en commun) permet à l'algorithme d'aligner ces deux formes.
c. **Comparaison de la distance archiphonologique** : comparaison à partir des représentations archiphonologiques ; par exemple : « chéte » - /ʃet/ - /ʃEt/ || *jette* - /ʒɛt/ - /ʒEt/. La proximité archiphonologique (moins de la moitié des phonèmes en commun, mais plus de la moitié si les archiphonèmes sont pris en compte) permet à l'algorithme d'aligner ces deux formes.

Ces comparaisons sont hiérarchisées dans l'ordre où elles sont exposées ici (Figure 52). Ainsi, nous faisons par exemple l'hypothèse que privilégier les erreurs n'influençant pas la phonologie (comparaison phonologie stricte, comme « jète » / *jette*) est plus pertinent que privilégier les erreurs impactant peu la graphie mais modifiant la phonologie (comparaison de la distance graphique, « chette » / *jette*).

[90] Le choix de la formule déterminant ou non l'alignement représente une première approximation, réalisée et modifiée afin de correspondre au mieux aux phénomènes rencontrés lors de la conception de l'algorithme. Il convient de l'évaluer, mais par manque de temps nous n'avons pas encore pu réaliser cette évaluation.

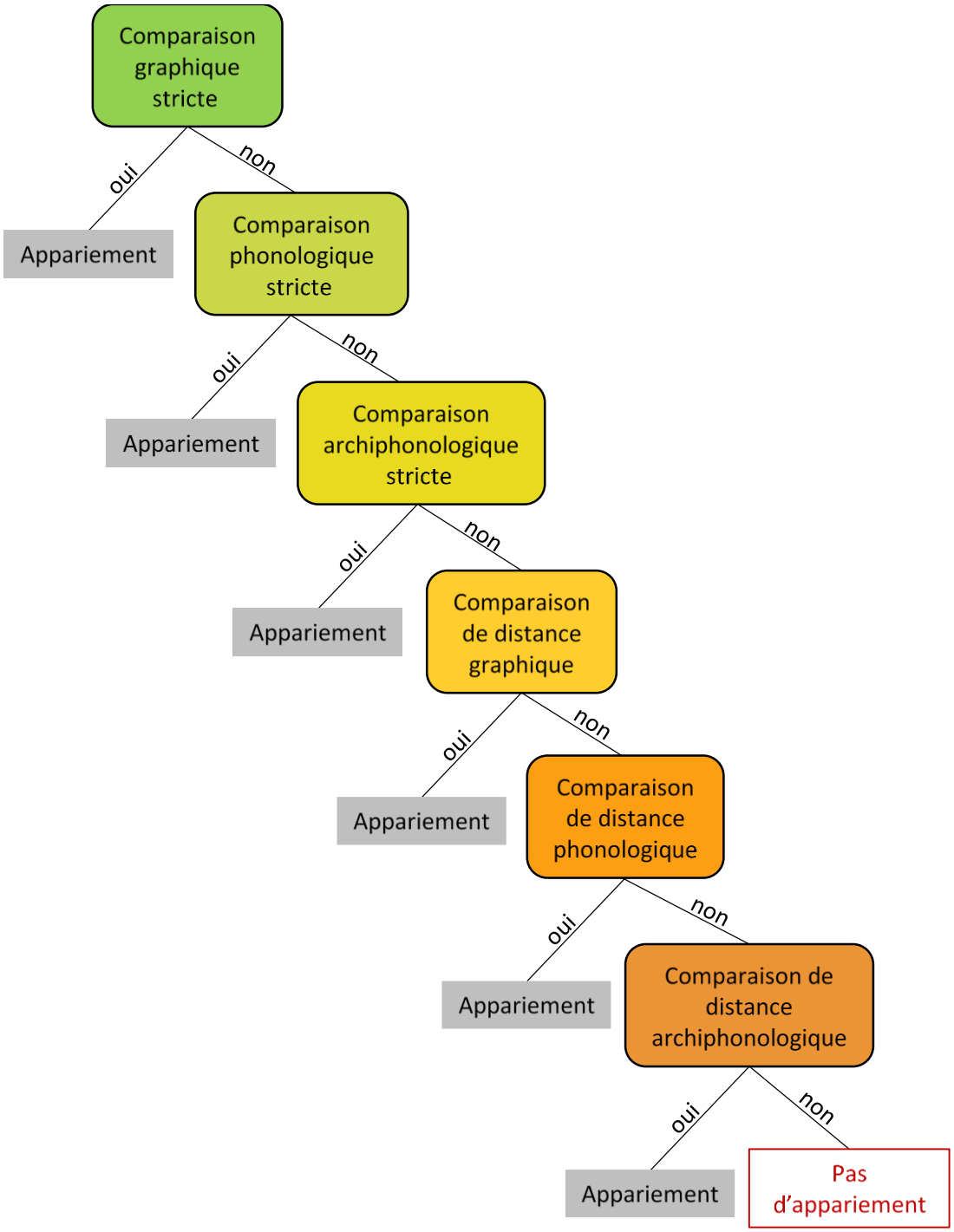


Figure 52 : Ordre des comparaisons

4.2. Parcours algorithmique

4.2.1. Parcours séquentiel (A)

L'algorithme développé en premier lieu est un algorithme séquentiel qui permet de gérer les différentes spécificités des écrits d'enfants, à savoir les phénomènes d'hyposegmentation, d'hypersegmentation, d'omission et d'insertion.

1. Hypersegmentation. En cas d'hypersegmentation, une forme normée est transcrite par plusieurs segments.

- Exemple 1 : « quelle » / *qu'elle* (1927, CM1)

  Alignement : quelle || qu' + elle

- Exemple 2 : « vatan » / *va-t'en* (982, CE1)

  Alignement : vatan || va + - + t' + en

2. Hyposegmentation. En cas d'hyposegmentation, un seul segment transcrit plusieurs formes normées.

- Exemple 1 : « a pré » / *après* (835, CP)

  Alignement : a + pré || après

- Exemple 2 : « trenc-forma » / *transforma* (582, CE2)

  Alignement : trenc + - + forma || transforma

3. Omission. Sont qualifiés de segments omis les segments non produits par l'élève mais nécessaires à la structure de la phrase, comme c'est souvent le cas du *ne* de négation, par exemple « il avait rien a manger » / *Il n'avait rien à manger* (2981, CE1). Dans ce cas, une forme normée, ne correspond à aucun segment transcrit.

- Exemple 1 : « il avait rien a manger » / *Il n'avait rien à manger*
  - Alignement : Ø || n'

4. Insertion. Plus rare, le phénomène d'insertion renvoie au cas où l'élève a produit des segments non pertinents dans la structure de la phrase, par exemple l'ajout d'un élément marquant une liaison, comme dans l'exemple « Et dans un n'autre monde » / *Et dans un autre monde* (2937, CM1). Dans ce cas, un segment transcrit ne correspond à aucune forme normée.

- Exemple 1 : « un n'autre monde » / *un autre monde*
  - Alignement : n' || Ø

Pour prendre en compte ces phénomènes, nous utilisons des fenêtres glissantes. Ce concept est repris à Y. Scherrer (2007) qui s'est lui-même inspiré de M. Jansche (2001). Alors que dans un algorithme classique d'alignement, les unités de comparaison, par exemple les segments, sont comparées les unes avec les autres, une à une (Figure 53), les fenêtres glissantes permettent de comparer une unité à plusieurs unités. Le nombre d'unités comparées est déterminé par la taille de la fenêtre glissante (Figure 54).

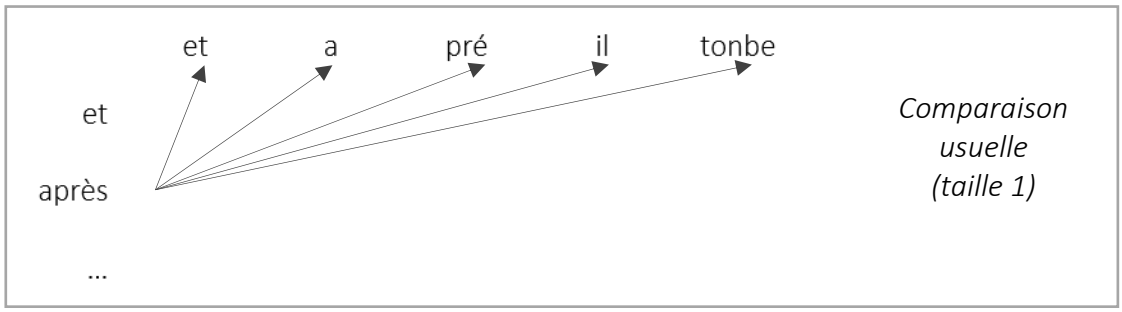


Figure 53 : Comparaison sans fenêtre glissante (= fenêtre glissante de taille 1)

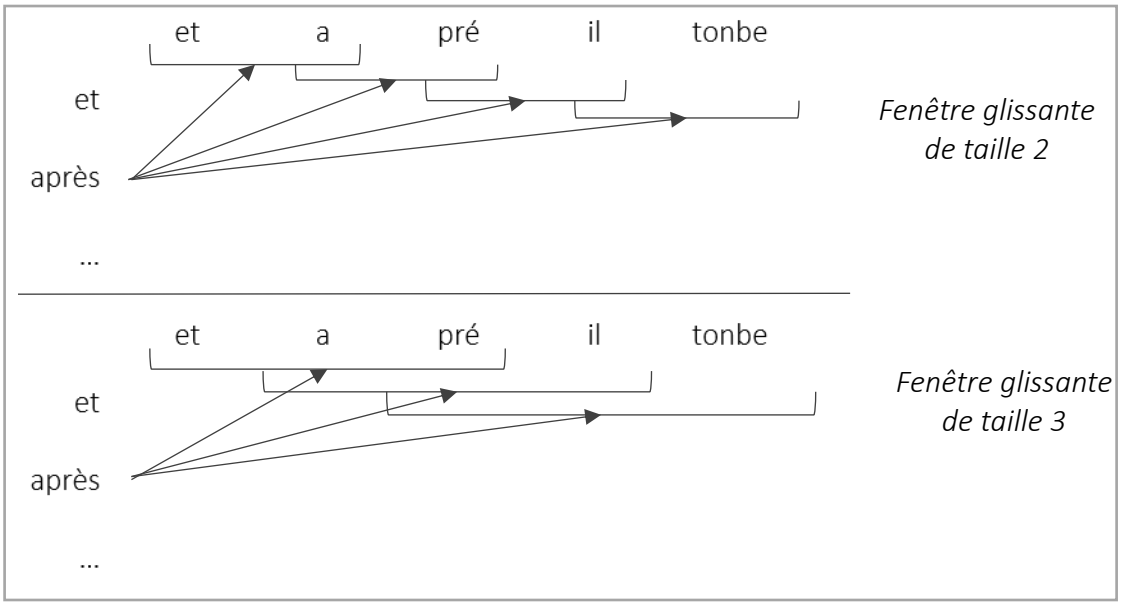


Figure 54 : Comparaison avec fenêtre glissante

L'algorithme d'alignement opère donc une comparaison entre les différentes représentations des tokens (représentation graphique, représentation phonologique et représentation archiphonologique, axe vertical dans la figure 55) et une comparaison linéaire des chaines de tokens (axe horizontal dans la figure 55).

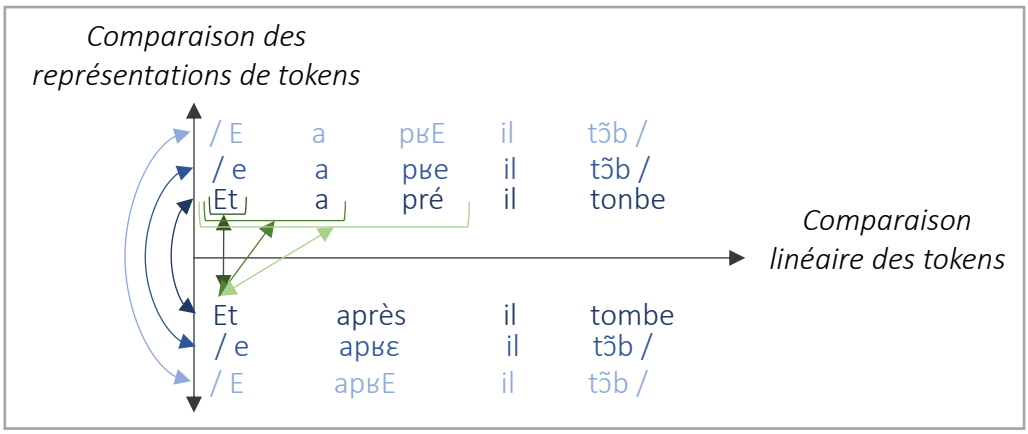


Figure 55 : Axes de comparaison de l'algorithme d'alignement

La figure 56 représente les différentes étapes de l'algorithme séquentiel. Chacun des blocs (numérotés de 1 à 7) y représente une nouvelle étape de comparaison. Ces étapes sont effectuées de manière séquentielle : les étapes ne sont effectuées qu'à condition que la précédente soit un échec[91]. À chacune de ces étapes, les six comparaisons explicitées à la section précédente (Figure 52) sont effectuées.

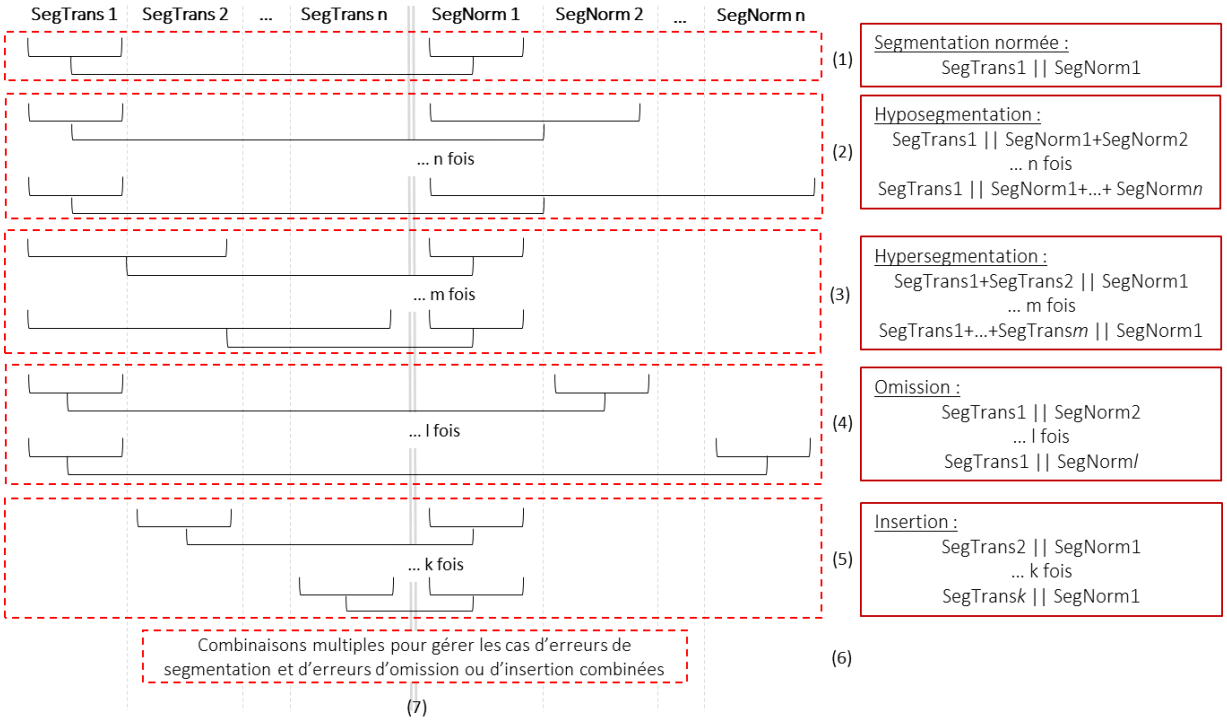


Figure 56 : Schéma séquentiel de l'algorithme d'alignement A

L'algorithme utilisé est un algorithme séquentiel qui considère chaque forme l'une après l'autre et ne s'interrompt qu'une fois l'ensemble des formes de chacune des versions examinées. Il procède comme suit :

(1) Dans le cas le plus courant, l'algorithme procède à une comparaison des formes de même rang (comparaison du premier segment transcrit avec la première forme normalisée, comparaison du deuxième segment transcrit avec la deuxième forme normalisée, etc.). Si les formes sont jugées équivalentes, la segmentation est jugée normée, leur alignement est sauvegardé et l'algorithme interrompt le processus séquentiel et considère les segments suivants. À chaque équivalence trouvée, dans cette étape ou dans une des étapes suivantes, ces mêmes opérations seront appliquées : sauvegarde et interruption de la séquence ;

[91] L'ordre d'exécution de ces modules a fait l'objet d'une évaluation (cf. Chapitre 9).

```
POUR (n de 0 à longueur(Normalisation)) {
  SI (COMPARER(Transcription[n],Normalisation[n]) == TRUE) {
    ENREGISTRER(Transcription[n],Normalisation[n]) ;
  }
}
```

Algorithme 1 : Extrait d'algorithme de gestion de la segmentation normée (bloc 1)

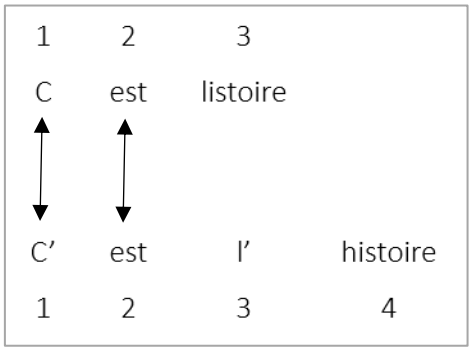


Figure 57 : Exemple de parcours en cas de segmentation normée

(2) et (3) Afin de gérer les erreurs d'hyposegmentation, comme « sereveille » (*se réveille*), et d'hypersegmentation, comme « dé sida » (*décida*), nous avons dû implémenter une *fenêtre glissante* qui envisage la comparaison de plusieurs segments produits à une forme normalisée (cas d'hypersegmentation) et inversement (cas d'hyposegmentation). L'algorithme ainsi développé autorise une fenêtre de longueur maximale 4 (cf. chapitre 9) pour une détermination de cette longueur), considérant qu'une forme produite donnée peut correspondre à quatre formes normées et inversement ;

```
POUR (n de 0 à longueur(Normalisation) {
  SI (COMPARER(Transcription[n], Normalisation[n].Normalisation[n+1]) == TRUE) {
    ENREGISTRER(Transcription[n], Normalisation[n].Normalisation[n+1]) ;
  }
  SI (COMPARER(Transcription[n], Normalisation[n].Normalisation[n+1].Normalisation[n+2])
    == TRUE) {
    ENREGISTRER(Transcription[n], Normalisation[n].Normalisation[n+1] .Normalisation[n+2]) ;
  }
  SI (COMPARER(Transcription[n].Transcription[n+1],Normalisation[n]) == TRUE) {
    ENREGISTRER(Transcription[n].Transcription[n+1],Normalisation[n]) ;
  }
  SI (COMPARER(Transcription[n].Transcription[n+1].Transcription[n+2],Normalisation[n]) ==
      TRUE) {
    ENREGISTRER(Transcription[n].Transcription[n+1].Transcription[n+2],Normalisation[n]) ;
  }
}
```

Algorithme 2 : Extrait d'algorithme de gestion de l'hyposegmentation et de l'hypersegmentation (blocs 2 et 3)

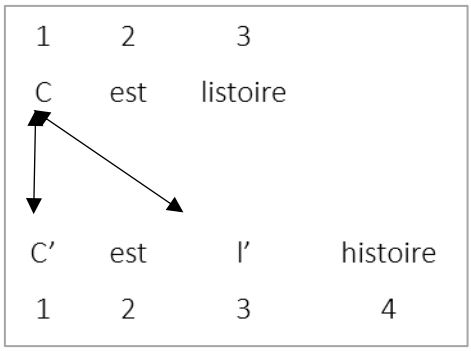


Figure 58 : Exemple de parcours en cas d'hyposegmentation

(4) et (5) Bien que peu de modifications syntaxiques aient été apportées à la normalisation, il se peut que celle-ci comporte tout de même quelques suppressions ou ajouts. L'algorithme permet donc des comparaisons entre la forme produite de rang *n* et la forme normée de rang *n+1*, *n+2* ou *n+3* (cas d'insertion, exemple : « il prend le chat le chat » pour *Il prend le chat.*) et inversement (cas d'omission, exemple : « il a pas vu » pour *Il n'a pas vu*) ;

```
POUR (n de 0 à longueur(Normalisation) {
  POUR (i de 0 à 3) {
    SI (COMPARER(Transcription[n],Normalisation[n+i]) == TRUE) {
      ENREGISTRER(Transcription[n],Normalisation[n+i]) ;
    }
    SI (COMPARER(Transcription[n+i],Normalisation[n]) == TRUE) {
      ENREGISTRER(Transcription[n+i],Normalisation[n]) ;
    }
  }
}
```

Algorithme 3 : Extrait d'algorithme de gestion de l'omission et de l'insertion (blocs 4 et 5)

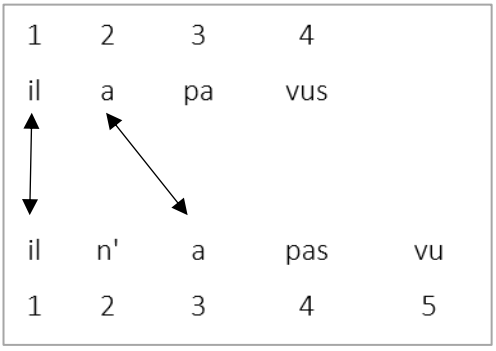


Figure 59 : Exemple de parcours en cas d'omission

(6) Il peut arriver également que les erreurs de segmentation et les erreurs d'omission ou d'insertion soient combinées. L'algorithme permet donc, en combinant les procédures 2, 3, 4 et 5 une recherche de ce type d'erreurs (exemple : « il apa » pour *Il n'a pas vu*).

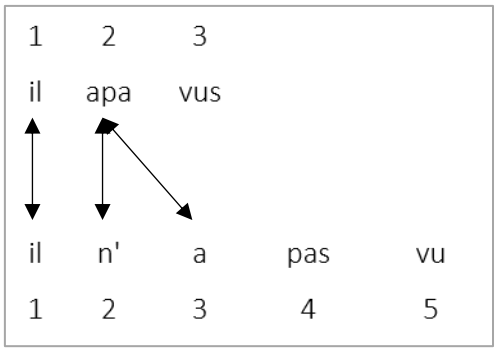


Figure 60 : Exemple de parcours en cas d'omission et d'hyposegmentation combinées

(7) Enfin, l'algorithme se déployant de façon linéaire, il permet d'aligner les segments restants, qui ne peuvent être alignés par les mesures explicitées ici.

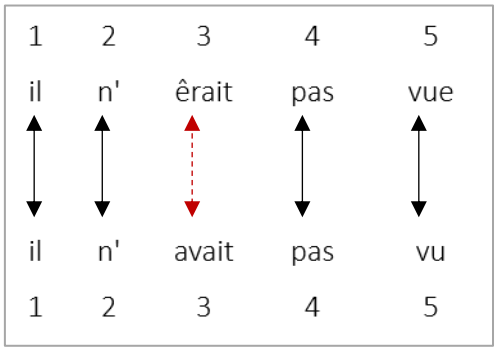


Figure 61 : Exemple d'alignement par défaut

Dans l'exemple de la figue ci-dessus, l'aligneur est en mesure d'aligner les tokens *il – il* ; *n' – n'* ; *pas – pas* ; *vue – vu*. En revanche, il n'est pas en mesure de reconnaitre la correspondance entre les tokens *êrait* et *avait*, mais, les autres tokens étant alignés par ailleurs, il va réaliser un alignement par défaut.

Si ces étapes permettent la comparaison de plusieurs formes entre elles, chaque comparaison est effectuée selon différentes mesures afin d'améliorer les possibilités d'identifier les équivalences.

4.2.2. Parcours matriciel (B)

Le second algorithme développé est un algorithme de programmation dynamique. Il reprend la structure de l'algorithme *Wagner-Fischer* (1974) et propose un parcours matriciel de la production transcrite et de la production normée. Cet algorithme est souvent utilisé pour comparer des chaines de caractères en prenant pour unité de comparaison le caractère. Nous proposons de l'utiliser ici en prenant pour unité les tokens (formes, segments, ponctuants). Ainsi, chaque segment transcrit peut être comparé à l'ensemble des formes normées et inversement. Chacune de ces comparaisons comportent plusieurs étapes :

4.2.2.1. Calcul de la distance d'édition de Levenshtein.

Cette première étape de comparaison permet d'observer les quatre opérations incluses dans la notion de distance d'édition :

- L'omission d'une forme ;
- L'insertion d'une forme ;
- La substitution d'une forme par une autre ;
- L'identité entre les deux formes.

Néanmoins, contrairement à l'algorithme originel (cf. Équation 1), l'opération de substitution ne repose pas uniquement sur des indices graphiques mais également sur des indices phonologiques. En effet, nous nous appuyons sur les six modes de comparaisons détaillés précédemment (comparaison graphique stricte et relative, comparaison phonologique stricte et relative, comparaison archiphonologique stricte et relative). À chacune de ces comparaisons est associé un poids différencié (cf. Équation 2). Ce poids est déterminé selon les priorités des comparaisons et permet de privilégier les substitutions dans l'ordre donné à la section 4.1. La substitution graphique stricte a donc le poids le plus faible, tandis que la substitution archiphonologique relative a le poids le plus élevé.

$$d(\mathrm{i}+1,\mathrm{j}+1) = \min\begin{pmatrix} \mathrm{d}(\mathrm{i}+1,\mathrm{j})+1, \\ \mathrm{d}(\mathrm{i},\mathrm{j}+1)+1, \\ \mathrm{d}(\mathrm{i},\mathrm{j})+\mathrm{cost}(\mathrm{s_i},\mathrm{t_j}) \end{pmatrix}, \mathrm{cost}(\mathrm{a},\mathrm{b}) = \begin{Bmatrix} 0 \text{ si } \mathrm{a}=\mathrm{b} \\ 1 \text{ sinon} \end{Bmatrix}$$

Équation 1: Calcul de la distance de Levenshtein classique

$$d(i+1,j+1) = \min\begin{pmatrix} d(i+1,j)+1, \\ d(i,j+1)+1, \\ d(i,j)+cost(s_i,t_j) \end{pmatrix}, cost(a,b) = \begin{Bmatrix} 0\ si\ a=b \\ > 0\ et < 1\ si\ a \sim b \\ 1\ sinon \end{Bmatrix}$$

Équation 2 : Calcul de la distance de Levenshtein avec poids

Dans ces équations :

- a = b signifie que le résultat de la comparaison graphique stricte est positif. Dans ce cas, le cout est de 0 ;
- a ~ b signifie que le résultat d'une des cinq autres comparaisons (phonologique et archiphonologique strictes, graphique, phonologique et archiphonologique relatives) est positif. Dans ce cas, le cout est inclus entre 0 et 1 ;
- le cout est de 1 lorsque toutes les comparaisons ont échoué.

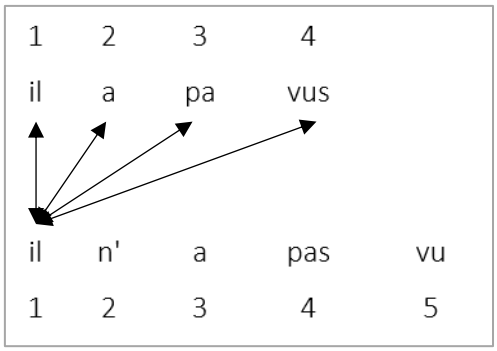


Figure 62 : Exemple de parcours matriciel

	il	n'	a	pas	vu
il	0	1	2	3	4
a	1	1	1	2	3
pas	2	2	2	1	2
vus	3	3	3	2	2

Tableau 17 : Résultat matriciel du calcul d'édition classique

	il	n'	a	pas	vu
il	0	1	2	3	4
a	1	1	1	2	3
pas	2	2	2	1	2
vus	3	3	3	2	1,3

Tableau 18 : Résultat matriciel du calcul d'édition pondéré

Les tableau 17 et tableau 18 présentent le résultat des différents algorithmes sous forme matricielle. L'alignement final (recherche du moindre cout) est surligné.

4.2.2.2. Calcul de la distance d'édition de Damereau-Levenshtein (opération d'inversion)

Outre les opérations d'identité, d'omission, d'insertion et de suppression, la distance de Damereau-Levenshtein prend également en compte l'opération d'inversion. Un calcul algorithmique permet de prendre en compte cette opération (cf. Équation 3). Cette opération permet d'associer un cout à la comparaison d'un segment *n* (« " » dans la transcription, Figure 63) et d'une forme *m+1* (*dit* dans la normalisation, Figure 63), ainsi qu'à la comparaison d'une forme *m* (" dans la normalisation, Figure 63) et d'un segment *n+1* (« dit » dans la transcription, Figure 63).

$$d(i+1,j+1) = \min \begin{pmatrix} d(i+1,j)+1, \\ d(i,j+1)+1, \\ d(i,j)+cost(s_i,t_j) \\ d(i-1,i-1)+cost_2(s_i,t_j) \end{pmatrix}, cost(a,b) = \begin{Bmatrix} 0\ si\ a=b \\ >0\ et<1\ si\ a\sim b \\ 1\ sinon \end{Bmatrix}$$

- $, cost_2(a_i,b_j) = \begin{Bmatrix} 0\ si\ cost(a_i,b_{j+1})=0\ et\ cost(a_{i+1},b_j)=0 \\ <1\ si\ cost(a_i,b_{j+1})<1\ et\ cost(a_{i+1},b_j)<1 \\ 1\ sinon \end{Bmatrix}$

Équation 3 : Calcul de la distance de Damereau-Levenshtein avec poids

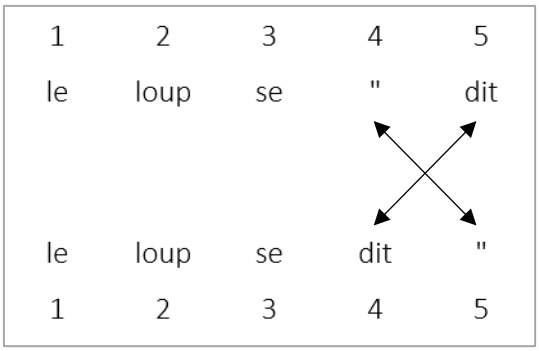


Figure 63: Exemple de recherche d'inversion

	le	loup	se	dit	"
le	0	1	2	3	4
loup	1	0	1	2	3
se	2	1	0	1	2
"	3	2	1	1	1
dit	4	3	2	1	1

Tableau 19 : Résultat matriciel du calcul de la distance d'édition de Damereau-Levenshtein

4.2.2.3. Prise en compte des phénomènes d'hyposegmentation et d'hypersegmentation

Les algorithmes classiques de calcul de distance permettent de faire des comparaisons caractère par caractère ou token par token mais ne prennent pas en compte les problèmes de segmentation. Deux étapes de comparaison ont donc été ajoutées pour prendre en compte ces spécificités des corpus scolaires. Ces étapes permettent d'accorder un poids différencié aux comparaison entre plusieurs tokens (« J' » et « uste » avec *Juste*, Figure 64).

$$d(i+1,j+1) = \min \begin{pmatrix} d(i+1,j)+1, \\ d(i,j+1)+1, \\ d(i,j)+cost(s_i,t_j) \\ d(i,j)+cost(t_j,t_j.t_{j+1}.\text{etc}..t_{j+n}) \\ d(i,j)+cost(s_i.s_{i+1}.\text{etc}..s_{i+n},t_j) \end{pmatrix}, cost(a,b)$$

$$= \begin{Bmatrix} 0\ si\ a=b \\ >0\ et<1\ si\ a\sim b \\ 1\ sinon \end{Bmatrix}$$

Équation 4 : Calcul des phénomènes d'hypersegmentation et d'hyposegmentation

L'équation ci-dessus reprend à la fois le calcul de la distance de Levenshtein classique et le calcul de la distance entre des tokens agglutinés (donnés par $t_j, t_j . t_{j+1} . \text{etc.} . t_{j+n}$ et $s_i . s_{i+1} . \text{etc.} . s_{i+n}, t_j$). Ce poids est toujours supérieur à une comparaison réussie entre des tokens non agglutinés, de sorte que l'aligneur priviliégie l'alignement des tokens uns à uns plutôt que d'un token avec des tokens agglutinés.

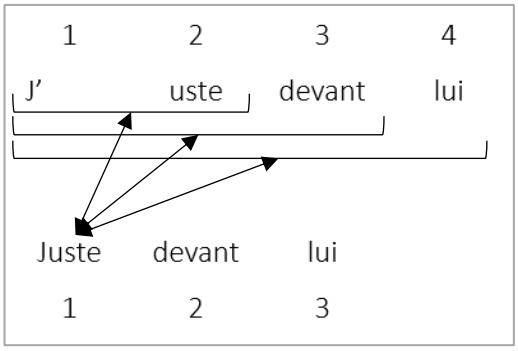


Figure 64 : Exemple de recherche d'hypersegmentation

Dans l'exemple donné à la figure 64, l'algorithme compare d'abord la forme *Juste* avec les segments « J' », « uste », « devant » et « lui ». Puis, il réalise une comparaison de la forme *Juste* avec les segments « J'uste », puis « J'ustedevant » et enfin « J'ustedevantlui ». Ces comparaisons lui permettent de reconnaitre une correspondance phonétique entre la forme *Juste* et les segments agglutinés « J'uste », un poids en conséquence est attribué dans la matrice de parcours.

	Juste	devant	lui
J'	0,6	1,6	2,6
uste	0,6	1,6	2,6
devant	1,6	0,6	1,6
lui	2,6	1,6	0,6

Tableau 20 : Résultat matriciel du calcul de la distance d'édition avec recherche d'hypersegmentation

5. Conclusion

L'approche par comparaison que nous adoptons pour l'exploration du corpus *Scoledit* nécessite l'élaboration de différents outils, à commencer par un algorithme d'alignement des formes normalisées et des segments transcrits. Pour réaliser cet alignement, nous appuyons sur différents indices, principalement graphiques et phonologiques. De plus, deux parcours algorithmiques différents ont été élaborés. L'ensemble de ces éléments ont fait l'objet d'une évaluation, détaillée dans le chapitre suivant.

Chapitre 9 - Évaluation de l'aligneur *AliScol*

1. Élaboration des données de référence 181
2. Mesures d'évaluation 182
3. Evaluation d'*AliScol* 184
4. Conclusion 196

Différentes évaluations existent selon le degré de conception d'un produit (Hirschman & Mani, 2003 ; Ozdowska, 2007), que celui-ci soit au stade de recherche, au stade de prototypage, ou au stade de commercialisation. Ce qui intéresse ici, c'est l'évaluation du premier stade, le stade de recherche au cours duquel sont évalués les performances du système en tant que technologie et non en tant qu'application opérationnelle destinée à des utilisateurs (Chaudiron, 2004, cité dans Ozdowska, 2007).

Pour réaliser l'évaluation de systèmes d'alignement, la méthode la plus fréquente est de comparer les sorties du système à évaluer avec des données de référence, aussi appelées corpus-étalon (*golden standard* en anglais). Ces données de référence sont construites manuellement par des humains et correspondent aux données que le système à évaluer devrait idéalement produire.

1. Élaboration des données de référence

Il existe principalement deux méthodes pour élaborer ces données (Marcus et al., 1993). La première méthode, que nous appelons *méthode manuelle*, consiste à effectuer manuellement la tâche que l'on cherche à évaluer à partir de données non transformées. La seconde, que nous appelons *méthode par correction*, consiste à effectuer automatiquement la tâche que l'on cherche à évaluer puis à réaliser une correction manuelle de ces données.

1.1. Méthode manuelle

Cette méthode a été utilisée dans plusieurs projets, comme dans le projet *BLINKER* (Melamed, 1998a et 1998) qui a permis la constitution d'un corpus parallèle de référence pour l'anglais et le français. La méthodologie adoptée était la suivante. En premier lieu les concepteurs du projet ont rédigé un guide d'annotation, une annotation étant un lien entre un mot anglais et sa traduction française. Puis à l'aide du logiciel *BLINKER*, une interface d'assistance à l'annotation spécialement conçue pour le projet, deux équipes de trois à quatre annotateurs ont annoté manuellement une partie du corpus. Une comparaison automatique de leur annotation a

permis de faire émerger les différences. Après analyse manuelle et discussion entre les annotateurs, le choix des meilleures annotations a été fait et le guide a été adapté. La suite du corpus a été annoté selon le même procédé. Enfin, un accord inter-annotateurs a été calculé entre les différentes annotations proposées.

Faisant suite à ce projet, F. Och et H. Ney (2000) reprennent le constat (Melamed, 1998b) que réaliser un alignement automatique n'est pas toujours aisé et représente souvent une tâche ambigüe. Ils proposent une méthode quelque peu alternative dans laquelle, en plus de réaliser un alignement, les annotateurs étiquettent ces alignements selon qu'ils sont S(ûrs) ou P(robables). Les différences d'alignement des annotateurs leur sont présentées et il leur est demandé d'y apporter une correction dans la mesure du possible. Aucun accord inter-annotateurs n'est ensuite calculé, mais les alignements jugés sûrs par les deux annotateurs sont repris et sont étiquetés comme S(ûrs), les autres alignements sont étiquetés comme P(robables).

1.2. Méthode par correction

Cette méthode nécessite d'effectuer automatiquement la tâche au préalable à l'aide de l'outil développé (Ozdowska, 2007 ; Ahrenberg, Andersson, & Merkel, 2000) ou à l'aide d'un outil déjà existant (Marcus *et al.*, 1993 ; Véronis, 2000a). Idéalement cette correction est réalisée par plusieurs annotateurs et un accord inter-annotateurs est calculé.

L'étude réalisée par Marcus *et al.* (1993) montre que cette deuxième méthode serait à la fois plus fiable, plus rapide et générerait un taux de désaccord plus faible entre les annotateurs que la première méthode. Cependant, dans leur étude, les annotateurs sont plus entrainés qu'à la deuxième tâche car la première. Melamed (1998b) met également en avant la nécessité de disposer d'un système suffisamment efficace pour obtenir une bonne approximation pour pouvoir utiliser cette méthode.

2. Mesures d'évaluation

Une fois les données de référence établies, il est alors possible de les comparer aux données produites par le système à évaluer. Il y a donc généralement plusieurs mesures d'évaluation à produire :

- L'accord inter-annotateurs qui permet d'évaluer la fiabilité des données de référence ;
- Le score d'évaluation qui mesure la qualité du système d'alignement que l'on souhaite évaluer.

2.1. Accord inter-annotateurs

Dans tout processus d'annotation ou d'alignement manuel, il y a une part de subjectivité (Ozdowska, 2007). En effet, ce processus est soumis à une certaine part d'interprétation, qui peut différer selon les annotateurs et annotatrices. Le calcul de l'accord inter-annotateurs (aussi appelé accord interjuge ou accord intersubjectif) permet d'estimer le taux de subjectivité dans les données de référence élaborées. Il permet donc d'évaluer la fiabilité de ces données et de s'assurer que la méthode peut être répliquée (Melamed, 1998b).

Il existe plusieurs méthodes pour calculer l'accord inter-annotateurs. La méthode la plus simple est celle proposée pour le projet *ARCADE* (Véronis & Langlais, 2000). L'accord inter-annotateurs y est calculé à partir d'un indice de Dice[92], selon la formule suivante :

$$\text{Accord} = 2 \times \frac{\text{Nombre d'alignements en commun}}{\text{Nombre total d'alignements (pour les deux annotateurs)}}$$

Une autre méthode a été choisie pour élaborer le *NUS Corpus of Learner English* (Dahlmeier et al., 2013). Celle-ci repose sur le calcul du kappa de Cohen[93] (cité également par Artstein & Poesio, 2008). Ce calcul prend en compte la possibilité que deux annotateurs soient en accord de manière fortuite et prend à la fois en paramètre la probabilité d'un accord entre les deux annotateurs et la probabilité d'un accord dû au hasard.

Dans leurs projets respectifs, F. Och et H. Ney (2000), R. Mihalcea et T. Pedersen (2003) choisissent de produire une version de référence à partir des alignements communs et étiquetés comme sûrs par les deux annotateurs (cf. 1.1), ils choisissent donc de ne pas calculer d'accord inter-annotateurs mais de tirer parti des points d'accords entre les annotateurs.

2.2. Score d'évaluation

Selon O. Kraif (2001), depuis les travaux d'évaluation menés au sein du projet *ARCADE* (Véronis & Langlais, 2000), un relatif consensus s'est établi autour des mesures de **précision** et de **rappel**. Ces deux mesures sont souvent combinées au sein de la F-mesure.

Habituellement, le rappel est défini comme suit :

$$\text{Rappel} = \frac{Nombre\ d'alignements\ corrects}{Nombre\ d'alignements\ de\ référence}$$

[92] Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3), 297-302.
[93] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1), 37-46.

Il indique la proportion d'alignements corrects identifiés parmi la totalité des alignements à identifier. Cette notion est souvent associée à la notion de silence qui représente le nombre d'alignements qui auraient dû être reconnus mais qui ne l'ont pas été.

La précision est généralement définie comme suit :

$$\text{Précision} = \frac{Nombre\ d'alignements\ corrects}{Nombre\ d'alignements\ proposés}$$

Elle indique la proportion d'alignements corrects parmi les alignements identifiés. Cette notion est souvent associée à la notion de bruit qui représente le nombre d'alignements identifiés mais qui n'auraient pas dû l'être.

Généralement, plus la précision d'un système croît, plus son rappel diminue, et inversement (Kraif, 2001). La F-mesure, proposée par van Rijsbergen[94] pour l'évaluation de systèmes de recherche d'information, permet d'englober ces deux mesures dans un même calcul afin de déterminer une mesure d'efficacité globale (Véronis & Langlais, 2000). Elle est calculée à partir d'un indice de Dice selon la formule suivante:

$$\text{F-mesure} = 2 \times \frac{précision \times rappel}{précision + rappel} = 2 \times \frac{Nombre\ d'alignements\ en\ commun}{Nombre\ total\ d'alignements\ (référence + système\ à\ évaluer)}$$

Notons cependant que réaliser de tels travaux d'évaluation est couteux car la constitution de données de référence est très chronophage (part de travail manuel non négligeable). Grâce aux projets de grande envergure, tels que les projets *BLINKER* et *ARCADE* et *ARCADE 2*, des corpus de référence sont désormais disponibles pour l'évaluation de certaines tâches d'alignement, comme l'alignement français – anglais. Cependant, lorsque ces données ne sont pas accessibles, la tâche de constitution des données de référence fait partie intégrante du processus d'évaluation. Dans le cadre des corpus scolaires, il n'existe pas de tels corpus de référence. Nous avons donc dû élaborer nos propres données et notre propre méthodologie d'évaluation.

3. Evaluation d'*AliScol*

Nous avons adopté une méthode par correction des sorties de l'aligneur, similaire à l'évaluation proposée dans le projet *ALIBI* (Ozdowska, 2007). Pour commencer, nous avons choisi un échantillon de productions parmi les productions de notre corpus, que nous avons appelé *corpus de référence*. Il aurait sans doute été plus pertinent de choisir des données non issues de notre corpus pour cela, afin de tester notre système sur des données proches mais non

[94] van Rijsbergen, C. J. (1979). *Information Retrieval*. 2nd edition, London: Butterworths.

identiques à celles utilisées pour son élaboration. Néanmoins, comme nous l'avons vu au chapitre 3, il existe encore peu de corpus scolaires, c'est-à-dire comparables au nôtre, numérisés et disponibles, particulièrement pour les premières années d'école primaire. Nous avons donc pris soin, de ne pas utiliser les mêmes productions lors de la phase d'élaboration du système et lors de la phase de test. Les premières sont rassemblées dans le corpus de travail tandis que les secondes sont rassemblées dans le corpus de référence.

3.1. Composition du corpus de référence

Le corpus de référence est un sous-corpus élaboré dans le but d'évaluer l'outil conçu. À nouveau, nous avons fait le choix d'élaborer un corpus longitudinal. Le corpus de référence se compose de quatre productions écrites pour 50 élèves du CP au CM1, soit 200 productions.

Ce sous-corpus n'intervient qu'à la fin du processus. Le recueil et la numérisation du corpus ayant évolué entre le début et la fin du développement de l'aligneur, des productions de CM1 ont pu être incluses dans ce corpus. En revanche, les productions de CM2 n'étaient pas encore numérisées et n'ont pu être prises en compte dans le corpus de référence.

Les identifiants des élèves ont été sélectionnés aléatoirement parmi les élèves du corpus longitudinal. Le corpus de référence se compose de la façon suivante :

Niveau scolaire	*Nombre de productions*	*Nombre de tokens*
CP	50	1 573
CE1	50	4 547
CE2	50	7 451
CM1	50	8 889
Total	200	22 460

Tableau 21 : Caractéristiques du corpus de référence

Une fois les productions du corpus de référence choisies, il a été aligné automatiquement, à l'aide du système que nous avons développé et que nous avons présenté précédemment : *AliScol*. Là encore, pour des raisons d'objectivité, il aurait sans doute été préférable d'utiliser un autre système d'alignement, mais le domaine des corpus scolaires ayant encore été peu étudié en TAL, nous n'en avions pas à disposition.

Une fois l'alignement automatique calculé, les sorties du système ont été corrigées manuellement. Cette correction s'est faite par quatre annotateurs, répartis en deux équipes de deux. Ils avaient à leur disposition un guide d'alignement (disponible en annexe 10) construit au fur et à mesure avec eux, selon les difficultés rencontrées.

La première équipe s'est réparti le corpus de référence et a corrigé une première fois la sortie de notre système. Il n'a pas été possible de leur faire corriger à tous les deux l'ensemble du corpus de référence. Néanmoins nous leur avons fait corriger au préalable un échantillon du corpus de référence afin d'approximer l'accord inter-annotateurs. Puis, nous les avons fait travailler de façon conjointe afin qu'ils puissent se concerter en cas de doute.

Quelques mois plus tard, une seconde équipe d'annotateurs a travaillé à la constitution de ces données de référence. Il était prévu initialement de réaliser une deuxième correction de la sortie de l'aligneur afin de procéder avec une méthodologie similaire au projet *ARCADE* (Véronis & Langlais, 2000). Malheureusement, les résultats de la première phase de correction n'étaient pas suffisamment bons pour être exploitables. Ce constat peut s'expliquer par différents facteurs. En premier lieu, la première équipe d'annotateurs a travaillé avant la fin du développement de l'aligneur, les sorties provisoires n'étaient peut-être pas de qualité suffisante pour être utilisées dans le cadre d'une méthode par correction. De même, le guide de normalisation a également évolué après ce premier travail, modifiant ainsi les données d'entrée. Enfin, la tâche d'alignement était peut-être trop dense et pas suffisamment comprise par les annotateurs. Nous avons cherché à évaluer l'ensemble de la sortie de l'aligneur (alignement, étiquettes morphosyntaxiques, types d'erreur, etc.), ce qui représentait sans doute une tâche trop conséquente.

Pour toutes ces raisons, nous avons fourni à la deuxième équipe une version actualisée de la sortie de l'aligneur. De plus, les différents types de données à évaluer ont été segmentées en fichiers séparés (un fichier pour l'alignement, un fichier pour les étiquettes d'erreurs et un fichier pour les étiquettes morphosyntaxiques), ce qui permet une tâche de correction moins dense. Comme précédemment, après un premier échantillon commun, les données ont été réparties entre les deux membres de l'équipe. Puis, l'une et l'autre se sont mutuellement corrigées.

Cette méthodologie est moins rigoureuse que celle que nous aurions voulu adopter initialement et ne permet pas vérifier la réplicabilité du guide d'alignement manuel. Une campagne d'évaluation ultérieure sera donc nécessaire pour une évaluation plus précise de l'algorithme *AliScol*.

Le tableau 21 inventorie le taux de différence entre la première correction, réalisée par la première équipe, et la deuxième correction, réalisée par la deuxième équipe.

Niveau	*Différences entre les deux phases de correction*
CP	7,6 %
CE1	6,8 %
CE2	6,1 %
CM1	3,6 %
CP-CM1	*5,4 %*

Tableau 22 : Évaluation des différences entre la 1ère correction et la 2ème correction

À la lecture de ce tableau, il apparait que les différences entre les deux corrections proposées sont importantes. Nous faisons l'hypothèse que celles-ci s'expliquent principalement par la différence entre les fichiers initiaux, fournis dans des formats divergents et issus de deux versions différentes de l'algorithme.

3.2. Comparaisons des différentes versions de l'aligneur

Comme nous l'avons présenté au chapitre précédent, deux versions de l'aligneur ont été développées (une version séquentielle et une version matricielle). Chacune de ces versions doit être évaluée.

3.2.1. Méthode séquentielle

L'algorithme propose de s'appuyer sur différents indices de comparaison (cf. Chapitre 8 - 4.1) et différents modules de comparaison (cf. Chapitre 8 - 4.2.1). Chacune de ces alternatives est également évaluée. Le tableau 22 recense les différents éléments évalués, au fur et à mesure de leur ajout dans l'architecture de l'algorithme. Le tableau 23 présente les résultats de l'évaluation de chacun de ces niveaux d'élaboration. Les mesures de précision, de rappel et la F-mesure, présentées précédemment, ont été utilisées.

Niveau	*Type de comparaison*	*Exemple*
Niveau 1	Comparaison graphique stricte	et || et
Niveau 2	Niveau 1 + Comparaison phonologique et archiphonologique strictes	miole || miaule
Niveau 3	Niveau 2 + Comparaisons graphiques, phonologiques et archiphonologiques relatives	crocette || croquettes
Niveau 4-A	Niveau 3 + recherche des hypersegmentations et des hyposegmentations	sais || c'est a prés || après
Niveau 4-B	Niveau 3 + recherche des hyposegmentations et des hypersegmentations	
Niveau 5-A	Niveau 4-A + recherche des insertions et des omissions	
Niveau 5-B	Niveau 4-A + recherche des omissions et des insertions	|| n' meilleur ||
Niveau 5-C	Niveau 5-A, mais recherche des omissions et des insertions avant les hyposegmentations et les hypersegmentations	
Niveau 6	Niveau 5-A + recherche des insertions / omissions / hyposegmentations / hypersegmentations combinés	Il apa vu || Il n'a pas vu

Tableau 23 : Différents niveaux de l' algorithme séquentiel évalués

Modes de comparaison	*Rappel*	*Précision*	*F-Mesure*
Niveau 1	22,0 %	19,5 %	20,7 %
Niveau 2	22,0 %	19,6 %	20,7 %
Niveau 3	21,7 %	19,1 %	20,3 %
Niveau 4-A	23,0 %	20,4 %	21,6 %
Niveau 4-B	22,9 %	20,4 %	21,5 %
Niveau 5-A	91,2 %	89,1 %	90,1 %
Niveau 5-B	87,9 %	84,8 %	86,3 %
Niveau 5-C	91,1 %	88,3 %	89,7 %
Niveau 6	91,5 %	89,4 %	90,4 %

Tableau 24 : Résultats de l'évaluation de chaque mode de comparaison

Alors que prendre en compte les cas d'hypersegmentation et d'hyposegmentation n'apporte qu'une progression relative de l'algorithme, l'ajout de la recherche des segments omis et insérés apportent un bond qualitatif certain (de 21,5 % à 90,1 %). Ce résultat ne signifie pas que les phénomènes d'omission ou d'insertion de mots sont nombreux, mais un tel calcul permet à la fois de prendre en compte les balises ajoutées dans la normalisation

(<segmentation/> par exemple) et de gérer les cas d'hypersegmentation et d'hyposegmentation qui n'ont pu être détectés précédemment. L'algorithme échoue par exemple à reconnaitre les segments « tu ne » (Tableau 25), comme des segments hypersegmentés de la forme normée *une*, ce qui peut générer de nombreuses erreurs dans la suite de l'algorithme (Tableau 25). À partir du niveau 5 (Tableau 26), « tu » peux être identifié comme un segment ajouté pour éviter de répercuter des erreurs.

Forme normée	*Forme transcrite*	*Type d'erreur ou de réussite*
Il	Il	normé
était	êtais	archiphonologie normée
une	tu	hypersegmentation
une	ne	hypersegmentation
fois	foit	phonologie normée

Tableau 25 : Alignement d'un extrait de la production 93-CE1 (résultat attendu)

Forme normée	*Forme transcrite*	*Type d'erreur ou de réussite*
Il	Il	normé
était	êtais	archiphonologie normée
une	tu	non normé
fois	ne	non normé
un	foit	non normé
petit	un	non normé

Tableau 26 : Alignement d'un extrait de la production 93-CE1 (sortie de l'aligneur niveau 4)

Forme normée	*Forme transcrite*	*Type d'erreur ou de réussite*
Il	Il	normé
était	êtais	archiphonologie normée
	tu	inséré
une	ne	approximation graphique
fois	foit	phonologie normée

Tableau 27 : Alignement d'un extrait de la production 93-CE1 (sortie de l'aligneur niveau 5)

L'évaluation montre également que rechercher en même temps des phénomènes d'hyposegmentation et d'hypersegmentation et des phénomènes d'omission et d'insertion augmente les performances de l'algorithme. Le tableau 27 présente un exemple de phénomène mieux pris en compte dans le niveau 6 de l'algorithme.

Forme normée	*Forme transcrite*	*Type d'erreur ou de réussite*
plus	plu	Phonologie normée
,		Omission
il	iltonbe	Hyposegmentation
tombe	iltonbe	Hyposegmentation
des	de	Approximation graphique

Tableau 28 : Alignement d'un extrait de la production 93-CP (sortie de l'aligneur niveau 6)

L'évaluation de la méthode séquentielle permet donc d'établir l'ordre d'exécution des modules le plus pertinent :

1. Recherche des phénomènes d'hyposegmentation ;
2. Recherche des phénomènes d'hypersegmentation ;
3. Recherche des phénomènes d'omission ;
4. Recherche des phénomènes d'insertion ;
5. Recherche des phénomènes combinés d'omission et d'hyposegmentation ou d'hypersegmentation ;
6. Recherche des phénomènes combinés d'insertion et d'hyposegmentation ou d'hypersegmentation.

L'évaluation des niveaux 1 à 3 tendent à montrer qu'utiliser les différents modes de comparaison ne semble pas faire beaucoup de différences. Nous avons voulu vérifier ce fait dans la version finale de l'algorithme. Nous avons donc élaboré un 7e niveau, similaire au 6e, mais qui n'inclut que la comparaison graphique stricte, excluant les comparaisons phonologiques et archiphonologiques et les comparaisons relatives. L'évaluation de ce niveau (Tableau 28) semble confirmer ce constat.

Modes de comparaison	*Rappel*	*Précision*	*F-Mesure*
Niveau 6	91,5 %	89,4 %	90,4 %
Niveau 7	90,49 %	88,30 %	89,4 %

Tableau 29 : Résultats de l'évaluation du niveau 7

Pour développer cet algorithme, nous nous sommes appuyée sur des fenêtres glissantes qui permettent de comparer plusieurs segments les uns avec les autres. Nous avons également évalué l'empan de ces fenêtres afin de déterminer les tailles de fenêtre idéales.

Deux types de fenêtres ont été évaluées : la fenêtre utilisée pour détecter les cas d'hyposegmentation et d'hypersegmentation, appelée fenêtre de segmentation, et la fenêtre utilisée pour détecter les cas d'omission et d'insertion, appelée fenêtre de tokens. Par défaut, lors du développement de l'algorithme des fenêtres de taille 3 ont été utilisées.

Dans un premier temps, les variations de taille de la fenêtre de segmentation ont été évaluées, de 2 à 5. Les résultats sont reportés dans le tableau 29. La taille optimale semble être de 4 tokens, même si les différences semblent peu significatives. Puis, les variations de taille de la fenêtre de tokens ont été évaluées. La taille semblant être optimale à 3, seules les tailles 2 à 4 ont été évaluées.

Fenêtre de segmentation	*fenêtre de tokens*	*Rappel*	*Précision*	*F-Mesure*
2	3	91,3 %	89,1 %	90,2 %
3	3	91,5 %	89,4 %	90,4 %
4	3	91,6 %	89,4 %	90,5 %
5	3	91,6 %	89,4 %	90,5 %

Tableau 30 : Évaluation des variations de longueur de la fenêtre de segmentation

Fenêtre de segmentation	*Fenêtre de tokens*	*Rappel*	*Précision*	*F-Mesure*
4	2	91,5 %	89,5 %	90,5 %
4	3	91,6 %	89,4 %	90,5 %
4	4	91,2 %	88,1 %	89,6 %

Tableau 31 : Évaluation des variations de longueur de la fenêtre de tokens

Il semble qu'une taille de fenêtre de tokens de 3 soit la plus adaptée, mais il est possible que la variation entre les tailles 2 et 3 soit uniquement due à l'échantillon utilisé pour l'évaluation et ne reflète pas un réel gain.

3.2.2. Méthode matricielle

Comme pour l'approche séquentielle, l'approche matricielle repose sur différents modes de comparaison, que nous avons évalués. Comme précédemment, nous avons défini différents niveaux d'élaboration de l'algorithme (Tableau 31). Notons que les phénomènes d'insertion et d'omission sont recherchés par défaut dans l'algorithme de la distance d'édition de

Levenshtein. De plus, le parcours étant matriciel, l'ordre d'exécution de ces modules n'a pas d'influence sur le résultat.

La méthode matricielle a également permis de rechercher un phénomène que nous n'avions pas inclus dans la méthode séquentielle, la recherche des inversions grâce à l'algorithme de distance de Damereau-Levenshtein (niveau 5). Les résultats de l'évaluation de ces niveaux sont disponibles dans le tableau 32.

Niveau	*Type de comparaison*	*Exemple*
Niveau 1	Comparaison graphique stricte	et || et
Niveau 2	Niveau 1 + Comparaison phonologique et archiphonologique strictes	miole || miaule
Niveau 3	Niveau 2 + Comparaisons graphiques, phonologiques et archiphonologiques relatives	crocette || croquettes
Niveau 4	Niveau 3 + recherche des hypersegmentations et des hyposegmentations	sais || c'est a prés || après
Niveau 5	Niveau 4 + recherche des inversions (distance de Damereau-Levenshtein)	amanite-tue mouche || amanite tue-mouches

Tableau 32 : Différents niveaux de l'algorithme séquentiel évalués

Modes de comparaison	*Rappel*	*Précision*	*F-Mesure*
Niveau 1	13,8 %	13,3 %	13,6 %
Niveau 2	13,8 %	13,3 %	13,6 %
Niveau 3	95,2 %	94,9 %	95,0 %
Niveau 4	97,1 %	96,5 %	96,8 %
Niveau 5	97,0 %	96,5 %	96,8 %

Tableau 33 : Résultats de l'évaluation de chaque mode de comparaison

Contrairement à la méthode séquentielle, l'ajout des comparaisons relatives créé un bond qualitatif de l'algorithme. Mais rappelons que la méthode matricielle inclut la prise en compte des phénomènes d'omission et d'insertion. Il semble donc intéressant de réévaluer la méthode séquentielle pour savoir si dans cette méthode les comparaisons relatives ont autant d'influence lorsque les phénomènes d'insertion et d'omission sont pris en compte. Un niveau 5-D a donc été mis en place, similaire au niveau 5-A (Tableau 22) mais ne permettant que des comparaisons graphiques strictes. Les résultats de l'évaluation sont reportés dans le Tableau 33.

Modes de comparaison	*Rappel*	*Précision*	*F-Mesure*
Niveau 5-A	91,2 %	89,1 %	90,1 %
Niveau 5-D	90,5 %	88,2 %	89,3 %

Tableau 34 : Résultats de l'évaluation du mode de comparaison (niveau 5-A et 5-D)

Au vu des résultats, les modes de comparaison n'ont pas le même impact dans la méthode séquentielle.

Le tableau 32 montre également le peu d'apport de la prise en compte des phénomènes d'inversion, via la distance de Damereau-Levenshtein. Ce constat peut s'expliquer par le peu de phénomènes d'inversion présents dans le corpus.

Contrairement à la procédure séquentielle, il n'est nul besoin d'une fenêtre pour les erreurs d'omission et d'insertion puisque l'ensemble des formes des productions sont comparées les unes avec les autres. En revanche, la recherche des cas d'hypersegmentation et d'hyposegmentation a dû être limitée à quelques segments, ce que nous appelons fenêtre d'hyposegmentation et fenêtre d'hypersegmentation, afin de limiter le temps d'exécution et le besoin de mémoire de l'algorithme. Les deux phénomènes étant traités indépendamment, il peut leur être attribué des fenêtres de taille différente, qui peuvent être évaluées de manière indépendante.

Par défaut, nous avions choisi d'utiliser des fenêtres de taille 4. Ce sont ces tailles de fenêtres qui ont été utilisées pour effectuer les évaluations comprises dans le tableau précédent.

Fenêtre d'hyposegmentation	*fenêtre d'hypersegmentation*	*Rappel*	*Précision*	*F-Mesure*
2	4	97,0 %	96,5 %	96,7 %
3	4	97,1 %	96,5 %	96,8 %
4	4	97,1 %	96,5 %	96,8 %
5	4	91,6 %	96,5 %	94,0 %

Tableau 35 : Évaluation des variations de longueur d'hyposegmentation

Fenêtre d'hyposegmentation	*fenêtre d'hypersegmentation*	*Rappel*	*Précision*	*F-Mesure*
4	2	97,1 %	96,5 %	96,8 %
4	3	97,1 %	96,5 %	96,8 %
4	4	97,1 %	96,5 %	96,8 %
4	5	97,1 %	96,5 %	96,8 %

Tableau 36 : Évaluation des variations de longueur d'hypersegmentation

Le tableau 34 montre qu'une fenêtre de taille 3 est optimale pour la gestion des cas d'hyposegmentation, tandis que le tableau 35 qu'une fenêtre de taille 2 suffit pour la gestion des cas d'hypersegmentation. Dans la mesure où augmenter la taille des fenêtres n'augmente pas les performances de l'algorithme, nous privilégions les fenêtres les plus courtes afin de limiter le temps d'exécution.

3.2.3. Evaluation de l'algorithme par niveau scolaire

Les versions optimales de l'algorithme semblent donc être les versions 4 et 5 de la méthode matricielle (niveau 4 et niveau 5, tableau 31), c'est-à-dire les versions basées sur un algorithme de distance de Levenshtein, incluant des modules de gestion d'hypersegmentation de taille 2 et d'hyposegmentation de taille 3 et, pour la version 5, un calcul de distance de Damereau (gestion des inversions). Ces deux versions sont basées sur des comparaisons graphiques et phonologiques strictes et relatives. Ces deux versions produisent un rappel de 97,1 % et une précision de 96,5 %, soit une F-mesure de 96,8 %.

À partir de la version 4, les performances de l'algorithme, niveau par niveau, ont été évaluées. Ces résultats sont répertoriés dans le tableau 36.

Niveau scolaire	*Rappel*	*Précision*	*F-Mesure*
CP	96,1 %	96,1 %	96,1 %
CE1	96,9 %	96,1 %	96,5 %
CE2	96,9 %	96,3 %	96,6 %
CM1	97,6 %	97,1 %	97,4 %
CP-CM1	97,1 %	96,5 %	96,8 %

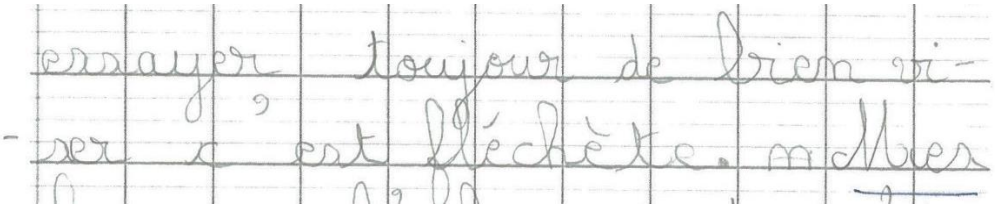
Tableau 37 : Résultats de l'évaluation par niveau scolaire

Plus le niveau scolaire augmente, plus les performances de l'algorithme augmentent. Ce fait semble concomitant avec la diminution des erreurs dans les productions (estimée de manière intuitive à partir de l'observation de productions).

3.3. Observations des erreurs de l'aligneur *AliScol*

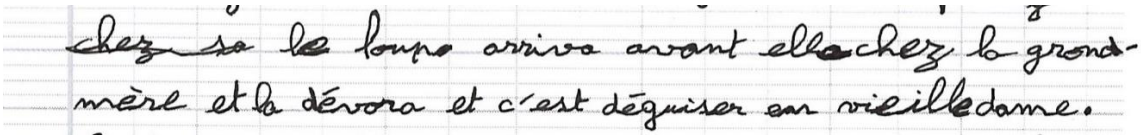
Une observation des erreurs produites par l'aligneur *AliScol* permet de distinguer différentes catégories principales d'erreurs :

- **Les erreurs issues de l'étape de transcription**. Ces erreurs sont issues d'erreurs commises lors de l'étape de transcription. Par exemple, l'alignement échoue régulièrement à distinguer les segments uniques tronqués par une fin de ligne (« viser », Figure 66) des formes composées séparées par une fin de ligne (« grand-mère », Figure 67). Une différenciation de ces deux cas de figures a été mise en place dans la transcription, mais celle-ci n'est pas toujours respectée et entraine des erreurs d'alignement.



Transcription :
« [...] essayer toujour de bien vi -/- ser c'est fléchète. [...] »

Figure 66 : Extrait de la production 94-CE1



Transcription :
« [...] <revision/>le loup arriva avant elle chez la grand-/ mère et la dévora et c'est déguiser en vieille dame. [...] »

Figure 67 : Extrait de la production 1160-CE2

- **Les erreurs issues de l'étape de normalisation**. Il peut arriver que des erreurs soient commises lors de la normalisation, mais il arrive également que l'algorithme échoue à prendre en compte certaines balises issues de la version normalisée, par exemple les balises <meta></meta>, ajoutée après le développement de l'aligneur. Une fois repérées, ces erreurs peuvent être facilement corrigées.
- **Les erreurs issues de l'étape de tokenisation**. Généralement, les systèmes de tokenisation prennent en compte un certain nombre d'expressions polylexicales et de formes composées, comme *c'est-à-dire*, *d'accord* ou encore *grand-mère*, qui sont considérées comme un seul token Néanmoins, en présence d'erreurs, ces formes ne peuvent plus être identifiées et sont donc considérées comme plusieurs tokens, ce qui génère des problème d'alignement (Tableau 37).

Forme normée	*Segment transcrit*	*Catégorie d'erreur*
grand-mère	grande	Hypersegmentation
grand-mère	-	Hypersegmentation
grand-mère	mère	Hypersegmentation

Tableau 38 : Exemple d'alignement (extrait de la production 772-CE2)

- Les erreurs issues de l'étape d'alignement. L'aligneur lui-même génère des erreurs. Elles se produisent généralement lorsqu'une combinaison d'erreurs apparait, par exemple une combinaison d'erreur orthographique et d'hypersegmentation (Tableau 38).

Sortie de l'aligneur		*Alignement attendu*	
Et	Et	Et	Et
bien	bein	bien	bein
en		en	alui
lui	alui	lui	alui
faisant	faisen	faisant	faisen
peur	peure	peur	peure

Tableau 39 : Exemple d'alignement (extrait de la production 93-CE1)

4. Conclusion

Une première campagne d'évaluation a permis de mesurer les performances des différentes versions de l'algorithme. Cette étude a permis de comparer l'approche séquentielle et l'approche matricielle et de montrer la plus grande efficacité de cette dernière approche. Au cours de cette étude, la pertinence de la prise en compte des différents phénomènes d'erreurs (hypersegmentation, omission, inversion, par exemple) a été évaluée. L'étude a montré que si prendre en compte les phénomènes d'hypersegmentation et d'hyposegmentation pouvait améliorer quelque peu les performances de l'algorithme, la prise en compte des phénomènes d'inversion ne modifie pas beaucoup ses performances.

Néanmoins, la modification continue des conventions de normalisation n'a pas permis de mener une évaluation aussi rigoureuse que nécessaire. Une deuxième campagne d'évaluation plus conséquente, après stabilisation des conventions, permettrait de pallier ce manque.

Partie 4 - Application de l'approche par comparaison : autres exemples

Chapitre 10 - Comparer formes transcrites et formes normalisées, quelques enseignements 199
Chapitre 11 - Appliquer la méthode par comparaison aux graphèmes 213
Chapitre 12 - Analyser la morphographie verbale grâce à l'approche par comparaison 227

Chapitre 10 - Comparer formes transcrites et formes normalisées, quelques enseignements

1. Erreurs de segmentation en mots 200
2. Comparaison des formes au niveau graphique et phonologique 203
3. Conclusion 210

L'approche par comparaison, que nous avons adoptée pour ce travail sur les corpus d'écrits scolaires (cf. Chapitre 4 - 3), repose sur un alignement des éléments transcrits et des éléments normés. De ce fait, elle nécessite des outils d'alignements. L'algorithme d'alignement *AliScol* présenté précédemment (cf. Chapitre 8) permet d'aligner les segments transcrits et les formes normées. À ce titre, il représente un élément essentiel de l'approche par comparaison appliquée au niveau lexical.

L'application de cette approche au niveau lexical nous permet de nombreuses observations, comme l'étude du lexique, la répartition des catégories syntaxiques utilisées ou encore l'étude des réussites et des erreurs des apprenants. Dans ce chapitre, nous proposons de nous pencher sur ce dernier aspect.

Les erreurs d'hyposegmentation et d'hypersegmentation étant une des spécificités des corpus scolaires, il semble intéressant de les observer brièvement dans un premier temps. Dans un second temps, nous nous pencherons également sur les réussites et les erreurs d'ordre orthographique que permet de décrire l'algorithme *AliScol*. Celui-ci étant basé sur des indices graphiques et phonétiques, il permet de distinguer les erreurs relevant de différences purement graphiques et les erreurs relevant de différences graphiques qui engendrent des modifications phonétiques. En revanche, les erreurs d'omission et d'insertion étant peu présentes dans notre corpus, nous ne nous y attarderons pas dans ce chapitre.

Les données que nous présentons ici reposent sur l'observation d'un échantillon du corpus longitudinal présenté précédemment (cf. Chapitre 5). Ce corpus est composé des 480 productions, dont la transcription et la normalisation ont toutes été vérifiées. Quelques caractéristiques de ce corpus sont données dans le Tableau 39.

Niveau scolaire	*Nombre de productions*	*Nombre de tokens*	*Nombre moyen de formes par production*
CP	120	3 597	30,0
CE1	120	9 754	81,3
CE2	120	15 791	131,6
CM1	120	19 099	159,2
Total	*480*	*48 241*	*100,5*

Tableau 40 : Caractéristiques du corpus de travail de 480 productions

1. Erreurs de segmentation en mots

Que l'élève sache ou non transcrire correctement les correspondances phonographiques, il est possible qu'il ne sache pas correctement la segmenter. Ainsi, il peut être tenté d'agglutiner deux formes entre elles ou, à l'inverse, de tronquer une forme en deux segments distincts. Nous distinguons donc deux types d'erreurs :

- Les erreurs d'hyposegmentation, c'est-à-dire la présence de plusieurs formes agglutinées. Exemple : « il séfémal. » (*il s'est fait mal.*, 1154-CP)
- Les erreurs d'hypersegmentation, c'est-à-dire une forme graphiée en plusieurs segments distincts. Exemple : « il jouer tout le jour en semble » (*ils jouaient tout le jour ensemble*, 1345-CE1)

Parfois, les erreurs d'hyposegmentation et d'hypersegmentation sont mêlées, c'est-à-dire que la forme peut être à la fois agglutinée avec la forme suivante et en même temps divisée en deux segments distincts, comme dans l'exemple : « le loup n'ais pas d'utuou rucé » (*le loup n'est pas du tout rusé*, 1224-CE2). Mais à l'heure actuelle, pour des raisons de complexité algorithmique, l'aligneur *AliScol* ne prend pas encore en compte ce type d'erreurs, nous n'en parlerons donc pas dans ce chapitre[95].

1.1. Recensement des phénomènes d'hypersegmentation et d'hyposegmentation

Les cas d'hyposegmentation et d'hypersegmentation sont reportés dans le tableau 40. Pour chaque niveau et chaque erreur sont donnés :

[95] Ce phénomène mérite tout de même une attention ultérieure. On en observe 26 occurrences dans le corpus de référence, soit 52 tokens concernés (0,23 % des tokens du corpus de référence).

- le nombre d'occurrences du phénomène concerné dans le corpus étudié (nombre absolu de segments) ;
- la fréquence d'apparition du phénomène concerné dans le corpus étudié (en pourcentage de segments).

Les alignements de signes de ponctuation et de balises ont été exclus du nombre total d'alignements.

Type d'erreurs de segmentation	*CP*	*CE1*	*CE2*	*CM1*	*CP-CM1*
Hyposegmentation	92 (3,5 %)	135 (1,8 %)	133 (1,1 %)	149 (1,1 %)	509 (1,4 %)
Hypersegmentation	33 (1,2 %)	61 (0,8 %)	84 (0,7 %)	84 (0,6 %)	262 (0,7 %)
Total d'alignements	2 665	7 379	11 625	14 116	35 795

Tableau 41 : Occurrences des phénomènes d'hyposegmentation et d'hypersegmentation

Une précision s'impose sur la façon de comptabiliser ces phénomènes. Pour évaluer l'alignement produit par l'algorithme *AliScol* nous nous sommes appuyée sur un alignement segment par segment (Tableau 41). Pour compter le nombre de phénomènes d'hyposegmentation et d'hypersegmentation nous nous sommes en revanche appuyée sur un alignement par séquences de segments (Tableau 42), afin de compter le nombre de phénomènes d'hypersegmentation et d'hyposegmentation et non le nombre de segments impliqués dans ces phénomènes. Le nombre total d'alignements peut donc varier quelque peu[96].

Forme normée	*Segment transcrit*	*Étiquette orthographique*	*Segmentation en mots*
Edgard	Edgard	Normé	Normé
eut	eu	Phonologie normée	Normé
tellement	t'	Phonologie normée	Hypersegmentation
tellement	elle	Phonologie normée	Hypersegmentation
tellement	ment	Phonologie normée	Hypersegmentation
peur	peur	Normé	Normé

Nombre d'alignements : 6
Hypersegmentation : 3 (50 %)

Tableau 42 : Exemple d'alignement par segments

[96] Pour plus d'explications au sujet de la différence entre ces deux types d'alignements, il est possible de se référer à l'article de S. Ozdowska (2007, p. 109-112).

Forme(s) normée(s)	*Segment(s) transcrit(s)*	*Étiquette orthographique*	*Segmentation en mots*
Edgard	Edgard	Normé	Normé
eut	eu	Phonologie normée	Normé
tellement	t'elle ment	Phonologie normée	Hypersegmentation
peur	peur	Normé	Normé

Nombre d'alignements : 4

Hypersegmentation : 1 (25 %)

Tableau 43 : Exemple d'alignement par séquences de segments

L'exemple présenté dans le tableau 41 et le tableau 42 inclut trois segments hypersegmentés, comptabilisés comme un seul phénomène d'hypersegmentation.

1.2. L'élision, un facteur d'hyposegmentation

Dans notre système de tokenisation (segmentation en formes), nous avons considéré que les apostrophes entre une forme élidée et une deuxième forme de catégorie différente (exemples : *c'est* et *l'arbre* par opposition à *aujourd'hui*) représentaient une marque de segmentation entre formes. Dans l'exemple *c'est*, nous sommes donc en présence de deux formes : le pronom *c'* et la forme verbale *est*. De ce fait, lorsque l'apostrophe est omise et qu'elle n'est pas remplacée par une espace, nous considérons qu'il s'agit d'une erreur d'hyposegmentation (par exemple, le segment transcrit « cest » équivalent aux deux formes *c'* et *est*). Les erreurs impliquant l'omission d'une apostrophe, comme dans l'exemple que nous venons de donner, sont très nombreuses parmi les erreurs d'hyposegmentation.

Forme élidée concernée	*Nombre d'occurrences d'erreurs*
s'	67
l'	61
qu'	42
c'	38
n'	30
d'	28
t'	16
m'	12
j'	9
Total	*303*

Tableau 44 : Formes élidées fréquemment impliquées dans des erreurs d'hyposegmentation (480 productions)

Le tableau 43 recense les formes élidées les plus fréquemment impliqués dans des erreurs d'hyposegmentation. Le segment *qu'* inclut les élisions des formes *parce que* et *jusque*.

Ce type d'erreurs d'hyposegmentation parait important à prendre en compte car il représente au moins 303 occurrences des 509 erreurs d'hyposegmentation recensées, soit 59,5 % des erreurs d'hyposegmentation.

Plus épisodiquement, on retrouve également des erreurs d'hypersegmentation causées par l'insertion d'une apostrophe (exemple : « l'aissait », *laissait*). Généralement, avec l'introduction d'apostrophes, on retrouve le déterminant *l'* ou des formes pronominales existantes (t', m', c', s', etc.).

2. Comparaison des formes au niveau graphique et phonologique

Le fonctionnement de l'aligneur *AliScol* s'appuie sur une comparaison graphique et phonologique des segments transcrits et des formes normées. Nous pouvons nous appuyer sur le résultat de ces comparaisons pour établir une typologie des réalisations graphiques selon les réalisations orthographiques des apprenants : les segments normés, les segments phonologiquement normés, les segments archiphonologiquement normés et les segments non normés.

2.1. Définition des catégories

2.1.1. Les segments normés

Les segments normés constituent la majeure partie des segments de notre corpus. Ce sont les segments ne comportant aucune erreur orthographique, pour lesquels les élèves ont su écrire correctement l'orthographe et la morphologie. Nous considérons comme segments normés les segments pour lesquels le segment produit par les élèves est identique à la forme attendue sans regard pour les considérations de majuscules. Les erreurs de segmentation en mots étant traitées dans une autre catégorie, des segments peuvent être considérés comme orthographiquement normés alors même qu'ils présentent ce type d'erreurs (hyposegmentation ou hypersegmentation).

2.1.2. Les segments phonologiquement normés

Les segments non normés mais phonologiquement normés, aussi appelés « segments phonologiquement normés », sont les segments pour lesquels l'orthographe n'est pas respectée mais dont la graphie reproduit l'ensemble des phonèmes attendus. Les représentations phonologiques de ces segments et des formes correspondantes sont identiques. Dans ce cas, nous supposons que l'élève n'a pas été en mesure de correctement

orthographier le mot écrit mais qu'il connait les correspondances phonie-graphie (entre sons et lettres ou groupes de lettres). Il est donc capable de transcrire les sons de l'oral à l'écrit et, en cas de difficulté, la lecture à haute voix de la production de l'élève permet de lever une grande partie des ambigüités.

2.1.3. Les formes archiphonologiquement normée

Les segments non normés mais archiphonologiquement normés, aussi appelés segments archiphonologiquement normés, sont les segments pour lesquels l'orthographe n'est pas respectée mais dont la graphie reproduit l'ensemble des sons attendus modulo certaines oppositions de voyelles.

Comme cela a pu être expliqué au chapitre 8, certaines oppositions phonologiques peuvent être ambigües voire inaudibles selon les contextes, soit parce que dans la variété de langues du locuteur (variété régionale, variété sociale, etc.) cette opposition est peu présente, soit parce que l'élève ne maitrise pas encore cette opposition. C'est le cas par exemple des oppositions /o/ et /ɔ/ d'une part, /e/ et /ɛ/ d'autre part. Certains auteurs, comme N. Catach (1980), proposent de neutraliser ces oppositions dans une représentation dite archiphonologique. Le tableau de ces oppositions est reporté ici.

Phonème 1	*Phonème 2*	*Archiphonème*
/e/	/ɛ/	/E/
/o/	/ɔ/	/O/
/ø/	/œ/	/œ/
/œ̃/	/ɛ̃/	/Œ̃/

Tableau 45 : Correspondance entre phonèmes et archiphonèmes

Les schwas finaux (*e* pouvant être muet ou prononcé selon les personnes, les régions ou les contextes, exemple : marche, transcrit /maʁʃø/ par *LIA-PHON* dans « le chat tombe de la marche et pleure. ») ont également été neutralisés. Ce choix est motivé par plusieurs facteurs : la prononciation ou non du schwa peut varier selon le contexte, une erreur sur le mot suivant peut donc la faire varier ; l'outil de phonétisation utilisé (*LIA-PHON*) n'inclut pas de symbole spécifique pour représenter le schwa. Un schwa peut donc ne pas être représenté ou être représenté par le phonème /ø/, ceci peut amener à tort l'algorithme à considérer certaines erreurs (exemple : « pleur » - *pleure*) comme des erreurs modifiant la phonologie.

Après transformation de la représentation phonologique en représentation archiphonologique, les formes archiphonologiquement normées sont les formes pour lesquelles les représentations archiphonologiques des segments et des formes sont identiques.

2.1.4. Les segments non normés

Les segments non normés ne correspondent à aucune des catégories précédentes. Ce sont les segments pour lesquels l'orthographe n'est pas respectée et dont la graphie ne reproduit pas l'ensemble des sons attendus.

Cette dernière catégorie regroupe deux types de segments. En premier lieu, elle rassemble les segments présentant une graphie proche de la graphie attendue mais qui ne correspond pas aux catégories précédentes (comparaisons relatives, « de » - *des* dans le tableau 45). En second lieu, elle contient les segments restants, dont l'algorithme n'a pas pu identifier d'alignements et qui sont classés comme *non normés* par l'aligneur (« les » - *escaliers* dans le tableau 45).

Id	*Niveau*	*Forme normée*	*Segment transcrit*	*Étiquette orthographique*	*Segmentation en mots*
93	CP	il	iltonbe	02-Phono	03-HypoSeg
93	CP	tombe	iltonbe	02-Phono	03-HypoSeg
93	CP	des	de	04-ApproxGraphique	01-Normé
93	CP	escaliers	les	07-Non normé	05-Non pertinent[97]
93	CP		seceille	08-Non pertinent	07-Inséré
93	CP	et	et	01-Normé	01-Normé

Tableau 46 : Exemple d'alignement produit par AliScol

Les segments identifiés comme *non normés* par l'algorithme peuvent aussi bien avoir été alignés correctement que mal alignés par celui-ci. Dans l'exemple donné dans le Tableau 45, l'aligneur échoue à identifier les tokens « seceille » et *escaliers* comme similaires, ce qui entraine des erreurs d'alignements. Cette marge d'erreur a été calculée au chapitre précédent (cf. Chapitre 9). La justesse des étiquettes orthographiques attribuées de manière automatique par l'aligneur est également à évaluer. Cette évaluation est présentée à la section 2.3.

[97] L'étiquette *non normé* est apposée uniquement dans les cas où l'algorithme échoue à identifier la correspondance entre deux tokens. Par défaut, les deux tokens sont alignés mais l'algorithme ne peut assurer que la segmentation en mots ait été correctement effectuée, l'étiquette non pertinent permet donc de spécifier cette indécision. Elle est également utilisée en cas d'insertion ou d'omission.

2.2. Données chiffrées

2.2.1. Données chiffrées longitudinales

Les données issues de l'algorithme concernant la performance orthographique sont reportées dans le tableau 46 et dans la figure 68. Le premier élément notable dans ces données est la prédominance des formes orthographiquement normées, et ce dès la première année d'apprentissage (au CP). Qui plus est, on note une progression durant ces quatre années d'apprentissage (presque 18 %).

Type d'erreur	*CP*	*CE1*	*CE2*	*CM1*	*CP-CM1*
Normé	1 721 (62,0 %)	5 441 (69,8 %)	9 662 (77,1 %)	12 150 (79,8 %)	28 974 (75,6 %)
Phonologie normée	491 (17,7 %)	1 177 (15,1 %)	1 419 (11,3 %)	1 588 (10,4 %)	4 675 (12,2 %)
Archiphonologie normée	129 (4,6 %)	305 (3,9 %)	318 (2,5 %)	376 (2,5 %)	1 128 (2,9 %)
Non normé	436 (15,7 %)	871 (11,2 %)	1 139 (9,1 %)	1 116 (7,3 %)	3 562 (9,3 %)
Total	*2 777*	*7 794*	*12 538*	*15 230*	*38 339*[98]

Tableau 47 : Répartition des réussites et des types orthographiques dans le corpus longitudinal

Un autre résultat particulièrement marquant est que, quelle que soit l'année d'apprentissage étudiée, les erreurs ne modifiant pas la phonologie sont plus importantes que les erreurs modifiant la phonologie, à plus forte raison lorsqu'on considère les erreurs archiphonologiques normées.

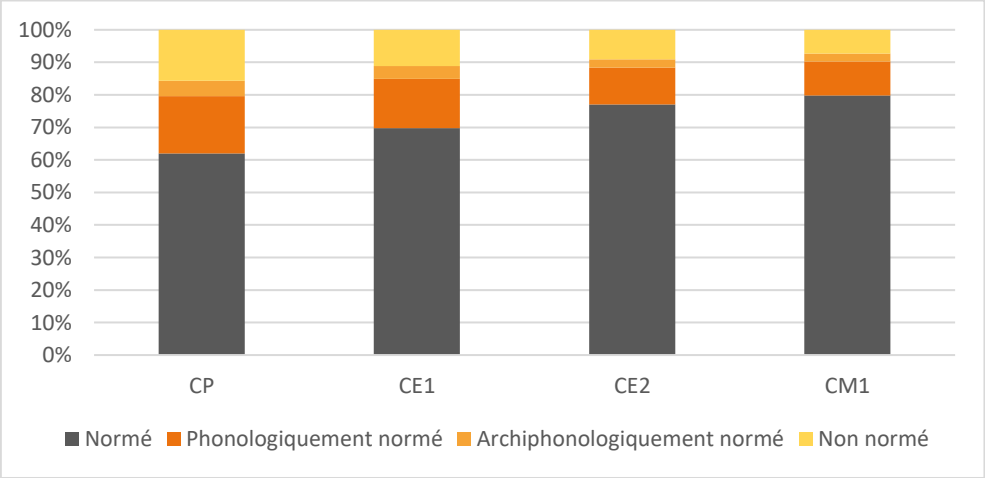


Figure 68 : Répartition des réussites et des types d'erreurs orthographiques année par année

[98] Ce nombre exclut les formes étiquetées comme normées par l'algorithme.

Dès la fin du CP, l'orthographe est maitrisée à plus de 60 % et les correspondances graphie-phonie à plus de 80 %. Notons cependant qu'il y a un réel enjeu autour des oppositions vocaliques que nous identifions comme des erreurs à archiphonologie normée. Les données issues de l'algorithme d'alignement ne permettent pas d'aller plus loin dans l'analyse de ces erreurs ; nous présentons au chapitre suivant un outil qui le permet.

2.2.2. Données chiffrées par catégories syntaxiques

Outre les types d'erreurs, l'algorithme *AliScol* renvoie également les lemmes et les catégories syntaxiques de chaque forme normée. En croisant ces deux types de données, il est possible d'étudier la répartition des échecs et des réussites orthographiques selon les catégories syntaxiques utilisées. La répartition des principales catégories syntaxiques est donnée dans le tableau 47 et la répartition des erreurs (en pourcentage) dans chacune de ces catégories est donnée dans la figure 69. Sont exclues du tableau 47 les catégories relevant de la ponctuation, pour lesquelles catégoriser les réussites en termes de phonologie n'auraient pas de sens, et les noms propres pour lesquels la norme orthographique peut ne pas exister ou de façon plus souple.

Catégorie syntaxique	*Nombre d'occurrences*	*Pourcentage d'occurrences*
NOM	7 443	21,1 %
VER	7 107	20,2 %
DET	5 513	15,7 %
PRO	5 089	14,4 %
PRP	3 109	8,8 %
KON	2 330	6,6 %
ADV	2 086	5,9 %
ADJ	1 385	3,9 %
Autre[99]	1 159	3,3 %
Total	*35 221*	*100 %*

Tableau 48 : Répartition des catégories syntaxiques (corpus longitudinal)

[99] La catégorie *Autres catégories* regroupe les nombres, les interjections, les abréviations et les symboles, ainsi que toutes les formes dont l'algorithme a échoué à étiqueter les catégories à l'aide de l'outil *TreeTagger*. Pour la plupart de ces formes, l'algorithme assigne une catégorie donnée par l'outil *LIA-PHON*, néanmoins ces catégories sont différentes de celles utilisées par *TreeTagger* et n'ont pas fait l'objet d'une évaluation, nous ne les prenons pas en compte ici.

Le tableau 47 fait apparaitre l'emploi massif des verbes et des noms par les élèves (un mot sur cinq). Il fait également apparaitre le nombre relativement élevé d'adjectifs utilisés par ceux-ci.

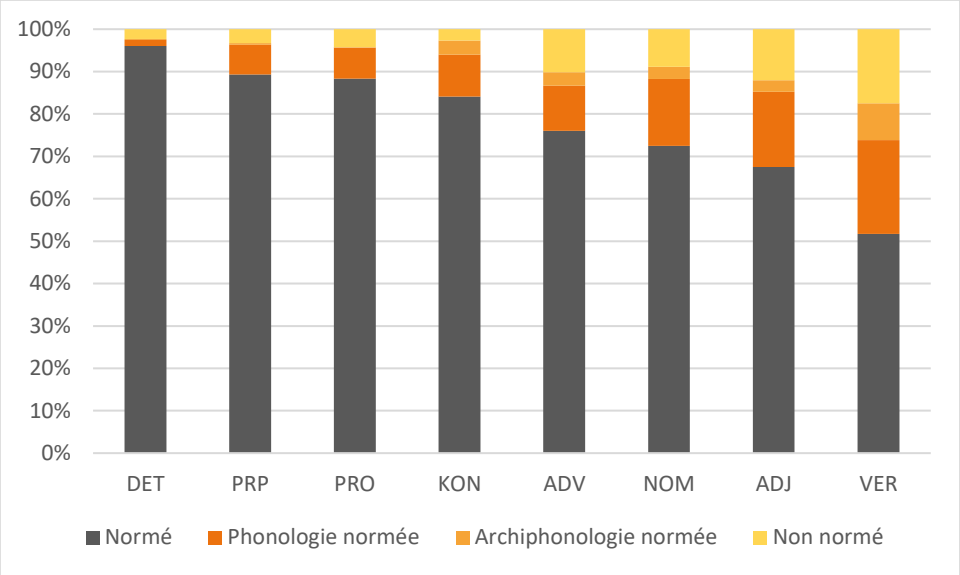


Figure 69 : Répartition des réussites orthographiques par catégorie

La figure 69 permet de mettre en évidence les catégories qui rencontrent le plus de résistance en termes d'apprentissage d'un point de vue orthographique. Comme on pouvait s'y attendre, ces catégories sont celles pour lesquelles existent des variations flexionnelles : les noms, les adjectifs et particulièrement les verbes. Ces résultats laissent à penser que l'orthographe flexionnelle est moins bien réussie par les apprenants que l'orthographe lexicale. Mais les données fournies par l'algorithme ne permettent pas de le vérifier, elles permettent uniquement de faire une analyse globale au niveau des unités lexicales, et ne permettent pas de distinguer, par exemple, les réussites orthographiques lexicales des réussites orthographiques grammaticales. C'est pourquoi l'équipe du projet *Scoledit* a conduit une recherche sur la morphographie verbale (la catégorie des verbes étant la catégorie présentant le plus faible taux de réussites orthographiques) permettant de détailler un peu plus ces résultats (cf. Chapitre 12).

2.3. Évaluation de la fiabilité des résultats

Les données que nous venons de présenter se basent uniquement sur les résultats de l'algorithme. Nous avons déjà mentionné qu'une analyse entièrement automatique ne peut être fiable à 100 %. L'évaluation présentée dans le chapitre 9 l'a d'ailleurs montré. Afin d'estimer la fiabilité des résultats présentés ici, il semble important également d'estimer la précision des étiquettes des erreurs orthographiques, des étiquettes des erreurs de segmentation en mots et des étiquettes morphosyntaxiques.

Pour évaluer la fiabilité des étiquettes, un procédé similaire à celui présenté précédemment (cf. Chapitre 9) a été utilisé. Le travail d'évaluation a été réalisé par deux équipes. Une première équipe a corrigé directement la sortie de l'aligneur (Tableau 48 pour un exemple). Cette équipe devait alors corriger simultanément l'ensemble des colonnes.

Identifiant de la production	*Identifiant du token normé*	*Lemme*	*Catégorie*	*Segment normé*	*Segment transcrit*	*Identifiant du token transcrit*	*Identifiant du segment transcrit*	*Réussite orthographique*	*Segmentation*
93_CE1	2	il	PRO:PER	Il	Il	2	1	01-Normé	01-Normé
93_CE1	3	être	VER:impf	était	êtais	3	1	03-Archi	01-Normé
93_CE1	4	un	DET:ART	une	tu	4	1	02-Phono	02-HyperSeg
93_CE1	4	un	DET:ART	une	ne	5	1	02-Phono	02-HyperSeg
93_CE1	5	foi|fois	NOM	fois	foit	6	1	02-Phono	01-Normé

Tableau 49 : Extrait du fichier de travail de l'équipe 1

Fâce à la complexité de la tâche, notamment en raison du nombre de colonnes trop important, les données de sortie de l'aligneur, corrigées par l'équipe 1, ont été séparées en deux fichiers distincts : un fichier contenant les étiquettes des différents types d'erreurs (Tableau 49) et un fichier contenant les étiquettes morphosyntaxiques (Tableau 50). L'équipe 2 a alors pu évaluer séparément ces deux fichiers.

IdProd	*SegNorm*	*SegTrans*	*StatutErreur*	*StatutSegm*
93_CE1	Il	Il	01-Normé	01-Normé
93_CE1	était	êtais	03-Archi	01-Normé
93_CE1	une	tu	02-Phono	02-HyperSeg
93_CE1	une	ne	02-Phono	02-HyperSeg
93_CE1	fois	foit	02-Phono	01-Normé

Tableau 50 : Extrait du fichier de travail pour les étiquettes d'erreur de l'équipe 2

IdProd	*Lemme*	*Catg*	*SegNorm*
93_CE1	il	PRO:PER	Il
93_CE1	être	VER:impf	était
93_CE1	un	DET:ART	une
93_CE1	fois	NOM	fois

Tableau 51 : Extrait du fichier de travail pour les étiquettes morphosyntaxiques de l'équipe 2

Une différence importante entre les corrections des deux équipes a pu être observée. Nous avons donc cherché à l'évaluer. Le tableau 51 recense le degré de différence entre les corrections proposées par l'équipe 1 puis les sur-corrections par l'équipe 2 (se reporter à la section 3 du chapitre 9 pour l'explication des calculs effectués).

Types d'étiquettes évaluées	*Degré de différence entre les deux équipes*
Étiquettes d'erreurs (orthographe et segmentation)	12,424 %
Étiquettes morphosyntaxiques	34,473 %

Tableau 52 : Degré de différence entre les corrections des deux équipes d'évaluation

Les résultats de ces calculs, reportés dans le tableau ci-dessus, confirment la grande différence de correction entre les deux équipes. Nous pouvons donc supposer que la taille des tableaux de données a influé sur la correction.

En prenant les résultats élaborés par l'équipe 2 pour référence, on obtient les mesures de précision et de rappel données dans le tableau 52. Les étiquettes d'erreurs orthographiques et d'erreurs de segmentation ont été calculées à partir du même fichier mais il parait intéressant de les distinguer.

Types d'étiquettes évaluées	*Précision*	*Rappel*	*F-mesure*
Étiquettes d'erreurs	88,1 %	88,1 %	88,1 %
➢ Étiquettes d'erreurs orthographiques	88,5 %	88,5 %	88,5 %
➢ Étiquettes d'erreurs de segmentation	91,8 %	91,7 %	91,7 %
Étiquettes morphosyntaxiques	94,1 %	93,7 %	93,9 %

Tableau 53 : Évaluation de la fiabilité des étiquettes apposées par l'aligneur

3. Conclusion

Bien que les résultats présentés ici n'atteignent pas une fiabilité de 100 %, les utiliser permet néanmoins de dessiner de grandes tendances afin de décider de nouvelles perspectives de

recherche. Ces perspectives de recherche doivent alors permettre de vérifier les tendances que l'algorithme a fait émerger. Nous pourrons alors étudier les phénomènes les plus saillants. C'est ce que nous proposons dans les chapitres suivants.

Le chapitre 11 permet de poursuivre l'analyse des réalisations orthographiques en termes de correspondances phonie-graphie en étudiant les productions à l'échelle du graphème. Ayant montré dans ce chapitre que la catégorie des verbes est la catégorie pour laquelle les apprenants rencontrent le plus de difficultés, le chapitre 12 présente une analyse plus détaillée de cette catégorie, et particulièrement de la morphographie verbale.

Chapitre 11 - Appliquer la méthode par comparaison aux graphèmes

1. Définition opératoire du graphème 215
2. Algorithme développé 222
3. Évaluation de l'algorithme 225
4. Conclusion 226

L'orthographe est la norme graphique, socialement et politiquement établie, qui définit la façon de transcrire les sons pour une langue donnée. Celle du français repose sur la correspondance entre les graphèmes (symbole graphique composé d'une ou plusieurs lettres) et les phonèmes (les sons de la langue). Ces correspondances sont appelées correspondances phonographiques. Elle a également recours à des signes, plus ou moins en lien avec la chaine sonore (plutôt moins que plus en français), qui renvoient à des significations linguistiques.

En d'autres termes, le fonctionnement du système d'écriture du français repose sur deux principes : le principe phonographique ou « écriture des sons » et le principe sémiographique ou « écriture du sens » (Fayol & Jaffré, 2014, p. 14). Pour reprendre les termes de M. Fayol et J.-P. Jaffré (2008, p. 232), la phonographie est « l'ensemble des procédés [...] qui permettent d'établir des correspondances entre des unités graphiques et des unités phoniques que sont les phonèmes et les syllabes ». La sémiographie regroupe « l'ensemble des unités pourvues de sens (morphèmes et mots) » (Ducard et al., 1995, p. 296).

Une orthographe est dite transparente lorsque les correspondances phonographiques qui la composent sont régulières, c'est-à-dire lorsqu'un graphème unique transcrit un phonème unique et inversement. Certains systèmes, comme l'italien, s'en approchent. À l'inverse, l'orthographe du français est peu transparente : en effet, si 96 % des correspondances graphèmes-phonèmes sont régulières, seulement 71 % des correspondances phonèmes-graphèmes le sont (Fayol & Jaffré, 2008). Cela signifie qu'il est possible de prédire à 96 % la prononciation d'un graphème ; en revanche, il n'est possible de prédire l'écriture d'un phonème qu'à 71 %. Ce phénomène s'explique par le fait que certains phonèmes du français peuvent être transcrits par plusieurs graphèmes, par exemple le phonème /o/ plus fréquemment écrit « o » peut s'écrire aussi « au ». Ce phénomène est appelé « polyvalence phonographique » (Fayol & Jaffré, 2008, p. 89). Une langue présentant une polyvalence phonographique importante est une langue à forte complexité phonographique.

Précédemment (Wolfarth, 2015), nous avions identifié trois raisons principales à cette complexité orthographique :

- Des raisons historiques (Catach, 1978) : pour écrire le français, c'est l'alphabet latin qui a été choisi. Cependant le latin est moins riche phonologiquement que le français. Un même signe a donc pu être utilisé pour transcrire deux sons différents. Par exemple, le signe *u* permettait à la fois de désigner le phonème /v/ et le phonème /y/ et des stratégies ont été développées pour distinguer des mots homographes. Par exemple, l'ajout d'un *h* devant le signe *u* pour distinguer *huitre* de *vitre*.
- Des différences de temporalité (Fayol & Jaffré, 2008) : l'oral évolue plus rapidement que l'écrit. Certaines différences de prononciation ont disparu mais l'écrit en garde encore des traces. Par exemple, il fut un temps où les graphies *o* et *au* se prononçaient différemment. Ce n'est plus le cas actuellement mais l'orthographe du français n'a pas accompagné ce changement.
- La distinction des homophones (Fayol & Jaffré, 2008) : dans un souci de faciliter la lecture, la norme écrite a tendance à différencier les homophones, à l'exemple du *p* de *corps* qui permet de faire le lien avec *corporel* et ainsi le distinguer de *cor* ou *cors*. Cette tendance s'est accrue entre le XIIe s. et le XIVe s. avec le développement de la lecture silencieuse et la nécessité d'une écriture et d'une orthographe plus iconique.

Pour toutes ces raisons, il ne suffit pas de connaitre la prononciation d'un mot pour savoir l'écrire. Cependant, certaines correspondances phonographiques sont plus évidentes (plus consistantes) que d'autres et peuvent être acquises plus tôt. C'est le cas par exemple du phonème vocalique /i/, très fréquent, transcrit par le graphème *i* dans 96 % des cas[100] ou du phonème consonantique /R/, très fréquent lui aussi, transcrit par le graphème *r*, tandis qu'il n'est transcrit par le graphème *rr* que dans 4 % des cas. Étudier ces mécanismes et ces différences permet donc de mieux comprendre les difficultés rencontrées par les apprenants.

L'algorithme d'alignement développé au chapitre 8 étudie les productions d'élèves à l'échelle des formes et ne permet donc pas d'étudier l'échelle des correspondances phonographiques. Afin de pallier ce problème, nous avons développé un nouvel algorithme, dont le

[100] Données issues de la base *Manulex-infra* : http://www.manulex.org/fr/infra/request.html [consulté le 08/09/2019].
Peereman, R., Lété, B., Sprenger-Charolles, L. (2007). Manulex-infra: Distributional characteristics of grapheme-phoneme mappings, infra-lexical and lexical units in child-directed written material. *Behavior Research Methods, 39*, 593-603.

fonctionnement est similaire au précédent, mais qui permet d'aligner la transcription des productions et leurs normalisations graphème par graphème.

Ce travail a été effectué lors d'un séjour de recherche à l'*Université Catholique de Louvain-La-Neuve*, en étroite collaboration avec Louise-Amélie Cougnon et assisté par les membres du *CENTAL*.

1. Définition opératoire du graphème

Comme nous venons de le dire, l'algorithme que nous souhaitons développer s'appuie sur un alignement en graphème. Il est donc nécessaire, en premier lieu, d'établir une définition opérationnelle du graphème, c'est-à-dire une définition basée sur des principes suffisamment précis pour qu'ils puissent être compilés en règles algorithmiques.

1.1. Définition du graphème dans la littérature

Les définitions de la notion de graphème sont nombreuses. J.-C. Pellat (1988) propose de les regrouper en quatre catégories selon leur façon d'envisager le graphème :

- Le graphème comme unité minimale du code écrit (Stetson, 1937 ; Haas, 1976, cités par Pellat, 1988). Selon ces définitions, le graphème est l'unité abstraite de l'écrit tandis que la lettre en est la réalisation en contexte. Le graphème prend donc la forme d'une lettre.
- Le graphème comme transcription d'un phonème (Imbs, 1971; Horejsi, 1972; Gak, 1976; Jakobson, 1976, cités par Pellat, 1988). La primauté est donnée à l'oral. Le graphème est alors défini par référence au phonème et il peut correspondre à une lettre ou à un groupe de lettres (*ai* dans *maison* par exemple).
- Le graphème comme unité minimale distinctive d'un système graphique donné (Uldall, 1944 ; Pulgram, 1951 ; Hjemslev, 1957 ; Gleason, 1961, cités par Pellat, 1988). En réaction au courant précédent, centré sur la langue orale au détriment de l'écrit, certains linguistes ont voulu réaffirmer la spécificité de l'écrit. Les définitions se rapportent alors à la sémiographie (au sens) ou à la morphographie, indépendamment de l'oral. Les graphèmes peuvent alors être une lettre ou un groupe de lettres, indépendamment de leur lien avec la chaine sonore (*t* dans *chat* par exemple).
- Le graphème comme unité graphique polyvalente (Fayol & Jaffré, 2008 ; Catach, 1979). Le graphème est vu comme la plus petite unité fonctionnelle de l'écrit, dont la fonction est soit de transcrire un son (*a* dans *chat*), soit de véhiculer un sens (*s* dans *chantes*). À nouveau, les graphèmes peuvent être une lettre ou un groupe de lettres.

Définition	*Nature*	*Critère*	*Principaux auteurs*
Unité minimale du code écrit	Lettre	graphique	Stetson (1937), Haas (1976)
Unité significative de l'écrit	Lettre ou suite de lettres	sémiographique	Uldall (1944), Pulgram (1951), Hjemslev (1957), Gleason (1961)
Correspondant graphique du phonème	Lettre ou suite de lettres	phonographique	Imbs (1971), Horejsi (1972), Gak (1976), Jakobson (1976)
Unité distinctive et significative de l'écrit	Lettre ou suite de lettres	phonographique et sémiographique	Fayol et Jaffré (2008), Catach (1979)

Tableau 54 : Synthèse des définitions du graphème (d'après Pellat, 1988)

Le tableau 53 récapitule les principes généraux des différentes définitions concurrentes. Pour notre travail, nous adoptons une définition proche de celle proposée par M. Fayol et J.-P. Jaffré (2008) et N. Catach (1979) qui prend en compte à la fois la dimension phonographique du graphème et sa dimension sémiographique, et plus particulièrement sa dimension morphographique.

Indépendamment de la fonction du graphème (phonographique ou sémiographique), N. Catach (1980) a défini quatre critères permettant d'identifier un graphème et de le distinguer des lettres hors système (étymologiques ou historiques) :

- La fréquence. Une lettre ou suite de lettres rencontrée fréquemment dans un texte est considérée comme un graphème. À l'inverse une lettre ou suite de lettres rencontrée peu fréquemment (dans les mots étrangers ou dans les toponymes par exemple) est jugée peu significative. Par exemple, la suite de lettres *oo* (*zoo* ou *foot*) est écartée du système des graphèmes du français.
- La cohésion et la stabilité. Une suite de lettres est considérée comme stable lorsque la composition de celle-ci ne varie pas, quelle que soit sa position dans le mot. Par exemple, la séquence *eau* (/o/ dans *oiseau* et dans *chapiteau*) est considérée comme stable, par opposition à la séquence *aut* (/o/ dans *saut* /ot/ dans *saute*). La séquence *eau* est un graphème, la séquence *aut* n'en est pas un.
- La signifiance ou la pertinence phonologique. Une suite de lettres peu signifiante, c'est-à-dire n'ayant pas d'incidence phonologique est considérée comme hors-système. Ainsi, pour N. Catach, la lettre *h* dans *rhume* est une lettre qui n'a plus de contact avec l'oral, détachée du système.
- La rentabilité linguistique ou la créativité linguistique. Une suite de lettres est dite rentable si elle peut être dérivée ou fléchie de manière sérielle et créative pour former

de nouveaux mots : par exemple le suffixe et graphème *eau* qui permet de former une série suffixale ouverte, à l'exemple de *chèvre* / *chevreau*.

Fonder un algorithme sur la notion de graphème nécessite d'en définir des critères opératoires. Le travail d'établissement de critère d'identification de graphème opéré par N. Catach, s'approche du travail qu'il nous a fallu réaliser. C'est pourquoi, même si nous n'utilisons pas la notion de lettres hors systèmes, les critères précités ont inspiré notre travail.

1.2. Définition du graphème adoptée pour l'algorithme *AliScol_Graph*

Dans le cadre de l'alignement réalisé par cet algorithme, nous avons distingué quatre types de composants :

- les **graphèmes** qui sont composés d'une ou plusieurs lettres ;
- les **ponctuants** qui sont composés d'un ou plusieurs signes de ponctuation, de marques de dialogue ou des caractères de séparation (apostrophe et trait d'union) ;
- les **balises** présentes dans la normalisation ;
- les **nombres** et les **symboles** qui sont composés de chiffres, de symboles (*€*, *%*, *+*, etc.) ou de lettres à caractère abréviatif (*h*, *km*, etc.).

1.2.1. Les graphèmes

Nous utilisons une définition proche de celles données par N. Catach (1980) et par M. Fayol et J.-P. Jaffré (2008). Un graphème est une lettre ou un groupement de lettres correspondant à un phonème (phonogramme) ou ayant une signification linguistique, le plus souvent sans lien avec la chaine sonore.

Dans le cadre de sa thèse soutenue en 2017, J. Riou a cherché à évaluer la part déchiffrable des textes en fonction des correspondances phonographiques vues précédemment avec les élèves. Ce travail, en partie automatisé et outillé, a nécessité le découpage de textes en graphèmes et, de ce fait, la mise au point d'une définition opératoire du graphème. La définition que nous proposons dans le cadre de notre travail s'inspire des principes proposés par J. Riou (2017, pp. 89-92) et sur la liste des graphèmes qu'il a établie (2017, pp. 93-94 ; reprise en Annexe 9). L'ensemble des critères retenus est explicité dans un protocole disponible en annexe (Annexe 10). Nous détaillons ces critères dans la suite de ce chapitre.

Pour identifier les graphèmes, la priorité est donnée aux sons plutôt qu'au sens. Nous nous appuyons donc en premier lieu sur un critère phonographique, pour repérer les phonogrammes, puis sur un critère morphographique, qui relève plus largement de la sémiographie, pour relever les morphogrammes sans correspondant phonique. Le premier critère permet d'identifier l'ensemble des graphèmes impliqués dans la transcription de la

chaine sonore (étape (1), Figure 70). À l'issue de l'application de ce critère, ne restent que les lettres (ou groupements de lettres) dites « muettes ». Afin de discriminer les graphèmes de ces groupements de lettres, le critère morphogrammique est invoqué (étape (2), Figure 70).

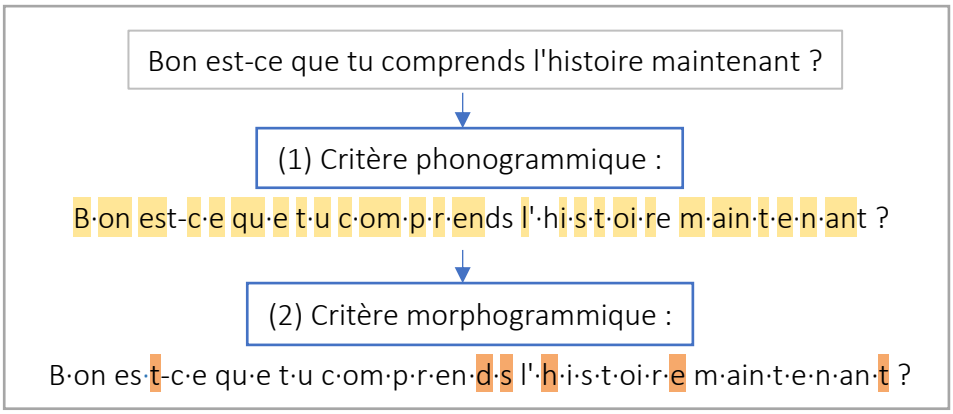


Figure 70 : Processus d'application des critères de discrimination des graphèmes

1.2.1.1. Application des règles de discrimination des graphèmes, le critère phonogrammique

Le critère phonogrammique est défini comme suit : un graphème est composé d'une lettre ou d'un groupement de deux ou plusieurs lettres et correspond à un phonème (Peereman et al., 2007).

Nous avons décliné ce principe en 4 principes opératoires :

<u>Principe 1 :</u> Deux ou plusieurs lettres transcrivant deux ou plusieurs sons sont considérées comme des graphèmes distincts.

(1) sac → s – a – c ; /sak/ (3 graphèmes correspondant à 3 phonèmes)

<u>Principe 2 :</u> À l'inverse, une suite de lettres transcrivant un son unique est considérée comme un seul graphème.

(2) raisin → r – ai – s – in ; /Rɛzɛ̃/ (4 graphèmes correspondant à 4 phonèmes)

<u>Principe 3 :</u> Si la délimitation de ces groupements de lettres est problématique, nous adoptons alors un principe d'indépendance. Selon ce principe, une lettre pouvant être considérée comme muette dans des contextes proches (*t* dans l'exemple (3)) est considérée comme un graphème indépendant. En revanche, une lettre qui ne peut pas ou que très occasionnellement être considérée comme une lettre muette dans un contexte proche (*a* dans l'exemple (4)(1) et *e* dans l'exemple (5)) est considérée comme intégrée au graphème.

(3) fait → f – ai – t ; /fɛ/

(4) pain → p – ain ; /pɛ̃/

(5) eau → eau ; /o/

La lettre *t* (exemple (3)) est muette dans de nombreux contextes, *est*, *rat*, *lit* par exemple ; elle est donc considérée comme un graphème indépendant. En revanche, la lettre *a* n'est muette que lorsqu'elle est suivie des suites *in* et *im* ; elle est donc considérée comme intégrée au graphème *ain* ou *aim*.

Dans ce dernier exemple (exemple (5)), la non indépendance du *e* est discutable. En effet, la lettre *e* peut être considérée comme muette dans d'autres contextes proches (exemple (6)), mais ici elle présente une valeur auxiliaire, c'est-à-dire qu'elle modifie la phonologie de la lettre *g.* De plus, nous considérons qu'il s'agit d'un contexte trop limité (uniquement après la lettre *g* et devant une voyelle), pour convenir au critère d'indépendance.

(6) mangea → m – an – g – e – a ; /mɑ̃ʒa/

À ces principes s'ajoutent un cas particulier, à savoir les séquences *e* + consonne (*r*, *s*, *t*, *z*, etc.). Selon les modèles linguistiques, ces séquences comme *et* ou *ez* (exemple : gagnez, poulet) sont considérées comme un unique graphème (Riou, 2017), correspondant au phonème /e/, ou comme deux graphèmes (Blanche-Benveniste & Chervel, 1969), le graphème *e,* correspondant au phonème /e/, et un graphème muet à valeur auxiliaire composé d'une consonne (*r*, *s*, *t*, *z*, etc.). Or, si certaines des consonnes qui composent la séquence (*s* et *t* par exemple) peuvent être considérées comme indépendantes en lettres muettes finales, ce n'est pas le cas du *e*, correspondant au phonème /e/. En effet, pour que la lettre *e* puisse prendre la valeur phonique /e/, des lettres contextuelles spécifiques sont nécessaires (doublement de la consonne ou consonne finale par exemple), si ce n'est pas le cas, un accent est nécessaire. Dans l'exemple, *nagez*, le *z* a donc pour fonction phonique de modifier la valeur phonique de *e*, au même titre que l'accent dans la lettre *é* (exemple (7)). Il nous semble donc nécessaire de pouvoir mettre en correspondance ces deux séquences, les séquences *e* + consonnes sont donc considérées comme un seul graphème (*er* dans l'exemple (8) et *ez* dans l'exemple (9)).

(7) nagé → n – a – g – é ; /naʒe/

(8) nager → n – a – g – er ; /naʒe/

(9) nagez → n – a – g – ez ; /naʒe/

Principe 4 : Il peut arriver qu'une seule lettre transcrive plusieurs sons (*x* dans l'exemple (10)), elle sera considérée comme un graphème.

(10) examina → e – x – a – m – i – n – a ; /ɛgzamina/

Note 1 : Précisons que selon ces principes, un même groupement de lettres peut contenir un nombre de graphèmes différent selon les contextes (*gu* dans les exemples (11) et (12) et *cc* dans les exemples (13) et (13)).

(11) aiguille → ai – g – u – i – ll – e (*gu* ; /gɥ/ : 2 graphèmes)

(12) anguille → an – gu – i – ll – e (*gu* ; /g/ : 1 graphème)

(13) vaccin → v – a – c – c – in (*cc* ; /ks/ : 2 graphèmes)

(14) accrédita → a – cc – r – é – d – i – t – a (*cc* ; /k/ : 1 graphème)

Note 2 : **les séquences *oi*, *oin*.** Les séquences oi (/wa/), oin (/wɛ̃/) qui transcrivent deux phonèmes sont considérées comme un graphème unique en raison du principe d'indépendance (principe 3). En effet, à l'exception de la séquence *oy* (/waj/, exemple (15)), la lettre *o* n'a jamais /w/ pour valeur phonique.

(15) voyage → v – o – y – a – g – e ; /vwajaʒ/

(16) poids → p – oi – ds ; /pwa/

(17) foin → f – oin ; /fwɛ̃/

1.2.1.2. Application des règles de discrimination des graphèmes, le critère morphogrammique

Le critère morphogrammique est défini comme suit : un graphème composé d'un groupement d'une ou plusieurs lettres correspond à une seule valeur morphologique.

Les principes définis pour étayer ce critère sont les suivants :

Principe 1 : Une lettre muette isolée est considérée comme un graphème (*e*, marque de la 3e personne du singulier du présent, dans l'exemple (18)).

(18) marche → m – a – r – ch – e;

Principe 2 : Une suite de lettres transcrivant une même information morphologique est considérée comme un seul graphème (*ent* comme marque de personne dans l'exemple (19)).

(19) vivaient → v – i – v – ai – ent (1 seul graphème muet)

Principe 3 : À l'inverse, deux ou plusieurs lettres encodant deux ou plusieurs informations morphologiques sont considérées comme des graphèmes distincts (*e* comme marque de temps et de mode et *nt* comme marque de personne dans l'exemple (20)).

(20) vivent → v – i – v – e – nt (2 graphèmes muets distincts)

Précisons ici, que selon le modèle linguistique utilisé (Pellat, 2009, par exemple), il est possible de considérer que la forme *vivent* ne contient aucune marque de temps, et que la séquence *ent* représente la marque de personne, comme dans l'exemple *vivaient*. Nous avons préféré considéré que la lettre *e* représentait une marque de temps, notamment pour des raisons de facilité de traitement informatique.

1.2.2. Les ponctuants

Sont reconnus comme ponctuants les signes de ponctuation, les marques de dialogue et les caractères de séparation (apostrophe et trait d'union).

Principe 1 : Les ponctuants se distinguent des lettres et constituent donc des unités à part.

(21) c'est → c – ' – es – t

Principe 2 : Lorsque plusieurs ponctuants ont même fonction (plusieurs signes de ponctuation successifs par exemple), ils sont considérés comme une seule unité d'alignement (exemple (22)).

(22) quoi !? → qu – oi – !?

Principe 3 : À l'inverse, s'ils n'ont pas même fonction (signe de ponctuation et marque de dialogue par exemple), ils sont considérés comme des unités distinctes (exemple (23)).

(23) !" → ! – "

1.2.3. Les balises

La normalisation comporte différentes balises qui sont considérées comme des unités d'alignement (exemple (24)).

(24) <segmentation/> → <segmentation/>

Deux balises distinctes consécutives sont considérées comme deux unités d'alignement distinctes (exemple (24)).

(25) <incomprehensible/> <nonfini/> → <incomprehensible/> – <nonfini/>

1.2.4. Les nombres et les symboles

Les nombres (un ou plusieurs chiffres) sont considérés comme des unités d'alignement (exemple (26)).

(26) 3 000 → 3 000

Les symboles (exemple (27)) et les lettres ou groupes de lettres (exemple (28)) représentant des unités ou des abréviations d'unités sont considérés comme des unités d'alignement.

(27) % → %

(28) km → km

Lorsqu'un nombre est suivi d'un symbole, d'une lettre ou d'un groupe de lettres représentant une unité, ils sont considérés comme des unités d'alignement distinctes.

(29) 3h → 3 - h

2. Algorithme développé

L'algorithme d'alignement développé, appelé *AliScol_Graph*[101], permet de découper chaque couple de formes (segment transcrit et forme normalisée) en graphèmes puis d'aligner ces graphèmes. Il s'appuie donc sur un alignement préalable des productions forme par forme. Il se déroule en quatre étapes (Figure 71).

1) **Extraction des formes transcrites et normalisées** : à partir de la sortie de l'aligneur *Aliscol*, les formes transcrites et normalisées sont récupérées.
2) **Prétraitements** : les formes transcrites et normalisées sont découpées en graphèmes grâce à l'appel de la fonction de segmentation en graphèmes de l'outil *LIA-PHON*. Les représentations phonologiques et archiphonologiques des graphèmes sont recalculées. Puis le découpage établi par l'outil *LIA-PHON* est corrigé pour correspondre à la définition du graphème que nous avons établie précédemment.
3) **Alignement des graphèmes** : les graphèmes des productions transcrites et les graphèmes des productions normalisées sont alignés à l'aide d'un calcul de distance de Levenshtein à couts variables.
4) **Affichage des productions** : le résultat de l'alignement est mis sous forme de tableau afin de pouvoir être affiché dans un fichier de sortie.

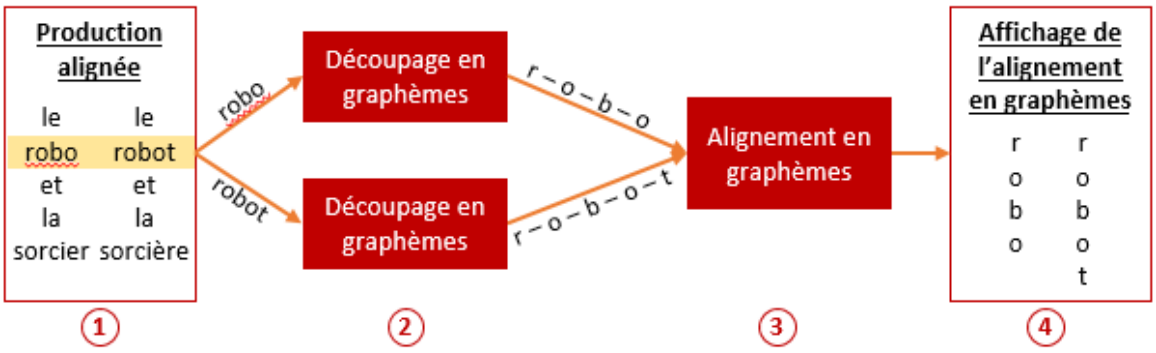


Figure 71 : Schéma général de l'algorithme d'alignement en graphèmes AliScol_Graph

2.1. Découpage en graphèmes (prétraitements, étape2)

2.1.1. Appel de *LIA-PHON*

Chaque couple de formes (segment transcrit et forme normalisée) est découpé en graphèmes à l'aide de l'outil *LIA-PHON*. En plus du découpage en graphèmes, cet outil donne également la

[101] Le code de cet algorithme, rédigé en langage Perl, est disponible à l'adresse https://claire.wolfarth.cf/.

représentation phonologique de chaque graphème. Un exemple de sortie de cet outil est donné dans le tableau 54.

Forme d'entrée	*Sortie LIA-PHON*
les	les lll#esei#
sorcière	sorcière sss#ooo#rrr#css#iyy#èai#rrr#e#

Où (èai#, par exemple) :

- è : forme graphique du graphème (1, 2 ou 3 caractères)
- ai : représentation phonologique du graphème (2 caractères)
- # : séparateur de graphèmes

Tableau 55 : Exemple de sortie de la fonction de découpage en graphème de l'outil LIA-PHON

2.1.2. Ajustement du découpage en graphèmes

La définition du graphème utilisée par l'outil *LIA-PHON* n'est pas exactement la même que celle que nous avons donnée précédemment. Une étape d'ajustement est donc nécessaire pour atteindre le découpage attendu.

Pour exemple, les séquences *et*, *er*, *es* et *ez* sont souvent scindées en deux graphèmes différents par l'outil *LIA-PHON*. Dans notre modèle, nous avons considéré que ces séquences, lorsqu'elles correspondaient aux phonèmes /e/ ou /ɛ/, constituaient un unique graphème. L'étape d'ajustement permet donc de modifier ce découpage.

L'outil *LIA-PHON* considère également que l'apostrophe se rattache au graphème précédent. Dans notre modèle, nous considérons qu'il s'agit d'un graphème à part, étant donné que l'apostrophe a souvent pour rôle de remplacer le graphème *e* (exemple *ce* découpé *c – e*).

Enfin, pour donner un dernier exemple, les consonnes doubles sont souvent considérées par l'outil *LIA-PHON* comme des graphèmes différents, un graphème muet et un graphème sonore. Dans ce cas, nous les considérons comme un seul graphème.

Forme	*Découpage LIA-PHON*	*Découpage après ajustement*
et	eei#t#	etei#
qu'	qu'kk#	qukk#'#
comme	ckk#ooo#m#mmm#e#	ckk#ooo#mmmm#e#

Tableau 56 : Quelques exemples d'ajustement du découpage en graphème

2.1.3. Calcul des représentations archiphonologiques

À partir de cette sortie, les représentations archiphonologiques des graphèmes sont calculées à l'aide du même module que celui utilisé dans l'algorithme *Aliscol* (cf. Chapitre 8 - 3). À l'issue de cette étape, pour chaque forme ou segment, nous disposons du découpage en graphèmes et de la représentation phonologique et archiphonologiques de chaque graphème.

sorcière	s - ss - ss	o - oo - au	r - rr - rr	c - ss - ss	i - yy - yy	é - ei - ei	r - rr - rr	e - -

Où :

- s - : forme graphique du graphème
- - ss - : représentation phonologique du graphème
- - ss : représentation archiphonologique du graphème

Tableau 57 : Données disponibles pour chaque forme ou segment après découpage en graphèmes et calcul de la représentation archiphonologique

2.2. Alignement des graphèmes (étape 3)

À l'issue de l'étape précédente, pour chaque forme normalisée et pour chaque segment transcrit, nous disposons d'un découpage en graphèmes. Il est donc désormais possible d'aligner ces différents graphèmes. Pour ce faire, nous utilisons un algorithme de calcul de distance d'édition de Levenshtein[102]. Tout comme dans l'algorithme d'alignement *Aliscol*, la variante de l'algorithme de distance d'édition que nous proposons prend en compte des critères graphiques et des critères phonologiques. Dans cet algorithme, le cout des substitutions ne modifiant pas la phonologie ou l'archiphonologie est nul, c'est-à-dire que les identités graphiques, phonologiques et archiphonologiques ont le même cout.

Type d'opération	*Cout*	*Exemple*
Identité graphique Identité phonologique et Identité archiphonologique	0	ai || ai ai /ɛ/ || è /ɛ/ ai /ɛ/ /E/ || é /e/ /E/
Autre substitution	1	b || d

Tableau 58 : Couts utilisés pour le calcul d'édition dans l'algorithme Aliscol_Graph

Une entreprise d'évaluation de l'impact de la variation des couts a été menée sur un échantillon restreint, mais elle a rapidement montré qu'attribuer un cout différencié aux substitutions ne modifiant pas la phonologie ou l'archiphonologie était néfaste aux performances de l'algorithme.

[102] Le calcul de distance d'édition de Levenshtein est expliqué au Chapitre 7 - 2.1.

3. Évaluation de l'algorithme

Une première évaluation de l'algorithme d'alignement en graphèmes a été réalisée. Cette évaluation, encore sommaire, a été réalisée sur un échantillon de 36 productions, soit 14 040 graphèmes. Le tableau 58 recense quelques caractéristiques du corpus d'évaluation.

Niveau	*Nombre de productions*	*Nombre de formes*	*Nombre de graphèmes*
CP	9	249	1 014
CE1	9	598	2 477
CE2	9	1281	5 409
CM1	9	1349	5 500
Total	36	3477	14 040

Tableau 59 : Caractéristiques du corpus d'évaluation

Pour procéder à l'évaluation, nous adoptons une méthode similaire à celle proposée pour le projet *ARCADE* (Véronis & Langlais, 2000). En premier lieu, un alignement est réalisé de façon indépendante par deux annotateurs humains[103] à partir d'un guide élaboré conjointement au préalable. À partir de ces deux alignements, un accord inter-annotateurs est calculé, selon la formule donnée au chapitre 9, afin de connaitre la fiabilité du guide établi. Les résultats de ce calcul sont reportés dans le tableau 59. Ces scores d'accord inter-annotateurs semblent suffisamment élevés pour valider le guide d'alignement.

Niveau	*Accord inter-annotateurs*
CP	92,9 %
CE1	92,1 %
CE2	90,4 %
CM1	94,1 %
CP-CM1	92,3 %

Tableau 60 : Mesures de l'accord inter-annotateurs

En second lieu, l'alignement réalisé par l'aligneur peut être calculé en comparant la sortie de l'aligneur avec l'alignement manuel le plus semblable. À nouveau, nous employons les mesures de précision et de rappel, ainsi que le calcul de F-mesure (cf. Chapitre 9), ces mesures sont reportées dans le tableau 60.

[103] Dans ce cas précis Louise-Amélie Cougnon et nous-même.

Niveau	*Rappel*	*Précision*	*F-Mesure*
CP	72,0 %	76,5 %	74,2 %
CE1	76,2 %	81,7 %	78,9 %
CE2	76,4 %	81,8 %	79,0 %
CM1	79,1 %	82,3 %	80,7 %
CP-CM1	77,1 %	81,6 %	79,3 %

Tableau 61 : Évaluation des performances de l'aligneur AliScol_Graph

Ces résultats révèlent des performances moins bonnes pour l'algorithme *AliScol_Graph* que pour l'algorithme *AliScol*. La principale explication réside sans doute dans l'investissement en temps très différent accordé aux deux algorithmes. Alors que l'algorithme *AliScol* a pu être travaillé pendant plusieurs années, le travail réalisé pour l'algorithme *AliScol_Graph* n'a duré que quelques mois.

Une observation à postériori de cet algorithme permettrait donc sans doute d'améliorer ses performances. Il est à noter qu'une grande partie des erreurs réalisées semblent être des erreurs de découpage en graphèmes, davantage que des erreurs d'alignement des graphèmes. Un travail supplémentaire sur le module de segmentation en graphèmes devrait donc permettre d'améliorer les performances de l'aligneur. L'aligneur échoue principalement à distinguer les séquences ambigües pouvant constituer un ou plusieurs graphèmes selon les contextes, comme ill (v-i-ll-e *vs* f-ill-e). De même, les formes ou segments contenant des graphèmes de type *e* + consonne présentent régulièrement des erreurs (« diser » segmenté « d-i-s-e-r » plutôt que « d-i-s-er »).

4. Conclusion

L'approche par comparaison peut également être appliquée à l'échelle des graphèmes mais nécessite des outils spécifiques comme un aligneur dont nous avons proposé ici une ébauche. Grâce à cet outil, différentes analyses linguistiques sont possibles comme l'étude de certaines correspondances phonographiques, particulièrement les correspondances impliquant plusieurs graphies pour un même son (/o/ - *o*, *au* ou *eau* par exemple) et inversement les correspondances impliquant un graphème et plusieurs phonèmes (*c* - /s/ ou /k/ par exemple).

Chapitre 12 - Analyser la morphographie verbale grâce à l'approche par comparaison

1. Quelle théorie du verbe ? 227
2. Caractérisation des verbes 229
3. Distinction des erreurs sur la base et sur la désinence 233
4. Conclusion 236

L'approche par comparaison peut également s'appliquer au niveau morphologique. Nous en donnons un exemple dans ce chapitre qui concerne la morphographie verbale[104]. La réalisation de cette étude se base sur un sous-corpus de 903 productions dont la répartition est résumée dans le tableau 61.

	Nombre de productions	*Nombre de formes*	*Longueur moyenne de production*	*Nombre de formes verbales*
CP	301	7 877	26,17	921
CE1	301	20 125	133,72	9 792
CE2	301	35 195	233,85	4 638
Tous niveaux (CP-CE2)	*903*	*63 197*	*131,25*	15 351

Tableau 62 : Données générales concernant le corpus recueilli

1. Quelle théorie du verbe ?

La tradition scolaire, du moins jusqu'aux programmes de 2008, classe les verbes en trois groupes selon leur terminaison infinitive et distingue radical et terminaison. Cette classification, héritée du latin distingue :

1) les verbes terminés par *-er* à l'infinitif (sauf parfois le verbe *aller*) ;
2) les verbes terminés par *–ir* qui ont une base longue en *-iss-* ;
3) tous les autres verbes, qui ne rentrent pas dans les deux premières catégories, et qui finissent par *-ir*, *-oir*, *-re*.

[104] Celui-ci reproduit en partie un article publié dans le n° 57 de la revue *Repères* (Wolfarth, Ponton, et al., 2018). L'étude qui y est présentée s'appuie à la fois sur des résultats issus de l'aligneur *AliScol* présenté au chapitre 8 et sur des résultats issus d'un algorithme original développé dans le cadre du projet *Scoledit*.

En raison du grand nombre de modèles de conjugaison qu'engendre cette répartition des verbes, elle ne nous parait pas opérationnelle pour l'étude de notre corpus, et plus particulièrement pour les traitements informatiques que l'on souhaite y appliquer. En effet, une telle conception de la flexion verbale entraine une parcellisation des verbes en près d'une centaine de modèles de conjugaison : on en compte 82 pour *Le nouveau Bescherelle, l'art de conjuguer* (1980) et pas moins de 140 pour *Conjugaison*, Le Robert et Nathan (1996) (Brissaud, 2002).

À la suite de nombreux linguistes (par exemple Martinet, 1979 ; Meleuc & Fauchart, 1999, Blanche-Benveniste, 2002 ; Pellat, 2009), nous optons pour un classement en deux catégories : les verbes en *–er* (dont l'infinitif se termine par /e/ à l'oral et *-er* à l'écrit) et les autres verbes (dont l'infinitif se termine par /R/ à l'oral) et nous proposons une entrée via les désinences (ou terminaisons) en fonction de la personne (P1 = première personne du singulier, P4 = première personne du pluriel). L'objectif de cette approche est de générer des modèles de désinences opérationnels pour le plus grand nombre de verbes. Selon cette conception, la conjugaison se réduit donc à un nombre restreint de modèles de désinences, résumés pour l'indicatif dans le tableau ci-dessous.

Temps	*Types de verbe*	*P1*	*P2*	*P3*	*P4*	*P5*	*P6*
Présent de l'indicatif	Verbes en -er	-e	-es	-e	-ons	-ez	-ent
	Autre verbes	-s	-s	-t (ou Ø)			
Imparfait	Tous les verbes	-ais	-ais	-ait	ions	-iez	-aient
Passé simple	Verbes en -er	-ai	-as	-a	-^mes	-^tes	-rent
	Autre verbes	-(v)s	-(v)s	- (v)t			

Tableau 63 : Désinences du présent de l'indicatif, de l'imparfait et du passé simple[105]

Ainsi, pour l'imparfait, nous considérons que tous les verbes à la deuxième personne (P2) de l'imparfait de l'indicatif ont pour désinence *–ais*, et tout ce qui vient avant cette désinence est considéré comme appartenant à sa base (ou radical), un même verbe pouvant posséder plusieurs bases (jusqu'à sept en se fondant sur l'oral, hors passé simple, infinitif et participe passé ; Dubois, 1967). Nous considérons donc qu'il n'y a pas de particularité de désinences pour les verbes du deuxième groupe, contrairement à certaines conceptions issues d'une certaine tradition grammaticale, qui distingue les verbes comme *fin-ir* / *fin-issais* des verbes comme *ven-ir* / *ven-ais* (par exemple dans le *Bescherelle, la conjugaison pour tous*, 1997). Pour le passé

[105] L'absence de marque pour la troisième personne du présent de l'indicatif ne concerne que certains verbes terminés en *-dre*.

simple, nous considérons qu'il n'y a qu'un seul modèle de conjugaison pour les personnes dites du pluriel (P4 à P6), qui se décline selon 4 voyelles données par la base du verbe (notées *v* entre parenthèses : *a, i, u, in*). Pour le présent, nous ne considèrerons que deux modèles. Pour la 3e personne, le *–t* peut être assimilé par la lettre finale de la base *–d*.

Le tableau 63 précise les différentes marques pour le mode participe passé :

Types de verbe	*Masculin Singulier*	*Féminin Singulier*	*Masculin Pluriel*	*Féminin Pluriel*
Verbes en *-er*	-é	-ée	-és	-ées
Verbes en *-oir*, *-oire*, *-ure*, *-dre*, *-aire*, *-aitre*, certains verbes en *-ire* et *-ivre*	-u	-ue	-us	-ues
Verbes en *-ir*, certains verbes en *-ire* et *-ivre*	-i	-ie	-is	-ies
Verbes en *-indre* et un petit nombre de verbes, comme *faire*, *dire*, *écrire*, *ouvrir*, *mourir*, *cuire*, *conduire*, *frire*, *traire*	-t	-te	-ts	-tes
Un petit nombre de verbes, comme *assoir*, *prendre* et *mettre*	-s	-se	-s	-ses

Tableau 64 : Les différentes marques du mode participe passé

En français, les participes passés se terminent majoritairement par une voyelle « nue », sans consonne. Les participes passés en -s et -t sont relativement peu nombreux mais ils concernent certains verbes très fréquents comme *faire* et *dire*.

2. Caractérisation des verbes

2.1. Répartition des temps verbaux

L'algorithme *AliScol* utilisé en combinaison de l'étiqueteur *TreeTagger* permet d'observer la répartition des temps verbaux dans les productions d'élèves. Étant donné la faible fréquence de certains tiroirs verbaux et du fort taux d'erreurs qu'ils engendraient via les traitements automatiques, nous avons choisi de nous concentrer sur les cinq tiroirs verbaux les plus fréquents : imparfait de l'indicatif, infinitif, participe passé, présent de l'indicatif et passé simple.

La figure 72, ci–dessous, précise la répartition des temps des verbes rencontrés selon le niveau de scolarité.

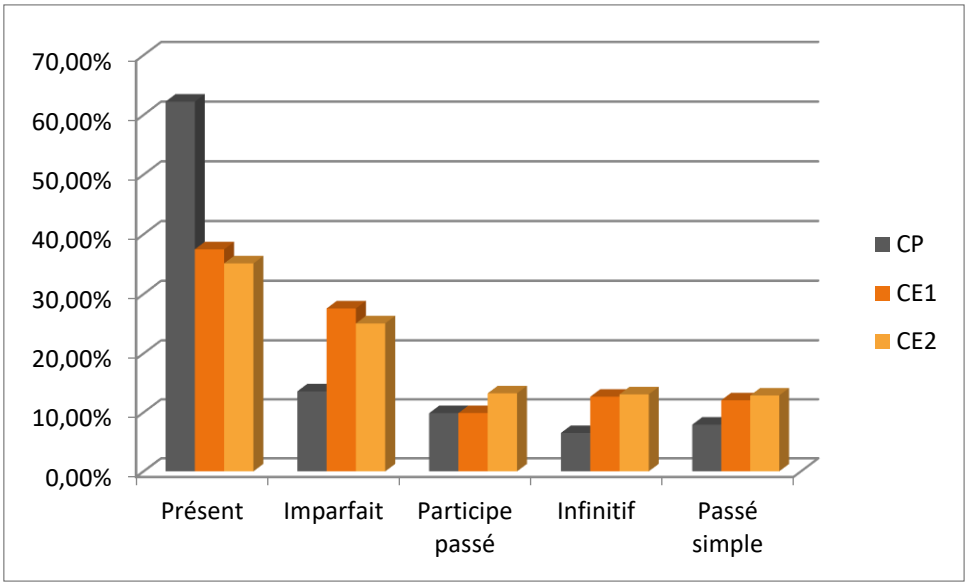


Figure 72 : Répartition des temps et modes verbaux selon le niveau de scolarité

Cette figure permet de mettre en lumière la répartition des temps verbaux que nous avons choisi d'étudier selon l'année de recueil. Nous noterons cependant que les consignes de production étaient différentes l'année de CP et les années suivantes ; il est possible que la consigne donnée en CP ait amené une prédominance de verbes au présent, représenté à 62 % contre 35 % et 37 % respectivement pour le CE1 et le CE2, ce qui conduit, pour ces deux niveaux, à une plus faible représentation des autres temps.

Au-delà de ces deux tiroirs verbaux particuliers, la répartition semble être relativement constante au cours des années, avec une prédominance de l'imparfait et du présent. On constate une présence plus importante de verbes au participe passé au CE2 qui peut laisser penser, nos données se basant sur les mots verbaux et non sur les formes verbales, qu'il s'agit en réalité de passé composé. Ces données convergent avec celles de M. Savelli, C. Brissaud, J.-P. Chevrot et V. Gounon, 2002) concernant la fréquence d'usage des tiroirs verbaux : dans leur étude sur le passé simple conduite auprès d'élèves de 3-6 ans (à l'oral), d'une classe de CM1 (à l'écrit) et de classes de 6e (à l'écrit), les tiroirs verbaux les plus utilisés dans le récit sont le passé simple (4,5 % *vs* 34,2 %), l'imparfait (13,8 % *vs* 19,2 %), le présent (61,5 % *vs* 27,8 %), qui diminue largement entre les 3-6 ans et les 6e, et le passé composé (12,3 % *vs* 9,1 %). De plus, on retrouve des dynamiques similaires à celles observées dans cette étude : un accroissement de l'usage du passé simple entre les 3-6 ans et les 6e et une nette diminution de l'usage du présent. Notons cependant que l'usage de l'imparfait progresse davantage dans notre corpus que dans l'étude de Savelli et ses collègues (2002).

2.2. Répartition des erreurs par temps

Comme nous l'avons évoqué précédemment, les indices utilisés dans la phase d'alignement nous permettent de répartir les formes en trois catégories : normées (formes correctes), non normées à phonologie normée[106] (forme incorrecte mais qui permet de restituer la forme sonore), non normées (forme incorrecte ne permettant pas de restituer la forme sonore). En appliquant cette classification aux verbes, nous obtenons les résultats suivants :

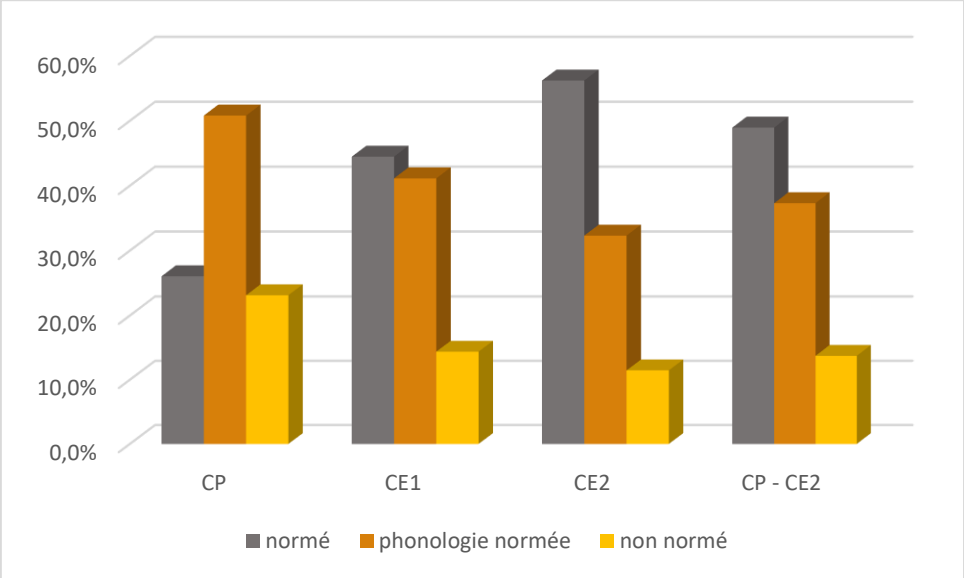


Figure 73 : Répartition des erreurs selon l'année d'apprentissage

À la fin du CP, 23 % des formes verbales écrites ne respectent pas la phonologie : on note donc une maitrise importante des formes phonologiques des verbes dès la fin de la première année. En revanche, l'orthographe des verbes est encore peu maitrisée dans la mesure où seuls 26 % des verbes sont correctement orthographiés. On note d'importants progrès en CE1 et en CE2 sur ce plan-là, puisqu'en fin de CE2 la majorité des verbes (56 %) sont correctement orthographiés. De plus, en fin de CE2, les formes non normées représentent moins de 12 % des verbes. Il y a donc une forte progression du CP au CE2, visible dans la figure 73 dans l'évolution des tailles des colonnes du CP au CE2.

[106] Pour cette étude, nous avons choisi de regrouper sous l'appellation *formes non normées à phonologie normée* à la fois les formes dont la phonologie est identique à celle de la forme normalisée et les formes dont l'archiphonologie (cf. chapitre 8 et chapitre 10 pour une explication de cette notion) est identique à celle de la forme normalisée. Ainsi, sont comptabilisés à l'identique les couples de formes « mangé » / *manger* et « mangeais » / *manger*.

Pour plus de détails, il nous est également possible d'obtenir une classification des verbes des cinq tiroirs verbaux qui nous intéressent. Les premières lignes du tableau 64 reprennent les données exposées dans la figure précédente (Figure 73).

Niveau		CP	CE1	CE2
Nb formes verbales		921	2792	4638
Ensemble des verbes	normé	26,05 %	44,48 %	56,21 %
	phonologie normée	50,81 %	41,14 %	32,32 %
	non normé	23,14 %	14,38 %	11,47 %
PRESENT	normé	30,30 %	52,86 %	64,64 %
	phonologie normée	49,04 %	37,40 %	27,38 %
	non normé	20,67 %	9,45 %	7,98 %
INFINITIF	normé	30,00 %	49,29 %	63,20 %
	phonologie normée	41,67 %	34,28 %	25,08 %
	non normé	28,33 %	16,43 %	11,72 %
PASSE SIMPLE	normé	26,76 %	45,24 %	50,25 %
	phonologie normée	54,93 %	36,61 %	29,98 %
	non normé	18,31 %	18,15 %	19,77 %
PART. PASSE	normé	18,68 %	35,14 %	43,32 %
	phonologie normée	54,95 %	44,20 %	42,18 %
	non normé	26,37 %	20,65 %	14,50 %
IMPARFAIT	normé	11,38 %	34,07 %	50,82 %
	phonologie normée	60,98 %	50,07 %	39,21 %
	non normé	27,64 %	15,86 %	9,97 %

Tableau 65 : Répartition des échecs et des réussites pour chaque tiroir verbal étudié[107]

On note une amélioration substantielle du CP au CE2 pour l'ensemble des tiroirs verbaux étudiés, de 26 % à 56 % de réussite. Pour chacun d'eux, les élèves progressent clairement. En CP, les deux tiroirs verbaux les mieux réussis sont le présent et l'infinitif, pour lesquels on atteint des scores de formes normées de 30 % ; en CE2, ce sont encore le présent et l'infinitif qui se détachent avec des scores dépassant respectivement 63 % et 64 %. Pour l'imparfait et le passé simple, près de la moitié des verbes présentent encore des erreurs orthographiques ou phonologiques, le participe passé étant un peu en retrait (43,3 % de formes normées).

[107] Les formes normées correspondent aux formes correctement orthographiées.

Au niveau de la progression du CP au CE2, l'imparfait se dégage nettement avec une marge de progression de +40 % des formes normées, contre une progression de 30 % en moyenne pour les autres temps. À l'inverse, le participe passé et le passé simple progressent de manière moins conséquente, avec une évolution respectivement de 24,6 % et 23,5 %.

3. Distinction des erreurs sur la base et sur la désinence

Enfin, il nous a paru intéressant de distinguer les erreurs qui portent sur les bases verbales des erreurs qui portent sur les désinences[108]. En se fondant sur le modèle théorique présenté à la section 1, cet algorithme permet pour toute forme verbale de distinguer sa désinence de sa base et de réaliser un alignement entre les bases, et plus particulièrement les désinences, des segments transcrits et des formes normalisées. Il permet ainsi de déterminer le ou les lieux des erreurs (base ou désinence). Les erreurs présentent sur les bases relèvent davantage de problème orthographiques, tandis que les erreurs présentent sur les désinences correspondent le plus souvent à des erreurs de morphologie verbale. Cet algorithme n'a pas encore pu être évalué, mais il permet tout de même de faire émerger certains phénomènes intéressants.

Les graphiques suivants permettent de visualiser la proportion de formes normées (en foncé), la proportion de formes présentant une erreur seulement sur la base, celle présentant une erreur et sur la base et sur la désinence et enfin celles avec une erreur seulement sur la désinence.

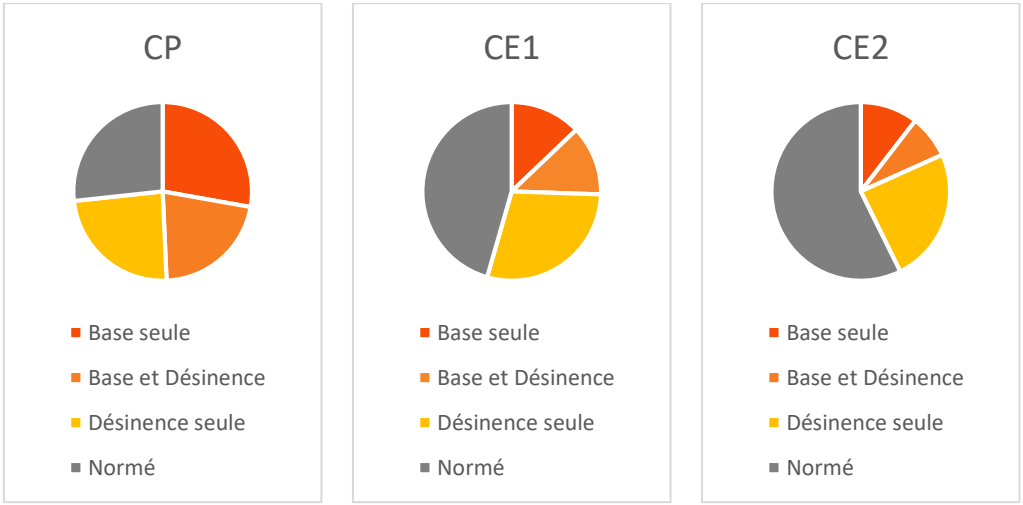


Figure 74 : Répartition des erreurs sur les bases et les désinences des formes verbales[109]

[108] Pour ce faire, le développement d'un algorithme spécifique, appelé *AliScol_Dés*, a été nécessaire. Cet outil a été développé par Claude Ponton, membre du projet *Scoledit*.

[109] Lecture des graphiques : le nombre d'erreurs sur les bases est donné par l'addition des erreurs sur la base seule et des erreurs à la fois sur la base et la désinence.

Il apparait sur ces graphiques qu'en classe de CP, les pourcentages de formes produites comportant une erreur sur les bases et les pourcentages de formes produites comportant une erreur sur les désinences sont proches. En effet, on observe 49,3 % d'erreurs sur les bases (donnée calculée en additionnant le pourcentage de formes comportant des erreurs sur leur base uniquement et le pourcentage de formes présentant à la fois des erreurs sur leur base et sur leur désinence), alors qu'on observe 45,5 % d'erreurs sur les désinences (donnée calculée de la même manière). Ce n'est plus le cas en classe de CE2, où le pourcentage d'erreurs sur les bases a sensiblement diminué pour n'être plus que de 18,3 %, tandis que le pourcentage d'erreurs sur les désinences est de 32,3 %. Les erreurs sur les bases diminuent donc plus fortement que les erreurs sur les désinences.

3.1. Le cas de l'imparfait

Nous proposons de nous pencher plus spécifiquement sur l'imparfait de l'indicatif (Figure 75), tiroir verbal quelque peu atypique dans la mesure où il est peu réussi en CP (11,4 %) et réussi à environ 50 % en CE2.

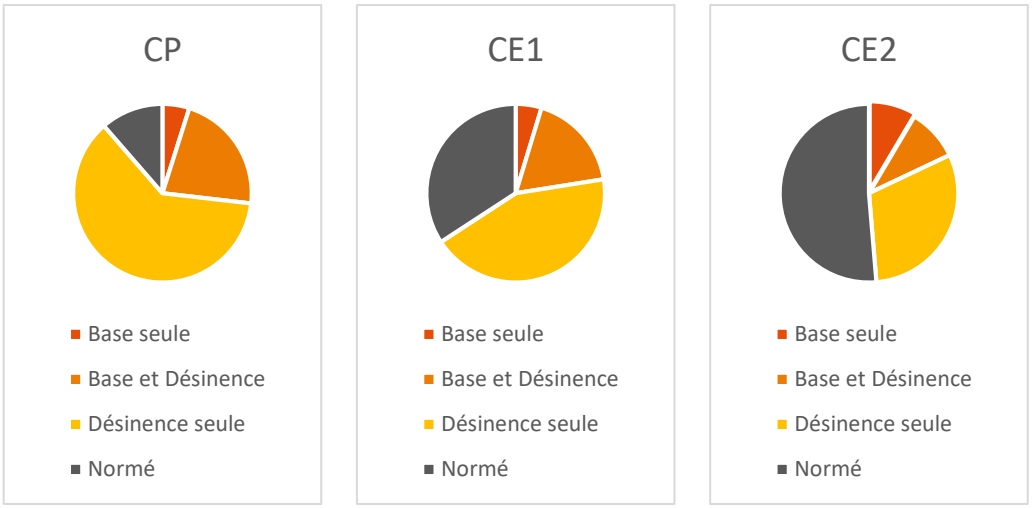


Figure 75 : Répartition des erreurs sur les bases et les désinences des formes verbales à l'imparfait

Globalement, ce tiroir verbal fait l'objet d'un grand nombre de formes présentant un problème de désinence et d'un nombre plus restreint de problèmes de base. En CP, par exemple, environ une forme sur quatre comporte une erreur sur la base (« alait » à la place de « allait » alors que l'ensemble des formes verbales en CP est affectée à près de 50 % (Figure 75). Cette tendance est moins nette en CE2, les verbes à l'imparfait présentant des proportions de formes avec problème sur la base proche de la moyenne des tiroirs verbaux.

L'imparfait est un tiroir verbal particulièrement notable du point de vue des erreurs de désinence. Alors que le pourcentage de mots verbaux comportant au moins une erreur de

désinence au CP est de 45,5 %, il est de 83,7 % pour les verbes à l'imparfait (*voulé* à la place de *voulait*), ce qui représente un écart conséquent. Cet écart a tendance à s'amenuiser dans les années suivantes, puisqu'en CE2, le pourcentage de mots verbaux comportant au moins une erreur est 32,3 % pour l'ensemble des tiroirs verbaux et de 40,2 % pour l'imparfait. Les désinences représentent donc un enjeu particulier dans l'apprentissage de l'imparfait.

3.2. Le cas du passé simple

Penchons-nous à présent sur un autre tiroir verbal moyennement réussi en CE2, le passé simple (Figure 76).

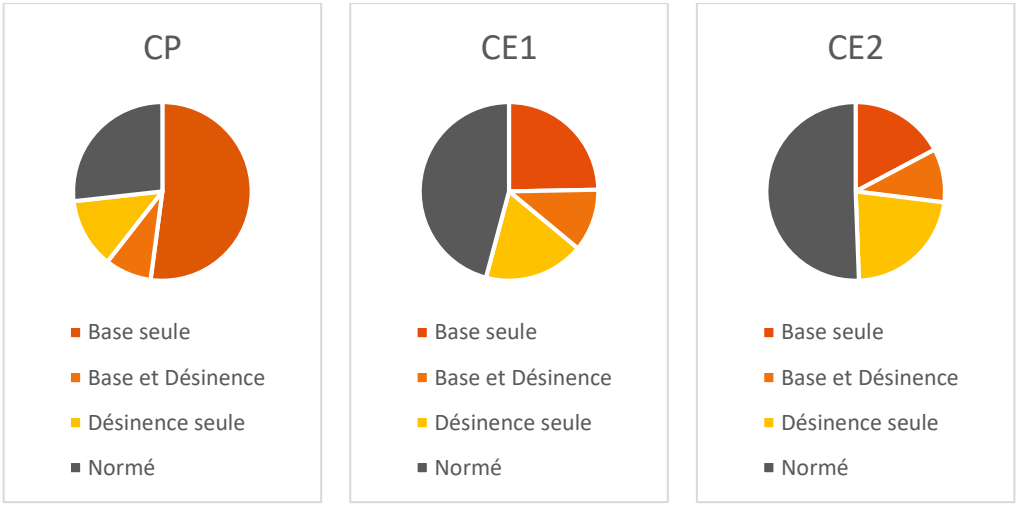


Figure 76 : Répartition des erreurs sur les bases et les désinences des formes verbales au passé simple

Alors que l'imparfait présentait un grand nombre d'erreurs de désinence, le passé simple est un temps qui concentre plus d'erreurs sur les bases, par exemple « rolacha » à la place de « relâcha » (élève 864) ou « prena » pour la forme attendue « prit » (élève 584, CP) qui comporte à la fois une erreur sur la base et sur la désinence. Au CP, 60,6 % des verbes au passé simple présentent des erreurs sur leur base. Par comparaison, on en dénombre 49,3 % sur l'ensemble des verbes aux temps considérés.

Il peut être intéressant de noter également que la progression des réussites est moins conséquente pour ce tiroir verbal que pour les autres tiroirs verbaux. En effet, sur l'ensemble des tiroirs verbaux considérés, on note une progression moyenne de plus de 30 %, de 26,7 % en CP à 57,3 % en CE2, tandis que la progression pour les verbes au passé simple est inférieure à 24 %, de 26,8 % en CP à 50,6 % en CE2.

Au vu de ces résultats, il semble que, plus que pour tout autre tiroir verbal, la formation de la base des verbes au passé simple soit particulièrement délicate pour les apprenants. Les erreurs relevées ici correspondent à la fois à des erreurs orthographiques et morphologiques,

puisqu'elles impacte souvent la formation de la base verbale qui peut être différente de celle du présent (exemple : « pren – a » au lieu de *pri – t*)[110].

4. Conclusion

Les résultats présentés ici, issus d'un premier traitement outillé par le TAL des formes verbales dans 903 textes d'élèves du CP au CE2, nous renseignent sur les zones de réussite et les points d'achoppement lors de l'apprentissage de la morphologie verbale. Grâce aux outils développés, ces résultats nous renseignent aussi sur la façon dont ce domaine peut outiller l'exploration linguistique et didactique de l'apprentissage de l'écriture.

Ces premiers résultats permettent également de dégager plusieurs pistes didactiques. La première concerne les tiroirs verbaux à travailler en priorité si l'on demande aux élèves de produire des récits. En CP, mais aussi en CE1 et en CE2, c'est le présent qui domine dans les usages examinés. Si 62 % des verbes utilisés en CP sont au présent (et cela peut être l'effet de la consigne, l'histoire du petit chat pouvant se transformer en une espèce de description des quatre images séquentielles proposées aux élèves), ce sont 35 % et 37 % des verbes qui sont respectivement au présent pour le CE1 et le CE2. Un premier examen des erreurs tend à montrer que les marques de personne non nécessaires à la phonographie (le *e* de *pleure vs* le *e* de *chante,* où la présence du *e* final influe sur la prononciation du *t*) pourraient faire l'objet d'un travail approfondi, en même temps que les règles phonographiques continuent à être travaillées au CP et au CE1. Les verbes en –er comme *pleurer* ou *tomber* sont en effet très employés dans les productions de CP de notre corpus.

En CE1 et en CE2, les élèves utilisent davantage les temps du passé pour raconter leur histoire. L'imparfait est le 2e tiroir verbal le plus employé, loin devant le passé simple. Il fait dès le CP, mais aussi en CE1 et en CE2, l'objet d'un grand nombre d'erreurs de désinence, qu'il conviendrait de sérier. Dans quel(s) jeu(x) d'oppositions homophoniques-hétérographiques entre l'imparfait en CP en production de texte ? et en CE2 ? Qu'est-il raisonnable, ou rentable, de travailler en CP et au CE de ce secteur des formes verbales en /E/ ? Des investigations plus poussées devraient permettre de répondre à cette question et ainsi de mieux utiliser le temps scolaire.

Parallèlement à l'augmentation du taux de réussite du CP au CE2, le nombre d'erreurs de transcription de la chaine sonore (formes non normées) diminue du CP au CE2, ainsi que celles qui respectent la phonologie de la forme attendue. Mais la proportion de formes non normées

[110] Le tiret marque la séparation entre la base et la désinence, telle qu'elle est calculée par l'algorithme.

est à un niveau élevé au CP (Figure 73). La piste didactique qui en découle est l'appui sur l'oral, pour ne pas dire la correction orale, au moment où les élèves écrivent.

En CE2, après le présent et l'imparfait, on enregistre autant d'occurrences d'infinitifs, de passés simples et participes passés. Globalement, au passé simple près, qui n'apparait pas comme surmobilisé par les élèves, les tiroirs verbaux les plus utilisés à l'écrit sont aussi ceux mis en évidence par J.-P. Adam et C. Blanche-Benveniste dans les corpus oraux (Adam, 1999). Il y a sans doute là matière à réflexion.

Si dans la dernière partie nous nous sommes concentrés uniquement sur l'imparfait et le passé simple, il est possible de poursuivre cette étude sur l'ensemble des tiroirs verbaux. De plus, coupler cet outil à l'outil d'alignement en graphèmes *AliScol_Graph* présenté au chapitre précédent devrait permettre d'analyser plus finement les erreurs pour percevoir plus précisément les points de tension et les sources d'erreurs.

Conclusion et perspectives

1. Travaux réalisés au cours de la thèse 239
2. Apports des travaux réalisés 240
3. Perspectives d'amélioration des ressources développées 240
4. Perspectives d'exploitation linguistique et didactique des ressources développées 242

1. Travaux réalisés au cours de la thèse

Cette thèse s'inscrit dans un contexte d'émergence de corpus d'écrits scolaires de plus en plus grands. Ce changement d'échelle s'accompagne d'un besoin d'outils spécifiques à ce type d'écrits. Par ailleurs, malgré le nombre croissant de corpus d'écrits scolaires disponibles, peu de corpus permettent d'approcher la notion de progressivité. Des corpus longitudinaux, c'est-à-dire réalisant le suivi d'une cohorte d'élèves à partir de conditions de recueil similaires année après année, sont nécessaires.

Face à ces constats, nous avons émis l'hypothèse que le champ du traitement automatique des langues et les méthodes qui y sont développées peuvent aider au développement et à l'exploitation des corpus d'écrits scolaires.

Pour répondre à ces nouveaux enjeux, nous avons donc adopté une méthodologie comprenant différentes étapes.

La première étape est la constitution d'un corpus d'écrits scolaires longitudinal de grande taille (plusieurs milliers de productions). Ce corpus contient des dictées et des textes produits par des élèves suivis durant toute leur scolarité à l'école primaire (CP-CM2). Le recueil ayant pris appui sur l'étude « Lire – Écrire au CP », la phase de réflexion sur le protocole s'en est trouvée limitée. Néanmoins, nous avons fait le choix de garder la consigne donnée en CE1 jusque dans les classes supérieures et de reproduire des conditions de recueil similaires. La constitution de ce corpus inclut à la fois son recueil, la numérisation et la transcription des productions, l'annotation des données linguistiques et la diffusion de la ressource ainsi constituée.

La deuxième étape consiste en l'élaboration d'une méthodologie d'exploration des écrits scolaires et la création d'outils permettant la mise en œuvre de cette méthodologie. Ces outils sont conçus pour être adaptés à ce type d'écrits et particulièrement aux écrits du corpus élaboré lors de l'étape 1. La méthodologie ainsi construite est appelée *approche par*

comparaison. Il s'agit de comparer des éléments transcrits (issus de la numérisation des production) et des éléments normalisés (issus de la normalisation des productions) de même niveau (syntaxique, lexical, morphologique, etc.) en vue de les analyser. Cette méthode nécessite à la fois la conception d'une normalisation des éléments comparés, conçue manuellement dans le travail de thèse que nous venons de présenter et vue comme une annotation déportée, et des outils de segmentation et d'alignement. Dans cette thèse, nous présentons l'adaptation de cette méthode à trois niveaux, à savoir les formes, les graphèmes et la morphographie verbale. Pour chacun de ces niveaux, le développement d'outils spécifiques a été nécessaire.

2. Apports des travaux réalisés

Le travail réalisé dans cette thèse participe à l'élaboration d'une ressource qui contribue au champ de la linguistique de corpus et qui est en cours de diffusion pour les communautés scientifiques et enseignantes. De plus, cette ressource contribue à pallier certains manques, à savoir le besoin d'un ensemble de données linguistiques numérisées suffisamment conséquent et le besoin de ressources permettant l'étude de la progressivité de la production de textes.

De plus, la méthode et les outils proposés au cours de cette thèse permettent une première étude de l'apport de l'introduction d'outils de traitement automatique des langues dans le champ des corpus d'écrits scolaires. Ils permettent de faire le lien avec des outils de TAL déjà existants mais peu adaptés à ce type d'écrits, comme les étiqueteurs morphosyntaxiques, et devraient permettre, dans un second temps, d'utiliser des outils tels que des concordanciers ou des outils d'analyses syntaxiques.

Bien que spécifiquement conçus pour la ressource élaborée au sein du projet *Scoledit* dans lequel s'inscrit cette thèse, les outils développés sont adaptables à d'autres ressources. En effet, ils s'appuient sur une normalisation, relativement spécifique au projet *Scoledit* par comparaison aux autres projets d'élaboration de corpus d'écrits scolaires. Néanmoins, il est relativement aisé de produire une version similaire à partir du formalisme d'annotation (balises contenant les formes normées) utilisé dans la plupart de ces projets (Boré & Elalouf, 2017 ; Doquet *et al.*, 2017, par exemple).

3. Perspectives d'amélioration des ressources développées

3.1. Le corpus

Le corpus présenté est encore en cours d'élaboration. En effet, bien que le recueil soit terminé depuis juillet 2018, la transcription et la normalisation des productions recueillies est toujours

en cours (cf. Tableau 65). À ce sujet, la priorité est accordée à la vérification des normalisations du corpus longitudinal de CP à CM1 (37 élèves concernés, soit 1 492 productions), puis à la transcription et la normalisation des productions de CM2 du corpus longitudinal de CP à CM2, afin de disposer de corpus longitudinaux exploitables pour réaliser des analyses linguistiques et didactiques. Dans un second temps, le même traitement pourra être appliqué aux autres productions et aux dictées.

	CP		*CE1*		*CE2*		*CM1*		*CM2*	
	C	*L*	*C*	*L*	*C*	*L*	*C*	*L*	*C*	*L*
Productions	975	373	743	373	415	373	420	373	334	373
Transcrits	970 (48 % val.)	373 (100 % val.)	738 (56 % val.)	373 (100 % val.)	415 (96 % val.)	373 (100 % val.)	419 (31 % val.)	373 (35 % val.)	334 (16 % val.)	332 (15 % val.)
Normalisés	717 (58 % val.)	373 (100 % val.)	540 (75 % val.)	373 (100 % val.)	410 (97 % val.)	373 (100 % val.)	388 (75 % val.)	373 (79 % val.)	331 (4 % val.)	332 (4 % val.)

C : corpus Complet
L : corpus longitudinal
val. : productions validées

Tableau 66 : Tableau d'avancement de la numérisation et l'enrichissement du corpus

À l'heure actuelle, seules les productions de CP et de CE1 sont visibles sur le site de visualisation initialement prévu (http://scoledit.org/scoledit/). Si la ressource n'est pas encore téléchargeable, les productions sont néanmoins consultables sur le site de travail (http://scoledit.org/scoledition/). Un travail sur les outils de diffusion est encore nécessaire pour rendre la ressource élaborée accessible plus largement.

3.2. Les outils développés

L'aligneur *AliScol*, mis au point dans ce travail, présente encore une marge de progression. Par exemple, certains phénomènes comme les phénomènes d'hyposegmentation et d'hypersegmentation ne sont pas pris en compte. Une étude plus détaillée sur ces phénomènes devrait permettre de développer l'algorithme dans ce sens. En outre, cet aligneur s'appuie sur plusieurs outils préexistants, comme *LIA-PHON* et *TreeTagger* ; il serait intéressant de mesurer les performances de ces outils par rapport à d'autres outils pour les écrits scolaires. L'outil *Talismane*, par exemple, est utilisé pour l'étiquetage de plusieurs corpus inclus dans le projet *E-CALM*. I. Falk et ses collègues (2014) répertorient un certain nombre d'outils d'étiquetage morphosyntaxique du français capable de s'adapter aux mots nouveaux. Une comparaison

entre les performances de l'outil *TreeTagger* et de ces outils pourrait donc être pertinente en vue d'améliorer l'outil *AliScol*. L'outil *TreeTagger* est un outil basé sur des mesures statistiques entrainées à partir de corpus. Ces corpus sont, par certains aspects, très différents des corpus d'écrits scolaires. Il pourrait donc être intéressant de réentrainer spécifiquement cet outil ou un autre plus adapté pour qu'il s'adapte aux écrits scolaires.

Le développement de l'outil *AliScol_Graph* a fait l'objet de moins d'explications détaillées ; il est également largement perfectible. Il est possible de poursuivre la réflexion en ce sens.

Une première évaluation des outils réalisés a été effectuée, néanmoins celle-ci présente des approximations et biais méthodologiques. Une évaluation plus approfondie est nécessaire, mais ne pourra se faire que lorsque des données en quantité suffisante auront gagné une certaine stabilité. De plus, l'un des enjeux de ce travail de thèse était la réalisation d'outils de traitement automatique facilitant l'exploitation de corpus d'écrits scolaires. S'il est possible d'affirmer que cet objectif est en partie atteint pour le corpus *Scoledit*, rien ne permet de dire qu'il l'est pour les corpus d'écrits scolaires en général sans une évaluation de ces outils à partir de données d'autres corpus. Cette évaluation n'a pu être menée pour le moment.

Suite à une collaboration franco-italienne, un premier test d'adaptation de l'outil réalisé à des données en italien est en cours et semble plutôt fructueux. Un test d'adaptation à des données en espagnol est également prévu.

3.3. Développement d'un outil de normalisation

La phase de normalisation des productions, nécessaire à l'approche adoptée dans le projet *Scoledit*, est pour l'heure entièrement manuelle et nécessite un temps très long. Ce travail peut probablement être en partie automatisé grâce à la conception d'un outil spécifique. Jusqu'à présent, la conception d'un tel outil n'était pas possible en raison du manque de données relatives aux phénomènes et aux erreurs spécifiques aux apprenants d'école primaire. Grâce aux ressources produites dans le cadre de ce travail de thèse, un tel travail parait maintenant envisageable. Plus particulièrement, l'extraction d'une liste de correspondances segments erronés / formes normalisées devrait permettre d'identifier un certain nombre de transformations récurrentes.

4. Perspectives d'exploitation linguistique et didactique des ressources développées

Le travail présenté a permis principalement l'élaboration de deux types de ressources : un corpus longitudinal d'écrits scolaires et des outils d'exploration des corpus d'écrits scolaires.

Comme nous l'avons montré dans la partie 3, l'utilisation de ces ressources devrait permettre de réaliser une description linguistique des productions d'apprenants en vue de leur exploitation didactique. Différents niveaux linguistiques peuvent être enquêtés grâce à ces outils. Nous nous contenterons ici de donner trois exemples : l'étude de la segmentation en mots, l'étude de la segmentation phrastique et l'étude de l'orthographe. Quelques pistes d'étude sont proposées dans cette thèse, mais un travail beaucoup plus conséquent est nécessaire pour pouvoir en extraire des pistes d'ordre didactique.

Au vu des spécificités des corpus scolaires, il parait intéressant d'étudier la segmentation en mots et plus particulièrement les phénomènes d'hyposegmentation et d'hypersegmentation. Nous avons d'ores et déjà pu identifier que les premiers étaient bien plus nombreux que les seconds. Néanmoins, il serait intéressant d'avoir une meilleure connaissance de leur contexte d'apparition afin de déterminer quels types de mots sont impactés. Cette connaissance permettrait d'identifier des pistes de travail didactique spécifique. Les données calculées sur les phénomènes d'hyposegmentation des formes élidées, où la présence d'une apostrophe génère des phénomènes d'hyposegmentation, a permis de faire émerger une première piste.

De la même façon, une description fine de la segmentation phrastique (phrases ou propositions) devrait permettre d'identifier les marqueurs de segmentation (ponctuation, connecteurs, etc.) utilisés par les apprenants. Dans ce cas précis, une étude longitudinale de l'apparition de ces marqueurs pourrait s'avérer particulièrement intéressante dans la mesure où les stratégies de segmentation employées par les apprenants sont très différentes d'un niveau scolaire à l'autre. Par exemple, le point est utilisé dès la classe de CP par certains élèves, tandis que la virgule n'apparait qu'à partir du CE1 ou du CE2. Etudier l'ordre d'apparition des signes de ponctuation permettrait donc de mieux cibler l'année et l'ordre d'étude en classe.

Une étude plus approfondie de l'orthographe pourrait également être menée, elle permettrait d'étudier les différentes stratégies d'écriture mises en œuvre par les apprenants pour orthographier des formes méconnues ou mal-connues. Le début de ce travail a pu être amorcé à travers l'étude de la morphographie verbale mais cela ne représente qu'une part minime du travail de description à fournir. De telles connaissances permettraient de mieux accompagner l'apprentissage de l'orthographe.

De plus, diverses études ont été menées pour étudier l'orthographe à partir de tests standardisés (cf. Brissaud, Fisher, & Negro, 2012 pour un exemple d'étude des finales verbales en /e/ - /ɛ/ à partir de ce type de tests). Les ressources que nous avons constituées devraient permettre de confronter les résultats obtenus dans ces études à ce que réalisent les élèves lors de tâches de production de texte.

Pour conclure, les études menées dans cette thèse ont pu montrer que le traitement automatique des langues pouvait représenter un apport pour la constitution d'écrits scolaires. Ce travail nécessite d'être poursuivi pour que cet apport soit effectif. Il a également permis d'appuyer la constitution d'une ressource qui devrait être accessible aux chercheurs et enseignants qui le souhaitent.

Bibliographie

Abeille, A., Clement, L., Kinyon, A., & Toussenel, F. (2001). Un corpus français arboré : Quelques interrogations. TALN 2001 Recital 2001 (Tours, 2-5 juillet 2001).

Adam, J.-P. (1999). La conjugaison des verbes : Virtuelle, attestée, défective. *Recherches sur le français parlé*, *15*, 87–112.

Adamson, G. W., & Boreham, J. (1974). The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Information storage and retrieval*, *10*(7-8), 253–260.

Ahrenberg, L., Andersson, M., & Merkel, M. (2000). A knowledge-lite approach to word alignment. In *Parallel Text Processing* (p. 97–116). New York: Springer.

Allauzen, A., & Schütze, H. (2018). Apprentissage profond pour le traitement automatique des langues. *Traitement Automatique des Langues*, 2(9), 8.

Almeida, J. J., Santos, A., & Simões, A. (2010). Bigorna : A toolkit for orthography migration challenges. *Seventh International Conference on Language Resources and Evaluation (LREC2010)*, 227–232. Luxembourg: European Language Resources Association (ELRA).

Anis, J. (2003). Communication électronique scripturale et formes langagières. *Actes des Quatrièmes rencontres Réseaux humains/Réseaux technologiques*, *31*.

Antoniadis, G., Echinard, S., Kraif, O., Lebarbé, T., Loiseau, M., & Ponton, C. (2004). NLP-based Scripting for CALL Activities. *Proceedings of the Workshop on eLearning for Computational Linguistics and Computational Linguistics for eLearning*, 18–25. Stroudsburg, PA, USA: Association for Computational Linguistics.

Antoniadis, G., Ponton, C., & Zampa, V. (2007). De la nécessité du TAL dans les EIAH en langues Les cas EXXELANT et MIRTO. *Environnements informatiques pour l'apprentissage humain*, 251-256. Lausanne: INRP.

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. Computational Linguistics, 34(4), 555–596.

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus Design Criteria. *Literary and Linguistic Computing*, *7*(1), 1-16.

Axelsson, M. W. (2000). USE-the Uppsala Student English corpus : An instrument for needs analysis. *ICAME journal*, *24*, 155–157.

Bachschmidt, P. (1997). Procédure de constitution d'un corpus attesté d'articles de recherche scientifique en vue d'une étude contrastive. ASp. la revue du GERAS, 15-18, 133-138.

Bahl, L., & Jelinek, F. (1975). Decoding for channels with insertions, deletions, and substitutions with applications to speech recognition. *IEEE Transactions on Information Theory*, *21*(4), 404–411.

Banerji, N., Gupta, V., Kilgarriff, A., & Tugwell, D. (2013). Oxford Children's Corpus : A corpus of children's writing, reading, and education. *Corpus Linguistics 2013*, 315-317. Lancaster: UCREL.

Baron, A., Rayson, P., Greenwood, P., Walkerdine, J., & Rashid, A. (2012). Children Online : A survey of child language and CMC corpora. *International Journal of Corpus Linguistics*, *17*(4), 443-481.

Beaufort, R. (2008). *Application des Machines à Etats Finis en Synthèse de la Parole* (Thèse de Doctorat). CEA, Saclay, France.

Beaufort, R. (2010). Composition filtrée et marqueurs de règles de réécriture pour une distance d'édition flexible. Application à la correction des mots hors-vocabulaire. *Traitement Automatique des Langues (TAL)*, *51*(1), 11–40.

Beaufort, R., & Roekhaut, S. (2011). Automation of dictation exercises. A working combination of CALL and NLP. *Computational Linguistics in the Netherlands Journal*, *1*, 2-20.

Beaufort, R., Roekhaut, S., Cougnon, L.-A., & Fairon, C. (2010a). A hybrid rule/model-based finite-state framework for normalizing SMS messages. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 770–779. Association for Computational Linguistics.

Beaufort, R., Roekhaut, S., Cougnon, L.-A., & Fairon, C. (2010b). Une approche hybride traduction/correction pour la normalisation des SMS. *Actes de la 17e conférence sur le traitement automatique des langues naturelles (TALN'10)*, 19–23. Montréal: École polytechnique de Montréal, Université de Montréal.

Beaufort, R., Roekhaut, S., & Fairon, C. (2008). Définition d'un système d'alignement SMS/français standard à l'aide d'un filtre de composition. *Proceedings of JADT 2008*, 155–166. Lyon: ENS Lettres et sciences humaines.

Benchiheub, M.-F., Berret, B., & Braffort, A. (2016). Collecting and Analysing a Motion-Capture Corpus of French Sign Language. *Workshop on the Representation and Processing of Sign Languages (LREC 2016)*. Présenté à Portoroz, Slovenia. Portoroz, Slovenia: European Language Resources Association (ELRA).

Berkling, K. (2001). SCoPE, syllable core and periphery evaluation : Automatic syllabification and foreign accent identification. *Speech Communication*, *35*(1-2), 125–138.

Berkling, K. (2016). Corpus for Children's Writing with Enhanced Output for Specific Spelling Patterns (2nd and 3rd Grade). *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 3200–3206. Portorož, Slovenia: European Language Resources Association (ELRA).

Berkling, K. (2018). A 2nd Longitudinal Corpus for Children's Writing with Enhanced Output for Specific Spelling Patterns. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, 2262-2268. Miyazaki, Japan: European Language Resources Association (ELRA).

Berkling, K., Fay, J., & Stuker, S. (2011). Speech Technology-based Framework for Quantitative Analysis of German Spelling Errors in Freely Composed Children's Texts. *Speech and Language Technology in Education*, 4.

Bernhard, D., & Steiblé, L. (2015). Quand l'oral se fait entendre à l'écrit : Alignement de lexiques en l'absence de normalisation graphique. *TALaRE 2015 - Traitement Automatique des Langues Régionales de France et d'Europe (TALN'22)*. Présenté à Caen. Caen: Université de Caen Basse-Normandie.

Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B., & Rauzy, S. (2006). Le CID - Corpus of Interactional Data - : Protocoles, conventions, annotations. *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA)*, *25*, 25-55.

Bessonnat, D. (1991). Enseigner la... « ponctuation » ? (!). *Pratiques*, *70*(1), 9-45.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243-257.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics : Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bigi, B., & Le, V.-B. (2008). Normalisation et alignement de corpus français et vietnamiens : Format et Logiciels. *9es Journées internationales d'Analyse statistique des Données Textuelles*, *2008*, 9e. Lyon: ENS Lettres et sciences humaines.

Bilger, M. (2000). *Corpus : Méthodologie et applications linguistiques*. Paris: H. Champion.

Blanchard, A., Kraif, O., & Ponton, C. (2009). Mastering Overdetection and Underdetection in Learner-Answer Processing : Simple Techniques for Analysis and Diagnosis. *CALICO Journal*, *26*(3), 592-610. Consulté à l'adresse JSTOR.

Blanche-Benveniste, C. (2002). Structure et exploitation de la conjugaison des verbes en français contemporain. *Le français aujourd'hui*, *4*, 11–22.

Blanche-Benveniste, C., & Chervel, A. (1969). *L'orthographe:* Paris: Maspero.

Boch, F., Grossmann, F., & Rinck, F. (2007). Conformément à nos attentes... : Les marqueurs de convergence/divergence dans l'article de linguistique. *Revue française de linguistique appliquée*, *12*(2), 109–122.

Bonnet, C. (1994). *Plume en main... Ou L'itinéraire de l'élève qui apprend à écrire*. Centre vaudois de recherches pédagogiques.

Bonnet, C., & Gardes-Tamine, J. (1990). *L'enfant et l'écrit : Récits, poésies, correspondances, journaux intimes*. Paris: Armand Colin.

Boré, C. (2007a). Corpus et genres scolaires : Affinités, difficultés. *Le francais aujourd'hui*, *159*(4), 19-28.

Boré, C. (2007b). La métamorphose d'un genre : Quelques descripteurs pour un genre scolaire de récit. In *Diptyque*: *Vol. 10. Construire et exploiter des corpus de genres scolaires* (Namur). Presses Universitaires de Namur.

Boré, C. (2007c). Les genres scolaires comme corpus, construction d'une problématique. In *Diptyque*: *Vol. 10. Construire et exploiter des corpus de genres scolaires*. Namur: Presses Universitaires de Namur.

Boré, C., & Elalouf, M.-L. (2017). Deux étapes dans la construction de corpus scolaires : Problèmes récurrents et perspectives nouvelles. *Corpus*, *16*, 31‑64.

Bouillon, P. (1998). *Traitement automatique des langues naturelles* (Duculot AUPELF-UREF). Paris: Duculot AUPELF-UREF.

Brissaud, C. (2002). Travailler la morphologie écrite du verbe au collège. *Le français aujourd'hui*, *4*, 59–66.

Brissaud, C., Fisher, C., & Negro, I. (2012). The relation between spelling and pronunciation : The case of French and the phonological variation/e/∼/ɛ/in different French dialects. *Written Language & Literacy*, *15*(1), 46–64.

Brown, P. F., Della-Pietra, S. A., Della-Pietra, V. J., & Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation. *Computational Linguistics*, *19*(2), 263–313.

Cappeau, P., & Roubaud, M.-N. (2018). *Regards linguistiques sur les textes d'élèves : De 5 à 12 ans*. Clermont-Ferrand: Presses universitaires Blaise Pascal.

Castagnet-Caignec, S. (2018). Traitement du temps dans des récits à visée littéraire chez les élèves de primaire et secondaire. *Repères. Recherches en didactique du français langue maternelle*, (57), 35‑56.

Catach, N. (1978). *L'Orthographe*. Paris: PUF.

Catach, N. (1979). Le graphème. *Pratiques*, *25*(1), 21–32.

Catach, N. (1980). *L'orthographe française : Traité théorique et pratique avec des travaux d'application et leurs corrigés* (Vol. 3). Paris: Nathan.

Catach, N. (1995). *L'orthographe française : Traité théorique et pratique avec des travaux d'application et leurs corrigés* (3. éd). Paris: Nathan.

Cendrowska, J. (1987). PRISM: An algorithm for inducing modular rules. *International Journal of Man-Machine Studies*, *27*(4), 349–370.

Charolles, M. (1988). La gestion des risques de confusion entre personnages dans une tâche rédactionnelle. *Pratiques*, *60*(1), 75‑97.

Chipere, N., Malvern, D., & Richards, B. (2004). Using a corpus of children's writing to test a solution to the sample size problem affecting type-token ratios. In *Studies in Corpus Linguistics*: *Vol. 17*. *Corpora and language learners* (p. 139‑147). Amsterdam, Philadelphia: John Benjamins Publishing Company.

Choudhury, M., Saraf, R., Jain, V., Mukherjee, A., Sarkar, S., & Basu, A. (2007). Investigation and modeling of the structure of texting language. *International Journal of Document Analysis and Recognition (IJDAR)*, *10*(3‑4), 157‑174.

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51‑89.

Church, K. W., & Mercer, R. L. (1993). Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *Comput. Linguist.*, *19*(1), 1–24.

Clanché, P. (1988). *L'enfant écrivain : Génétique et symbolique du texte libre*. Prais : Paidos Le Centurion.

Clanché, P. (2010). *Anthropologie de l'écriture et pédagogie Freinet*. Caen: Presses universitaires de Caen.

Condamines, A. (2005). Linguistique de corpus et terminologie. *Langages*, (157), 36-47.

Corbin, P. (1980). *De la production des données en linguistique introspective.* 121-179.

Cori, M. (2008). Des méthodes de traitement automatique aux linguistiques fondées sur les corpus. *Langages*, *171*, 95-110.

Cori, M., & David, S. (2008). Les corpus fondent-ils une nouvelle linguistique ? *Langages*, *171*, 111-129.

Cori, M., David, S., & Léon, J. (2008). Présentation : Éléments de réflexion sur la place des corpus en linguistique. *Langages*, *171*, 5-11.

Cougnon, L.-A., & Beaufort, R. (2009). SSLD: a French SMS to standard language dictionary. In *Cahiers du Cental*: *Vol. 7. eLexicography in the 21st century : New applications, new challenges* (Vol. 1, p. 33-42). Louvain-La-Neuve: Presses universitaires de Louvain.

Crystal, D. (2006). *Language and the Internet*. Cambridge: Cambridge University Press.

Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-aided error analysis. *System*, *26*(2), 163-174.

Dahlmeier, D., Ng, H. T., & Wu, S. M. (2013). Building a large annotated corpus of learner English : The NUS corpus of learner English. *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, 22–31. Atlanta, Georgia: Joel Tetreault, Jill Burstein, Claudia Leacock.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, *7*(3), 171–176.

Damper, R. I., Marchand, Y., Marsters, J.-D., & Bazin, A. I. (1997). Aligning text and phonemes for speech technology applications using an EM-like algorithm. *Journal of Sol-Gel Science and Technology*, *8*(2), 147–160.

Dasigi, P., & Diab, M. (2011). Codact : Towards identifying orthographic variants in dialectal arabic. *Proceedings of the 5th International Joint Conference on Natural Language Processing*, 318–326. Chiang Mai, Thailand: AFNLP.

David, J. (2000). Étudier les textes d'enfants : Revue de travaux. In *Savoirs en Pratique*. *Apprendre à lire des textes aux enfants* (p. 274-288). Paris: De Boeck Supérieur.

David, J., & Doquet, C. (2016). Les écrits d'élèves : Un corpus de référence pour le français contemporain. *Actes du Congrès Mondial de Linguistique, juillet 2016*, *27*, 11001. EDP Sciences.

De Haan, P. (2000). Tagging non-native English with the TOSCA-ICLE tagger. In *Mair, Ch.; Hundt, M. (ed.), Corpus linguistics and linguistic theory* (p. 69-79). Amsterdam: Rodopi.

De Vogüé, S., Espinoza, N., Garcia, B., Perini, M., & Marzena Watorek, F. (2017). Constitution d'un grand corpus d'écrits émergents et novices : Principes et méthodes. *Corpus*, *16*, 65-86.

Debili, F., & Sammouda, E. (1992). Aligning sentences in bilingual texts : French-English and French-Arabic. *Proceedings of the 14th conference on Computational linguistics-Volume 2*, 517–524. Association for Computational Linguistics.

Díaz-Negrillo, A., & Domínguez, J. F. (2006). Error tagging systems for learner corpora. *Revista española de lingüística aplicada*, *19*, 83–102.

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, *26*(3), 297-302.

Doquet, C., & David, J. (2018). Collecter, interpréter, enseigner l'écriture. Analyses linguistiques des écrits d'élèves. *Repères. Recherches en didactique du français langue maternelle*, *57*, 7-14.

Doquet, C., Enoiu, V., Fleury, S., & Maziotti, S. (2017). Problèmes posés par la transcription et l'annotation d'écrits d'élèves. *Corpus*, *16*, 133-156.

Dubois, J. (1967). *Grammaire structurale du français : Le verbe* (Vol. 2). Paris: Librairie Larousse.

Ducard, D., Honvault, R., & Jaffré, J.-P. (1995). *L'orthographe en trois dimensions*. Paris: Nathan pédagogie.

El Kaladi, A. (2007). Corpus et usages de corpus. *Les corpus en linguistique et en traductologie*, 33–47.

Elalouf, M.-L. (2005). *Écrire entre 10 et 14 ans un corpus, des analyses, des repères pour la formation*. Paris: Canopé - CRDP de Versailles.

Elalouf, M.-L. (2011). Constitution de corpus scolaires et universitaires : Vers un changement d'échelle ? *Pratiques. Linguistique, littérature, didactique*, *149-150*, 56-70.

Elalouf, M.-L., & Boré, C. (2007). Construction et exploitation de corpus d'écrits scolaires. *Revue francaise de linguistique appliquee*, *12*(1), 53-70.

Ernst-Gerlach, A., & Fuhr, N. (2006). Generating search term variants for text collections with historic spellings. *European Conference on Information Retrieval*, 49–60. Berlin: Springer.

Fabre, C. (1990). *Les brouillons d'écoliers ou l'entrée dans l'écriture*. Grenoble: Ceditel.

Fabre-Cols, C. (2000). *Apprendre à lire des textes d'enfants*. Bruxelles: De Boeck Supérieur.

Fairon, C., Klein, J.-R., & Paumier, S. (2006). *Le langage SMS : Étude d'un corpus informatisé à partir de l'enquête « Faites don de vos SMS à la science »*. Louvain-la-Neuve: UCL, Presses Univ. de Louvain.

Falk, I., Bernhard, D., Gérard, C., & Potier-Ferry, R. (2014). Étiquetage morpho-syntaxique pour des mots nouveaux. *1ème conférence sur le Traitement Automatique des Langues Naturelles, Jul 2014*. Présenté à Traitement Automatique des Langues Naturelles, Marseille, France.

Fayol, M. (2013). *L'acquisition de l'écrit*. Paris: Presses universitaires de France.

Fayol, M., & Jaffré, J.-P. (2008). *Orthographier*. Paris: Presses Universitaires de France.

Fayol, M., & Jaffré, J.-P. (2014). *L'orthographe* (1. éd). Paris: Presses Univ. de France.

Fenoglio, I. (2007). EDITE, un programme de génétique textuelle : Outil pour la reconnaissance de genres textuels et de styles d'auteurs. In *Dyptique*: *Vol. 10*. *Construire et exploiter des corpus de genres scolaires* (p. 109-123).

Filali, K., & Bilmes, J. (2005). A dynamic Bayesian framework to model context and memory in edit distance learning : An application to pronunciation classification. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 338–345. Ann Arbor, Michigan: Association for Computational Linguistics.

Fillmore, C. J. (1992). Corpus Linguistics' or'Computer-aided armchair linguistics'. In *Trends in Linguistics. Studies and Monographs*: *Vol. 65*. *Directions in corpus linguistics. Proceedings of Nobel Symposium* (p. 35-60). Berlin, New York: Mouton de Gruyter.

Flintham, R. (1995). Les relatifs which et that dans un corpus journalistique. *Cahiers Charles V*, 19(1), 139-165.

Fuchs, C. (Éd.). (1993). *Linguistique et traitements automatiques des langues*. Paris: Hachette Supérieur.

Gadet, F. (2017). *Les parlers jeunes dans l'Ile de France multiculturelle*. Paris: Ophrys.

Ganascia, J.-G., Fenoglio, I., & Lebrave, J.-L. (2004). Manuscrits, genèse et documents numérisés. *Document numerique*, *Vol. 8*(4), 91-110.

Garcia-Debanc, C. (2015). Le statut des textes d'élèves dans les recherches en didactique du français langue première : Approche historique. In *Littérature, linguistique et didactique du français : Les travaux Pratiques d'André Petitjean* (p. 195-204).

Garcia-Debanc, C., & Bonnemaison, K. (2014). La gestion de la cohésion textuelle par des élèves de 11-12 ans : Réussites et difficultés. *Actes du 4e Congrès Mondial de Linguistique Française (CMLF 2014), Juillet 2014*, *8*, 961–976. Berlin, Allemagne.

Garcia-Debanc, C., Ho-Dac, L.-M., Bras, M., & Rebeyrolle, J. (2017). Vers l'annotation discursive de textes d'élèves. *Corpus*, *16*, 157-184.

Garside, R., Leech, G. N., & McEnery, T. (1997). *Corpus annotation : Linguistic information from computer text corpora*. New York: Taylor & Francis.

Ghienne, C. (2015). Corpus de brouillons d'élèves de CM2/6e : Du recueil à l'analyse. *Actes du ICODOC 2015 : Colloque Jeunes Chercheurs du Laboratoire ICAR*.

Goigoux, R. (2016). *Étude de l'influence des pratiques d'enseignement de la lecture et de l'écriture sur la qualité des premiers apprentissages. Paris : MEN-ESR, Lyon : ENS-Lyon.*

Granger, S. (2004). Computer Learner Corpus Research : Current Status and Future Prospects. In *Applied Corpus Linguistics* (p. 123-145). Amsterdam: Rodopi.

Granger, S. (2007). Corpus d'apprenants, annotation d'erreurs et ALAO : Une synergie prometteuse. *Cahiers de lexicologie: Revue internationale de lexicologie et lexicographie*, *91*, 117-132.

Granger, S. (2013). Error-tagged Learner Corpora and CALL : A Promising Synergy. *CALICO Journal*, *20*(3), 465-480.

Granger, S., & Meunier, F. (1994). Towards a grammar checker for learners of English. In *Creating and Using English Language Corpora*. Amsterdam: Rodopi.

Granger, S., Vandeventer, A., & Hamel, M.-J. (2001). Analyse des corpus d'apprenants pour l'ELAO base sur le TAL. *Traitement automatique des langues*, *42*(2), 609.

Grésillon, A. (1994). *Éléments de critique génétique*. Paris: CNRS éditions.

Guillot, C., Lavrentiev, A., & Marchello-Nizia, C. (2007). *La Base de Français Médiéval (BFM) : États et perspectives*. F. Steiner.

Guimier de Neef, E., Debeurme, A., & Park, J. (2007). TILT correcteur de SMS: évaluation et bilan quantitatif. *Actes de la 14e conférence sur le traitement automatique des langues naturelles (TALN)*, 123–132.

Guimier de Neef, E. G., & Fessard, S. (2007). Évaluation d'un système de transcription de SMS. *Proceedings of the 26th International Conference on Lexis and Grammar, Bonifacio, France*.

Gunnarsson-Largy, C., & Auriac-Slusarczyk, E. (2013). *Écriture et réécritures chez les élèves : Un seul corpus, divers genres discursifs et méthodologies d'analyse*. Louvain-La-Neuve: Academia L'Harmattan.

Habert, B. (2000). Des corpus représentatifs : De quoi, pour quoi, comment. *Cahiers de l'Université de Perpignan*, *31*, 11–58.

Habert, B. (2005). *Instruments et ressources électroniques pour le français*. Paris: Editions Ophrys.

Habert, B., Fabre, C., & Issac, F. (1998). *De l'écrit au numérique : Constituer, normaliser et exploiter les corpus électroniques*. Paris: InterEditions.

Habert, B., Nazarenko, A., & Salem, A. (1997). *Les linguistiques de corpus*. Paris: Colin.

Hall, P. A., & Dowling, G. R. (1980). Approximate string matching. *ACM computing surveys (CSUR)*, *12*(4), 381–402.

Han, B., & Baldwin, T. (2011). Lexical Normalisation of Short Text Messages : Makn Sens a #Twitter. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, 368–378. Stroudsburg, PA, USA: Association for Computational Linguistics.

Heeringa, W., Kleiweg, P., Gooskens, C., & Nerbonne, J. (2006). Evaluation of string distance algorithms for dialectology. *Proceedings of the workshop on linguistic distances*, 51–62. Association for Computational Linguistics.

Heiden, S., Guillot, C., & Lavrentiev, A. (2010). *Manuel d'encodage XML-TEI des textes de la Base de Français Médiéval. 2002–2008*. Consulté à l'adresse http://bfm.ens-lyon.fr/IMG/pdf/Manuel_Encodage_TEI.pdf

Heift, T. (2010). Developing an Intelligent Language Tutor. *CALICO Journal*, *27*(3), 443-459.

Heift, T. (2017). History and Key Developments in Intelligent Computer-Assisted Language Learning (ICALL). In S. Thorne & S. May (Éd.), *Language, Education and Technology* (p. 1-12). Cham: Springer International Publishing.

Heift, T., & Schulze, M. (2007). *Errors and Intelligence in Computer-Assisted Language Learning : Parsers and Pedagogues*. Routledge.

Hirschman, L., & Mani, I. (2003). Evaluation. In *The Oxford Handbook of Computational Linguistics* (Oxford University Press). Oxford: Ruslan Mitkov.

Hulden, M., Alegria, I., Etxeberria, I., & Maritxalar, M. (2011). Learning word-level dialectal variation as phonological replacement rules using a limited parallel corpus. *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, 39–48. Association for Computational Linguistics.

Inkpen, D., Frunza, O., & Kondrak, G. (2005). Automatic identification of cognates and false friends in French and English. *Proceedings of the International Conference Recent Advances in Natural Language Processing*, *9*, 251–257. Varna, Bulgarie.

Jacques, M.-P. (2005). Pourquoi une linguistique de corpus ? In G. Williams (Éd.), *La linguistique de corpus* (p. 21-30). Rennes: Presses Universitaires de Rennes.

Jacques, M.-P. (2017). *La dynamique du texte, corpus, outils, analyses* (Thèse d'habilitation à diriger des recherches). Université Grenoble Alpes.

Jacques, M.-P., & Rinck, F. (2017). Un corpus de littéracie avancée : Résultat et point de départ. *Corpus*, *16*, 217-238.

Jansche, M. (2001). Re-engineering letter-to-sound rules. *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, 1–7. Association for Computational Linguistics.

Jansche, M. (2003). Parametric models of linguistic count data. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, 288–295. Association for Computational Linguistics.

Johannessen, J. B., Hagen, K., & Lane, P. (2002). *The performance of a grammar checker with deviant language input*. 1-8. Association for Computational Linguistics.

Jose, G., & Raj, N. S. (2014). Lexical normalization model for noisy SMS text. *2014 First International Conference on Computational Systems and Communications (ICCSC)*, 57-62.

Juel, C. (1988). Learning to Read and Write : A Longitudinal Study of 54 Children From First Through Fourth Grades. *Journal of Educational Psychology*, *80*(4), 437-447.

Jurafsky, D., & Martin, J. H. (2016). *Speech and language processing : An introduction to natural language processing, computational linguistics, and speech recognition*. Dorling Kindersley Pvt, Ltd.

Kempken, S., Luther, W., & Pilz, T. (2006). Comparison of distance measures for historical spelling variants. *IFIP International Conference on Artificial Intelligence in Theory and Practice*, 295–304. Boston, MA: Springer.

Kestemont, M., Daelemans, W., & De Pauw, G. (2010). Weigh your words—Memory-based lemmatization for Middle Dutch. *Literary and Linguistic Computing*, *25*(3), 287–301.

Kobus, C., Yvon, F., & Damnati, G. (2008a). Normalizing SMS : Are two metaphors better than one? *Proceedings of the 22nd International Conference on Computational Linguistics - COLING '08*, *1*, 441-448. Manchester, United Kingdom: Association for Computational Linguistics.

Kobus, C., Yvon, F., & Damnati, G. (2008b). Transcrire les SMS comme on reconnaît la parole. *Actes de la Conférence sur le Traitement Automatique des Langues (TALN'08)*, 128–138.

Kogkitsidou, E. (2018). *Communiquer par SMS : Analyse automatique du langage et extraction de l'information véhiculée* (Thèse de Doctorat). Université Grenoble Alpes.

Kondrak, G. (2000). A new algorithm for the alignment of phonetic sequences. *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, 288–295. Association for Computational Linguistics.

Kondrak, G. (2002). Determining recurrent sound correspondences by inducing translation models. *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, 1–7. Association for Computational Linguistics.

Kondrak, G., & Sherif, T. (2006). Evaluation of several phonetic similarity algorithms on the task of cognate identification. *Proceedings of the Workshop on Linguistic Distances*, 43–50. Association for Computational Linguistics.

Kraif, O. (2001). *Constitution et exploitation de bi-textes pour l'Aide à la traduction* (Thèse de Doctorat). Université de Nice Sophia Antipolis.

Kraif, O., & Ponton, C. (2007). Du bruit, du silence et des ambiguïtés : Que faire du TAL pour l'apprentissage des langues. *Actes de la Conférence sur le Traitement Automatique des Langues (TALN'07), Toulouse, 12-15 juin 2007*.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English.* (Vol. 143). Providence: Brown University Press.

Lallich-Boldin, G., & Maret, D. (2005). *Recherche d'information et traitement de la langue : Fondements linguistiques et applications*. Villeurbanne: Presses de l'Enssib.

Lamraoui, F., & Langlais, P. (2013). Yet another fast, robust and open source sentence aligner. Time to reconsider sentence alignment. *XIV Machine Translation Summit*.

Lavalley, R., Berkling, K., & Stüker, S. (2015). Preparing children's writing database for automated processing. *LTLT@ SLaTE*, 9–15.

Lecolle, M. (2001). Figures et référence plurielle, en corpus journalistique. *Cahiers de Grammaire*, 25, 29-52.

Léon, J. (2008). Aux sources de la « Corpus Linguistics » : Firth et la London School. *Langages*, *171*, 12-33.

Léon, J. (2015). *Histoire de l'automatisation des sciences du langage*. Paris: ENS Éditions.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, *10*, 707–710.

L'haire, S., & Vandeventer-Faltin, A. (2003). Diagnostic d´erreurs dans le projet FreeText. *Alsic. Apprentissage des Langues et Systèmes d'Information et de Communication*, *6*(2), 21-37.

Lyras, D. P., Sgarbas, K. N., & Fakotakis, N. D. (2007). Using the levenshtein edit distance for automatic lemmatization : A case study for modern greek and english. *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, *2*, 428–435. Patras, Greece: IEEE.

Ma, Y., Audibert, L., & Nazarenko, A. (2009). Ontologies étendues pour l'annotation sémantique. *20es Journées Francophones d'Ingénierie des Connaissances*, 205-216. Hammamet, Tunisie.

Mackay, W., & Kondrak, G. (2005). Computing word similarity and identifying cognates with Pair Hidden Markov Models. *Proceedings of the Ninth Conference on Computational Natural Language Learning*, 40–47. Association for Computational Linguistics.

Maire, H., Auriac-Slusarczyk, E., Slusarsczyk, B., Daniel, M.-F., & Thebault, C. (2018). Does One Stand to Gain by Combining Art with Philosophy? A Study of Fourth-Year College (13/14 Years of Age) Philosophical Writings Produced within the « Precphi/Philosophemes » Corpus. *Journal of Education and Learning*, *7*(4), 1-19.

Malrieu, D. (2007). Linguistique de corpus et caractérisation des ge nres : Un exemple d'analyse d'un conte didactique. In *Dyptique*: *Vol. 10*. *Construire et exploiter des corpus de genres scolaires* (p. 125-140).

Mann, G. S., & Yarowsky, D. (2001). Multipath Translation Lexicon Induction via Bridge Languages. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, 1–8. Stroudsburg, PA, USA: Association for Computational Linguistics.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). *Building a large annotated corpus of English : The Penn Treebank* (Technical Report N° MS-CIS-93-87; p. 313-330). University of Pennsylvania Department of Computer and Information Science.

Martinet, A. (1979). *Grammaire fonctionnelle du français*. Saint-Cloud; Paris: CRÉDIF ; Didier.

Marzal, A., & Vidal, E. (1993). Computation of normalized edit distance and applications. *IEEE transactions on pattern analysis and machine intelligence*, *15*(9), 926–932.

Masseron, C. (2004). Analyse critique de quelques dispositifs et activités d'écriture en DEUG. *Pratiques*, *121*(1), 58-80.

Mayaffre, D. (2005). Rôle et place du corpus en linguistique. Réflexions introductives. *Actes du colloque JETOU'2005*, 5–17. Toulouse: Université de Toulouse-Le Mirail.

McEnery, A. M. (2003). Corpus linguistics. In R. Mitkov (Éd.), *The Oxford handbook of computational linguistics* (p. 448-463). Oxford: Oxford University Press.

McEnery, Anthony M, & Wilson, A. (2001). *Corpus linguistics : An introduction.* Edinburgh: Edinburgh University Press.

Melamed, I. D. (1995). Automatic Evaluation and Uniform Filter Cascades for Inducing N-Best Translation Lexicons. *Proceedings of the 3rd Workshop on Very Large Corpora (WVLC3)*, *cmp-lg/9505044*. Boston.

Melamed, I. D. (1998a). *Annotation style guide for the blinker project*. Institute for Research in Cognitive Science Technical Report #98-06. University of Pennsylvania, Philadelphia, PA.

Melamed, I. D. (1998b). *Manual Annotation of Translational Equivalence : The Blinker Project*. Institute for Research in Cognitive Science Technical Report #98-07. University of Pennsylvania, Philadelphia, PA.

Meleuc, S., & Fauchart, N. (1999). *Didactique de la conjugaison : Le verbe" autrement"*. Paris, Toulouse: Bertrand-Lacoste; CRDP Midi-Pyrénées.

Meurers, D. (2015). Learner corpora and natural language processing. In S. Granger, G. Gilquin, & F. Meunier (Éd.), *The Cambridge Handbook of Learner Corpus Research* (p. 537-566). Cambridge: Cambridge University Press.

Meurers, W. D. (2013). On the Automatic Analysis of Learner Language : Introduction to the Special Issue. *CALICO Journal*, *26*(3), 469-473.

Mihalcea, R., & Pedersen, T. (2003). An evaluation exercise for word alignment. *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond*, 1–10.

Mitkov, Ruslan. (2003). *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press.

Mohri, M. (2003). Edit-distance of weighted automata : General definitions and algorithms. *International Journal of Foundations of Computer Science*, *14*(06), 957–982.

Muggleton, S., & De Raedt, L. (1994). Inductive logic programming : Theory and methods. *The Journal of Logic Programming*, *19*, 629–679.

Nagata, N. (2009). Robo-Sensei's NLP-Based Error Detection and Feedback Generation. *CALICO Journal*, *26*(3), 562-579.

Née, É. (Éd.). (2017). *Méthodes et outils informatiques pour l'analyse des discours*. Rennes: Presses universitaires de Rennes.

Nonnon, É. (2010). La notion de progression au cœur des tensions de l'activité d'enseignement. *Repères. Recherches en didactique du français langue maternelle*, (41), 5-34.

Och, F. J., & Ney, H. (2000). Improved Statistical Alignment Models. *Proceedings of the 38th Annual Meeting of the Association of Computational Linguistics (ACL)*, 440–447. Stroudsburg, PA, USA: Association for Computational Linguistics.

Ozdowska, S. (2007). Trois expériences d'évaluation dans le cadre du développement d'un système d'alignement sous-phrastique. *TAL*, *48*(1), 93–114.

Panckhurst, R. (2009). Short Message Service (SMS) : Typologie et problématiques futures. In *Polyphonies, pour Michelle Lanvin* (p. 33-52). Université Paul-Valéry Montpellier 3.

Panckhurst, R., Détrie, C., Verine, B., Lopez, C., Moïse, C., & Roche, M. (2014). Une grande collecte de SMS authentiques en français : Démarche, remarques et conseils. *Le français à l'université. Bulletin des départements de français dans le monde*, *19*(3).

Peereman, R., Lété, B., & Sprenger-Charolles, L. (2007). Manulex-infra : Distributional characteristics of grapheme—Phoneme mappings, and infralexical and lexical units in child-directed written material. *Behavior Research Methods*, *39*(3), 579–589.

Pellat, J.-C. (1988). Indépendance ou interaction de l'écrit et de l'oral ? Recensement critique des définitions du graphème. *Pour une Théorie de la Langue Ecrite (Table ronde internationale CNRS- HESO, Paris, 23-24 oct. 1986)*, 133-146. Paris: CNRS Editions.

Pellat, J.-C. (2017). *Quelle grammaire enseigner ?* Paris: Hatier.

Penloup, M.-C. (2000). *La tentation du littéraire. Essai sur le rapport a l'écriture littéraire du scripteur ordinaire*. Paris: Didier.

Penloup, M.-C. (2001). De quelques propriétés d'une pratique de lecture extrascolaire : Le courrier des lecteurs du journal Astrapi. *Repères. Recherches en didactique du français langue maternelle*, *23*(1), 75-91.

Perera, K. (1985). Grammatical Differentiation between Speech and Writing in Children Aged 8 to 12. In *The Writing of Writing* (p. 90-108). Buckingham: Open University Press.

Péry-Woodley, M.-P. (1995). Quel corpus pour quels traitements automatiques? *TAL. Traitement automatique des langues*, *36*(1-2), 213–232.

Pierrel, J.-M. (2000). *Ingénierie des langues*. Paris: Hermes.

Pierrel, J.-M. (2005). Linguistique de corpus et Traitement Automatique de la Langue. In J.-G. Ganascia (Éd.), *Communication et connaissances : Supports et médiations à l'âge de l'information*. Paris: CNRS Editions.

Pilz, T., Ernst-Gerlach, A., Kempken, S., Rayson, P., & Archer, D. (2008). The identification of spelling variants in English and German historical texts : Manual or automatic? *Literary and Linguistic Computing*, *23*(1), 65–72.

Pincemin, B. (2006). Concordances et concordanciers : De l'art du bon KWAC. In F. Rastier, M. Ballabriga, C. Duteil-Mougel, & B. Fouquié (Éd.), *XVIIe colloque d'Albi Lagages et signification—Corpus en Lettres et Sciences sociales : Des documents numériques à l'interprétation* (p. 33-42). Albi, France: CALS-CPST.

Poibeau, T. (2014). Le traitement automatique des langues pour les sciences sociales. *Réseaux*, *188*(6), 25-51.

Popin, J., & Thomasset, C. A. (1998). *La ponctuation*. Paris: Nathan.

Porta, J., Sancho, J.-L., & Gómez, J. (2013). Edit transducers for spelling variation in Old Spanish. *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013; May 22-24; 2013; Oslo; Norway. NEALT Proceedings Series 18*, 70–79. Linköping University Electronic Press.

Poudat, C., & Landragin, F. (2017). *Explorer des données textuelles : Méthodes - pratiques - outils* (1re édition). Paris: De Boeck supérieur.

Pravec, N. A. (2002). Survey of learner corpora. *ICAME Journal*, *26*, 81–114.

Quirk, R. (1968). The survey of English usage. In *Essays on the English language medieval and modern.* (p. 70–87). London: Longman.

Raccah, P.-Y. (2018). Le corpus de pommes d'Isaac Newton était-il un Grand Corpus ? *Corpus*, *18*.

Rastier, F. (2005). Enjeux épistémologiques de la linguistique de corpus. In *Rivages linguistiques*. *La linguistique de corpus* (Presses Universitares de Rennes, p. 31-45). Rennes.

Reuter, Y. (2006). Les récits sollicitant le vécu au CM2. Eléments d'analyse et de comparaison. *Repères. Recherches en didactique du français langue maternelle*, *34*(1), 111-139.

Riou, J. (2017). *Étude de l'influence de l'enseignement du code alphabétique sur la qualité des apprentissages des élèves de cours préparatoire* (Thèse de doctorat). Université Clermont Auvergne.

Ristad, E. S., & Yianilos, P. N. (1998). Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(5), 522–532.

Roubaud, M.-N. (2017). Le français écrit : Transcription et édition. Le cas des textes scolaires. *Corpus*, *16*, 113-132.

Sagot, B., & Boullier, P. (2008). SxPipe 2 : Architecture pour le traitement pré-syntaxique de corpus bruts. *Traitement Automatique des Langues*, *49*(2), 155-188.

Santiago Oriola, C. (1998). *Systeme vocal interactif pour l'apprentissage des langues. La synthese de la parole au service de la dictee* (Thèse de Doctorat). Toulouse 3.

Savelli, M., Brissaud, C., Chevrot, J.-P., & Gounon, V. (2002). L'apprentissage d'un temps peu enseigné : Le passé simple. *Le francais aujourd'hui*, *4*, 39–48.

Scherrer, Y. (2007). Adaptive string distance measures for bilingual dialect lexicon induction. *Proceedings of the ACL 2007 Student Research Workshop*, 55–60.

Scherrer, Y. (2008). *Transducteurs à fenêtre glissante pour l'induction lexicale*.

Simard, M., Foster, G. F., & Isabelle, P. (1993). Using cognates to align sentences in bilingual corpora. *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing-Volume 2*, 1071–1082. Toronto, Ontario: IBM Press.

Similowski, K., Pellan, D., & Plane, S. (2018). Que révèlent les traces de réécriture dans les brouillons d'élèves produisant des récits à partir de sources littéraires ? *Repères. Recherches en didactique du français langue maternelle*, *57*, 15-34.

Sinclair, J. M. (1996). *EAGLES Preliminary recommendations on Corpus Typology EAG--TCWG--CTYP/P Version of May*.

Singh, A. K., Subramaniam, S., & Rama, T. (2010). Transliteration as Alignment vs. Transliteration as Generation for Crosslingual Information Retrieval. *TAL*, *51*(2), 95–117.

Šmilauer, I. (2011). Description formelle et diagnostic automatique des erreurs en langue étrangère : Quelques perspectives pour les outils d'ELAO. In T. Ponchon & I. Labord-Milla (Éd.), *Sciences du langage et nouvelles technologies (ASL'09)* (p. 107-115). Lambert-Lucas.

Smith, N., McEnery, T., & Ivanic, R. (1998). Issues in Transcribing a Corpus of Children's Hanwritten Projects. *Literary and Linguistic Computing*, *13*(4), 217-225.

Soulé, Y., Kervyn, B., Geoffre, T., & Chabanne, J.-C. (2016). Évaluer la production d'écrit en fin du cours préparatoire (première primaire). De l'élaboration d'une épreuve de test à l'analyse des résultats obtenus. *L'évaluation en classe de français, outil didactique et politique*, 85–107.

Steuckardt, A. (2014). De l'écrit vers la parole. Enquête sur les correspondances peu lettrées de la Grande Guerre. *Actes du 4e Congrès Mondial de Linguistique Française (CMLF 2014), Juillet 2014*, *8*, 353-364.

Stosic, D., Marjanović, S., & Miletic, A. (2018). ParCoGLiJe Corpus parallèle pour l'étude des grands classiques de la littérature de jeunesse Objectif du projet. *Journée d'étude CORLI : Traitements et standardisation des corpus multimodaux et web 2.0.* Présenté à Paris. Paris.

Stüker, S., Fay, J., & Berkling, K. (2011). *Towards Context-Dependent Phonetic Spelling Error Correction in Children's Freely Composed Text for Diagnostic and Pedagogical Purposes*. *Twelfth Annual Conference of the International Speech Communication Association*, 4.

Tarrade, L. (2017). *Normalisation des messages issus de la communication électronique médiée* (Mémoire de Master 2). Université Grenoble Alpes.

Testenoire, P.-Y. (2017). Transcrire des écrits scolaires : Entre philologie et génétique textuelle. *Corpus*, *16*, 87-109.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Tono, Y. (2003). Learner corpora : Design, development and applications. *Proceedings of the Corpus Linguistics 2003 conference*, 800-809. Lancaster: University Centre for Computer Corpus Research on Language.

van Rooy, B., & Schäfer, L. (2003).s *An evaluation of three POS taggers for the tagging of the Tswana Learner English Corpus*. *Proceedings of the Corpus Linguistics 2003 conference*, 835-844.

Vénérin-Guénez, C. (2018). À la recherche de stratégies scripturales dans des rappels de récit par des élèves de cours moyen. *Repères. Recherches en didactique du français langue maternelle*, *57*, 83-98.

Verlinde, S. (2010). La conception de didacticiels intégrés d'aide à la lecture, à la traduction et à la rédaction. *Revue francaise de linguistique appliquee*, *15*(2), 53-65.

Véronis, J. (1988). Computerized correction of phonographic errors. *Computers and the Humanities*, *22*(1), 43–56.

Véronis, J. (2000a). Alignement de corpus multilingues. *Ingénierie des langues*, 151–171.

Véronis, J. (2000b). Annotation automatique de corpus : Panorama et état de la technique. In *Ingénierie des langues* (Vol. 4, p. 111-129).

Véronis, J., & Langlais, P. (2000). Evaluation of parallel text alignment systems. In *Parallel text processing* (p. 369–388). Springer.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, *13*(2), 260–269.

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

Williams, G. (Éd.). (2005). *La linguistique de corpus*. Rennes: Presses Univ. de Rennes.

Williams, G. (2006). La linguistique et le corpus : Une affaire prépositionnelle. *Texto*, 151-158.

Wolfarth, C. (2015). *Apport du TAL à la constitution et l'exploitation d'un corpus scolaire de cours préparatoire* (Mémoire de Master 2). Université Grenoble Alpes.

Wolfarth, C., Brissaud, C., & Ponton, C. (2018). Transcrire et normer un corpus scolaire : Pour quelles analyses ? In C. Brissaud, M. Dreyfus, & B. Kervyn (Éd.), *Repenser l'écriture et son évaluation au primaire et au secondaire* (p. 121–146). Presses universitaires de Namur.

Wolfarth, C., Ponton, C., & Brissaud, C. (2018). Gestion de la morphographie verbale en production d'écrits : Que peut nous apprendre un corpus longitudinal ? *Repères. Recherches en didactique du français langue maternelle*, *57*, 209–226.

Zargayouna, H., Roussey, C., & Ouardani, S. (2017). Semantic annotation from text : Use case of agricultural observation from alerte bulletin. *9ème atelier Recherche d'Information SEmantique (RISE 2017) adossé à la conférence IC 2017 de la Plateforme Francophone d'Intelligence Artificielle*, 8 p. Caen, France.

Zobel, J., & Dart, P. (1996). Phonetic string matching : Lessons from information retrieval. *Proceedings of the 19th International Conference on Research and Development in Information Retrieval*, *96*, 166–172. Zurich, Suisse: ACM Press.

Liste des tableaux

Tableau 1 : Correspondances classe – âge moyen des élèves – année de scolarité xviii
Tableau 2 : Liste non exhaustive de corpus scolaires non francophones existants..... 46
Tableau 3 : Liste non exhaustive de corpus scolaires existants pour le français langue première ou langue de scolarisation 50
Tableau 4 : Exemple d'éléments manipulables au moyen de l'approche par comparaison et des résultats attendus..... 83
Tableau 5 : Écoles participant au recueil Scoledit 90
Tableau 6 : Supports de l'épreuve de dictée..... 96
Tableau 7 : Caractéristiques du corpus longitudinal 101
Tableau 8 : Correspondance entre phonèmes et archiphonèmes 159
Tableau 9 : Les différentes représentations d'une forme, les exemples de faire et « fér » ... 159
Tableau 10 : Étiquetage morphosyntaxique à l'aide de l'outil TreeTagger 160
Tableau 11 : Représentation phonologique produite par l'outil LIA-PHON 161
Tableau 12 : Représentation phonologique produite par l'outil LIA-PHON 162
Tableau 13 : Représentation archiphonologique produite lors de la phase de prétraitement (exemple 1) 162
Tableau 14 : Représentation archiphonologique produite lors de la phase de prétraitement (exemple 2) 162
Tableau 15 : Exemple de tokenisations des outils LIA-PHON et TreeTagger 163
Tableau 16 : Données accessibles à la fin de la phase de prétraitements 164
Tableau 17 : Résultat matriciel du calcul d'édition classique..... 175
Tableau 18 : Résultat matriciel du calcul d'édition pondéré..... 175
Tableau 19 : Résultat matriciel du calcul de la distance d'édition de Damereau-Levenshtein 176
Tableau 20 : Caractéristiques du corpus de référence..... 183
Tableau 21 : Évaluation des différences entre la 1ère correction et la 2ème correction 185
Tableau 22 : Différents niveaux de l' algorithme séquentiel évalués 186
Tableau 23 : Résultats de l'évaluation de chaque mode de comparaison 186
Tableau 24 : Alignement d'un extrait de la production 93-CE1 (résultat attendu) 187
Tableau 25 : Alignement d'un extrait de la production 93-CE1 (sortie de l'aligneur niveau 4) 187
Tableau 26 : Alignement d'un extrait de la production 93-CE1 (sortie de l'aligneur niveau 5) 187
Tableau 27 : Alignement d'un extrait de la production 93-CP (sortie de l'aligneur niveau 6) 188

Tableau 28 : Résultats de l'évaluation du niveau 7 188
Tableau 29 : Évaluation des variations de longueur de la fenêtre de segmentation 189
Tableau 30 : Évaluation des variations de longueur de la fenêtre de tokens 189
Tableau 31 : Différents niveaux de l' algorithme séquentiel évalués 190
Tableau 32 : Résultats de l'évaluation de chaque mode de comparaison 190
Tableau 33 : Résultats de l'évaluation du mode de comparaison (niveau 5-A et 5-D)..... 191
Tableau 34 : Évaluation des variations de longueur d'hyposegmentation 191
Tableau 35 : Évaluation des variations de longueur d'hypersegmentation 192
Tableau 36 : Résultats de l'évaluation par niveau scolaire 192
Tableau 37 : Exemple d'alignement (extrait de la production 772-CE2)..... 194
Tableau 38 : Exemple d'alignement (extrait de la production 93-CE1)..... 194
Tableau 39 : Caractéristiques du corpus de travail de 480 productions 198
Tableau 40 : Occurrences des phénomènes d'hyposegmentation et d'hypersegmentation . 199
Tableau 41 : Exemple d'alignement par segments..... 199
Tableau 42 : Exemple d'alignement par séquences de segments 200
Tableau 43 : Formes élidées fréquemment impliquées dans des erreurs d'hyposegmentation (480 productions) 200
Tableau 44 : Correspondance entre phonèmes et archiphonèmes 202
Tableau 45 : Exemple d'alignement produit par AliScol..... 203
Tableau 46 : Répartition des réussites et des types orthographiques dans le corpus longitudinal 204
Tableau 47 : Répartition des catégories syntaxiques (corpus longitudinal)..... 205
Tableau 48 : Extrait du fichier de travail de l'équipe 1 207
Tableau 49 : Extrait du fichier de travail pour les étiquettes d'erreur de l'équipe 2 207
Tableau 50 : Extrait du fichier de travail pour les étiquettes morphosyntaxiques de l'équipe 2 208
Tableau 51 : Degré de différence entre les corrections des deux équipes d'évaluation 208
Tableau 52 : Évaluation de la fiabilité des étiquettes apposées par l'aligneur 208
Tableau 53 : Synthèse des définitions du graphème (d'après Pellat, 1988) 214
Tableau 54 : Exemple de sortie de la fonction de découpage en graphème de l'outil LIA-PHON 221
Tableau 55 : Quelques exemples d'ajustement du découpage en graphème 221
Tableau 56 : Données disponibles pour chaque forme ou segment après découpage en graphèmes et calcul de la représentation archiphonologique 222
Tableau 57 : Couts utilisés pour le calcul d'édition dans l'algorithme Aliscol_Graph..... 222
Tableau 58 : Caractéristiques du corpus d'évaluation 223

Tableau 59 : Mesures de l'accord inter-annotateurs 223
Tableau 60 : Évaluation des performances de l'aligneur AliScol_Graph 224
Tableau 61 : Données générales concernant le corpus recueilli 225
Tableau 62 : Désinences du présent de l'indicatif, de l'imparfait et du passé simple..... 226
Tableau 63 : Les différentes marques du mode participe passé 227
Tableau 64 : Répartition des échecs et des réussites pour chaque tiroir verbal étudié 230
Tableau 65 : Tableau d'avancement de la numérisation et l'enrichissement du corpus 239
Tableau 66 : Caractéristiques du corpus Scoledit..... 269
Tableau 67 : Caractéristiques du corpus longitudinal 269
Tableau 68 : Caractéristiques du corpus de développement 270
Tableau 69 : Caractéristiques du corpus de référence..... 270
Tableau 70 : Caractéristiques du corpus de travail 270
Tableau 71 : Caractéristiques du corpus d'évaluation 271
Tableau 72 : Caractéristiques du corpus de travail 271

Liste des figures

Figure 1 : Processus de génération (à gauche) et d'analyse (à droite) 27
Figure 2 : Production de texte en fin de CM2 de l'élève 2973 62
Figure 3 : Production de texte en fin de CP de l'élève 1637 64
Figure 4 : Exemple de version iconique (Boré & Elalouf, 2017, p. 36) 65
Figure 5 : Exemple de version commentée (Boré & Elalouf, 2017, p. 37) 65
Figure 6 : Exemple de transcription du projet Ecriscol (Doquet et al., 2017, p. 144) 66
Figure 7 : Exemple de transcription (Garcia-Debanc et al., 2017, p. 164) 67
Figure 8 : Exemple de transcription du corpus ÉMA 70
Figure 9 : Exemples d'annotations embarquées, corpus ÉMA 73
Figure 10 : Exemple issu du corpus BFM (Heiden et al., 2010) 76
Figure 11 : Exemples issus du corpus SMS4Science (Fairon et al., 2006) 76
Figure 12 : Approche par comparaison 82
Figure 13 : Répartition des écoles du projet Scoledit 89
Figure 14 : Images présentées aux élèves lors de la production écrite en CP 92
Figure 15 : Production de texte en CP - élève 1154 93
Figure 16 : Images présentées aux élèves lors de la production écrite en CE1, CE2, CM1 et CM2 94
Figure 17 : Production de texte en CE1 - élève 1154 95
Figure 18 : Ensemble des données recueillies lors des deux phases de recueil 100
Figure 19 : Répartition des écoles au sein des académies 102
Figure 20 : Composition sociale des écoles 102
Figure 21 : Répartition genrée des élèves 102
Figure 22 : Répartition des élèves par langue parlée à la maison 102
Figure 23 : Scan et transcription de la production de texte de CP de l'élève 1363 104
Figure 24 : Scan et transcription de la production de texte de CP de l'élève 2413 105
Figure 25 : Scan et transcription de la production de texte de CE2 de l'élève 1101 105
Figure 26 : Scan et transcription de la production de texte de CM2 de l'élève 103 106
Figure 27 : Scan et transcription de la production de texte de CE1 de l'élève 1283 107
Figure 28 : Scan et transcription de la production de texte de CE1 de l'élève 1283 107
Figure 29 : Extrait du scan et de la transcription de la production de texte de CE2 de l'élève 3003 108
Figure 30 : Scan et transcription de la production de texte de CP de l'élève 1573 108
Figure 31 : Scan et transcription de la production de texte de CP de l'élève 218 109

Figure 32 : Extrait du scan et de la transcription de la production de texte de CE2 de l'élève 1339 110
Figure 33 : Extrait du scan et de la transcription de la production de texte de CE2 de l'élève 239 110
Figure 34 : Scan et transcription de la production de texte de CP de l'élève 207 111
Figure 35 : Interface de transcription 112
Figure 36 : Interface de visualisation du corpus (ensemble du corpus + production particulière) 113
Figure 37 : Interface de visualisation de l'état d'avancement 114
Figure 38 : Interface de visualisation du corpus 115
Figure 39 : Exemple d'outil d'exploration du corpus 116
Figure 40 : Commentaires laissés dans l'espace de commentaires à propos de la transcription 117
Figure 41 : Utilisateurs du site Scoledit 118
Figure 42 : Exemple de trois versions d'une production (200, CE1) 130
Figure 43 : Approche par comparaison appliquée à l'identification des formes du corpus 138
Figure 44 : Alignement simple 140
Figure 45 : Alignement maximisé 140
Figure 46 : Matrice d'édition entre les formes pran et prend 143
Figure 47 : Calcul de distance entre deux séquences de caractères 145
Figure 48 : Schéma général de l'algorithme d'alignement 156
Figure 49 : Résultat de la comparaison graphique stricte 165
Figure 50 : Résultat de la comparaison phonologique stricte 165
Figure 51 : Résultat de la comparaison phonologique avec archiphonologique stricte 165
Figure 52 : Ordre des comparaisons 167
Figure 53 : Comparaison sans fenêtre glissante (= fenêtre glissante de taille 1) 169
Figure 54 : Comparaison avec fenêtre glissante 169
Figure 55 : Axes de comparaison de l'algorithme d'alignement 169
Figure 56 : Schéma séquentiel de l'algorithme d'alignement A 170
Figure 57 : Exemple de parcours en cas de segmentation normée 171
Figure 58 : Exemple de parcours en cas d'hyposegmentation 172
Figure 59 : Exemple de parcours en cas d'omission 172
Figure 60 : Exemple de parcours en cas d'omission et d'hyposegmentation combinées 173
Figure 61 : Exemple d'alignement par défaut 173
Figure 62 : Exemple de parcours matriciel 175
Figure 63: Exemple de recherche d'inversion 176

Figure 64 : Exemple de recherche d'hypersegmentation 177
Figure 65 : Résultat matriciel du calcul de la distance d'édition avec recherche d'hypersegmentation 177
Figure 66 : Extrait de la production 94-CE1 193
Figure 67 : Extrait de la production 1160-CE2 193
Figure 68 : Répartition des réussites et des types d'erreurs orthographiques année par année 204
Figure 69 : Répartition des réussites orthographiques par catégorie 206
Figure 70 : Processus d'application des critères de discrimination des graphèmes 216
Figure 71 : Schéma général de l'algorithme d'alignement en graphèmes AliScol_Graph..... 220
Figure 72 : Répartition des temps et modes verbaux selon le niveau de scolarité 228
Figure 73 : Répartition des erreurs selon l'année d'apprentissage 229
Figure 74 : Répartition des erreurs sur les bases et les désinences des formes verbales..... 231
Figure 75 : Répartition des erreurs sur les bases et les désinences des formes verbales à l'imparfait..... 232
Figure 76 : Répartition des erreurs sur les bases et les désinences des formes verbales au passé simple..... 233

Annexes

Table des annexes

Annexe 1 : Les différents corpus utilisés 269
Annexe 2 : Conventions nationales de constitution de corpus scolaires 273
Annexe 3 : Livrets de recueil 281
Annexe 4 : Ensemble des productions recueillies pour un élève (élève 96) 289
Annexe 5 : Guide de transcription 295
Annexe 6 : Guide de normalisation 301
Annexe 7 : French *TreeTagger* Part-of-Speech Tags 313
Annexe 8 : Table des correspondances des formats de transcription phonologique 315
Annexe 9 : Liste des correspondances graphophonémiques établie par J. Riou et R. Goigoux (2017) 317
Annexe 10 : Protocole d'alignement en graphèmes 319

Annexe 1 : Les différents corpus utilisés

Le **corpus *Scoledit*** est constitué de l'ensemble des productions recueillies dans le cadre du projet *Scoledit*. Il est composé de 4 931 textes et de 4 459 dictées de niveaux différents (CP – CM2).

Niveau scolaire	Nombre de textes	Nombre de dictées
CP	763	2 126
CE1	871	736
CE2	1 135	/
CM1	1 132	813
CM2	1 030	784
Total	*4 931*	*4 459*

Tableau 67 : Caractéristiques du corpus Scoledit

Le **corpus longitudinal** est le sous-corpus qui ne contient que les productions des élèves présents toutes les années du CP au CM1. Il est composé de 1 492 textes de niveaux différents (CP – CM2) produits par 373 élèves.

Niveau scolaire	*Nombre de productions*	*Nombre de tokens*
CP	373	11 141
CE1	373	27 504
CE2	373	48 654
CM1	373	58 952
Total	*1 492*	*146 251*

Tableau 68 : Caractéristiques du corpus longitudinal

Étude 1 : Alignement des formes (Chapitre 8, 9 et 10)

Dans le cadre de l'étude sur l'alignement des formes, différents corpus ont été utilisés :

1. Corpus de développement

Le corpus de développement (Chapitre 8) est composé de 53 textes de niveaux différents (CP – CE2), issus du corpus longitudinal.

Niveau scolaire	*Nombre de productions*	*Nombre de tokens*
CP	28	758
CE1	14	1 089
CE2	11	1 415
Total	*53*	*3 262*

Tableau 69 : Caractéristiques du corpus de développement

2. Corpus d'évaluation

Le corpus d'évaluation (Chapitre 9), aussi appelé **corpus de référence**, est composé de 200 textes produits par 50 élèves du CP au CM1 (un texte par niveau), issus du corpus longitudinal.

Niveau scolaire	*Nombre de productions*	*Nombre de tokens*
CP	50	1 573
CE1	50	4 547
CE2	50	7 451
CM1	50	8 889
Total	*200*	*22 460*

Tableau 70 : Caractéristiques du corpus de référence

3. Corpus de travail

Le corpus de travail (Chapitre 10) est composé de 480 textes produits par 120 élèves du CP au CM1 (un texte par niveau), issus du corpus longitudinal. Ce corpus inclut le corpus de référence.

Niveau scolaire	*Nombre de productions*	*Nombre de tokens*
CP	120	3 597
CE1	120	9 754
CE2	120	15 791
CM1	120	19 099
Total	*480*	*48 241*

Tableau 71 : Caractéristiques du corpus de travail

Étude 2 : Alignement des graphèmes (Chapitre 11)

Dans le cadre de l'étude sur l'alignement des graphèmes, un corpus d'évaluation a été constitué. Il est composé de 36 textes, produits par 9 élèves du CP au CM1 (un texte par niveau), issus du corpus longitudinal.

Niveau scolaire	*Nombre de productions*	*Nombre de tokens*	*Nombre de graphèmes*
CP	9	249	1 014
CE1	9	598	2 477
CE2	9	1281	5 409
CM1	9	1349	5 500
Total	*36*	*3477*	*14 040*

Tableau 72 : Caractéristiques du corpus d'évaluation

Étude 3 : Etude de la morphographie verbale (Chapitre 12)

Dans le cadre de l'étude sur la morphographie verbale, un corpus de travail a été défini. Il est composé de 903 textes de niveaux différents, produits par 301 élèves du CP au CE2, issus du corpus longitudinal.

Niveau scolaire	*Nombre de productions*	*Nombre de tokens*
CP	301	7 877
CE1	301	20 125
CE2	301	35 195
Total	*903*	*63 197*

Tableau 73 : Caractéristiques du corpus de travail

Annexe 2 : Conventions nationales de constitution de corpus scolaires

Règles pour la constitution d'un corpus d'écrits d'élèves

Pour chaque manuscrit, on fait 2 fichiers :

- Un fichier qui est une « simple » transcription et qui est destiné à être lu par des utilisateurs de la plate-forme de textes.
- Un fichier qui comporte les annotations et qui est sert de base à la création d'un autre fichier, de format différent, qui sera traité par un logiciel de textométrie.

1. PROCEDURE DE TRANSCRIPTION ET DE NORMALISATION DES COPIES D'ELEVES :

1. Recueil des données, des méta-données et des autorisations de diffusion.
2. Numérisation des données (+ anonymisation sur le scan mais garder la trace qu'il y a eu un masquage. Les infos pourront être reliées via les métadonnées) : V1
3. Informatisation des données primaires : transcription : V2
   - association/saisie des méta-données
   - saisir via un éditeur de texte brut[1] une forme au plus près de la copie scannée
   - élément nécessitant un codage particulier (cf. tableau : « éléments à transcrire »)
   - vérification de la transcription (script de validation ?)

Ces éléments de codage seront « traduits » en balises XML au moment du passage à la Version 3

4. Normalisation des données primaires → données secondaires : V 3
   - transformer les données primaires au format texte brut vers un format XML (norme TEIP5) et/ou un autre format facilitant l'annotation (Glozz ?)
   - annoter les éléments permettant de convertir les données primaires en données linguistiquement « normées » → besoin d'un typologie
     - présence d'une chaîne de caractère incorrecte : erreur d'orthographe lexicale, erreur d'orthographe grammaticale, erreur de segmentation, erreur typographique (majuscules), erreur de ponctuation (point en milieu de « phrase »), ...
     - manque d'éléments lexical ou typographique
     - autre ?
   - générer plusieurs versions ergonomiques pour explorer/interroger les données primaires et secondaires

1 Par exemple https://notepad-plus-plus.org/fr/ pour Windows et http://www.barebones.com/products/textwrangler/ pour Mac

EA 7345 *Clesthia*

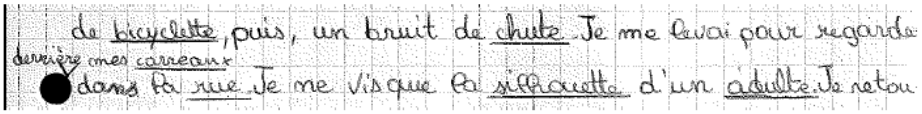
EMA
Ecole, Mutations, Apprentissages
EA 4507

FA 4671

2. TRANSCRIPTION

Principe n°1 : s'inscrire dans la tradition de la philologie et de la génétique textuelle, en adoptant les mêmes codes pour les opérations de base : suppression / ajout
- encadrement par des crochets pour les [élément supprimés]
- encadrement par des chevrons pour les <éléments ajoutés>

Principe n°2 : établir des transcriptions qui puissent être exploitées ensuite pour l'annotation des fichiers, en privilégiant la linéarisation du texte. Cela revient à indiquer des opérations avec des signaux spécifiques plutôt que placer les segments textuels exactement à la même place sur la transcription et sur le manuscrit. Par exemple, en cas d'ajout hors ligne sur le manuscrit, on rétablit la place que vient prendre l'élément ajouté dans le texte (dans la ligne) :

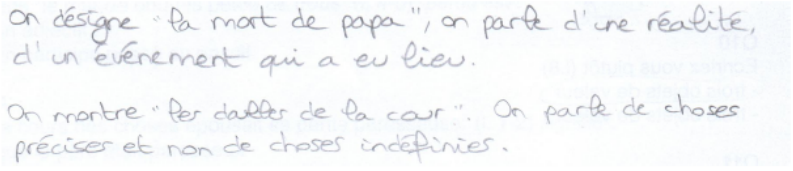


de bicyclette, puis, un bruit de chute. Je me levai pour regarder
<derrière mes carreaux> dans la rue. Je ne vis que la silhouette d'un adulte.

Segment illisible :
#XXX# mot illisible [#XXX#]
#X# partie de mot illisible (ex : fille#X#)

Anonymisation :
Dans le texte, on remplace les noms propres par des initiales.
Toponyme / Homme / Femme
Albert = A_h
Lucie = L_f
Paris = P_t

Délimitation des lignes :
Aller à la ligne à chaque fin de ligne du manuscrit et marquer la fin de ligne par /n
Marquer les retours à la ligne volontaire sur le manuscrit par §, combiné au /n : §\n



On désigne « la mort de papa », on parle d'une réalité, /n
d'un événement qui a eu lieu. §/n
On montre « les dalles de la cour ». On parle de choses /n

EA 7345 *Clesthia* — Ecole, Mutations, Apprentissages EA 4507 — EA 4671

précises et non de choses indéfinies. §/n

Tableau récapitulatif des règles de transcription :

Élément à transcrire	**Proposition de codage**
Ajout	<caractères ajoutés>
Suppression	[caractères supprimés]
Déplacement de texte	- écrire le texte à l'endroit où il apparaît sur la copie et l'encadrer des marques de suppression [] - identifier le segment. Ex : @1 [texte source] - localiser la cible du déplacement en indiquant l'ajout <@1>
Segment illisible Segments illisibles séparés par des blancs Paragraphe illisible (1 élément par ligne) Caractère illisible dans un mot	#xxx# #xxx# #xxx# #xxx# #x# inséré entre les caractères lisibles - ex. : for#x#t
Retours à la ligne, alinéas Retour à la ligne imposé (espace de la feuille) retour à la ligne volontaire saut de ligne alinéa en début de paragraphe	 \n retour à la ligne + § : \n§ ligne vide + § : \n§\n Tabulation : \t
Informations à modifier par le transcripteur (e.g. données privées, anonymisation) Noms propres modifiés : Noms de personne Toponymes	#code# où *code* est indiqué dans l'en-tête de la transcription. initiale du nom propre « _ » type de nom e.g. Madame A_n = Madame Albert, femme → Madame #A_f# (pour la mention de Madame Albert) e.g. S_n = Salem, homme → #S_h# (pour la mention de Salem) e.g. G_t = Garonne, toponyme → #G_t# (pour la mention de la Garonne)
Multi-écritures Plusieurs scripteurs : le professeur est identifié par P., les élèves différents sont identifiés par E (E1, E2…). Plusieurs moments d'écriture (retour a posteriori sur l'écrit) : identifiés par T1, T2 etc.	 si rature : [code#caractères raturés] si ajout : <code#caractères ajoutés> … où code est indiqué dans l'en-tête de la transcription e.g. P = enseignant → [P#caractères raturés] @1 [T2#texte source] dépl a posteriori par le même élève @1 [E2T2#texte source] dépl a posteriori par un autre élève @1 [PT2#texte source] dépl a posteriori par l'enseignant
Autre élément (grandes lettres, dessins, ...)	Tout élément non indiqué explicitement dans le guide n'est pas à transcrire mais à mentionner sous forme de commentaire libre dans l'en-tête. Cela permet d'avertir le lecteur de l'existence d'éléments non transcrits. {marge} élément verbal ne s'inscrivant pas tel quel dans le texte {T2#marge} élément verbal ne s'inscrivant pas tel quel dans le texte inscrit a posteriori #FIGURE# élément non verbal faisant partie de l'écrit {#FIGURE#} élément non verbal périphérique ne faisant pas partie de l'écrit

3. ANNOTATION :

Le fichier annoté n'est pas destiné à être vu par des utilisateurs du corpus mais à être lu par un logiciel de textométrie. Les conventions d'annotation correspondent en partie aux conventions de codage XML, ce qui signifie que certains signes, les chevrons par exemple, n'ont plus la valeur philologique (marquage d'un ajout) mais la valeur textométrique (balise XML).

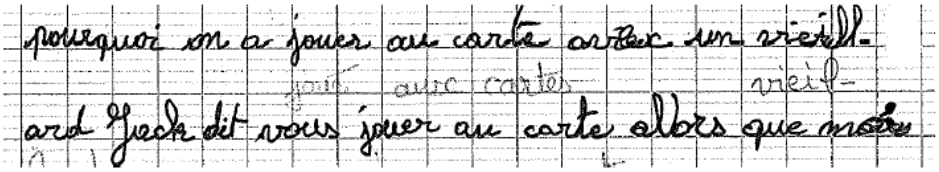
L'annotation linguistique concerne les erreurs orthographiques. On ne corrige pas la ponctuation.
Pour annoter un mot ou une suite de mots dont on souhaite rectifier l'orthographe, on procède de la manière suivante :

- isoler le segment à annoter entre deux signes < >
- placer ensuite le tiret bas _
- écrire le segment normé, entre les deux signes < >

Exemples :

- Erreur d'orthographe sur un seul mot non composé :
  - Les <pettit>_<petites> filles
- Erreur sur un mot composé :
  - <rez de chaussé>_<rez-de-chaussée>
- Erreur de segmentation :
  - <lape tifil>_<la petite fille>

Coupure de mot en fin de ligne sur le manuscrit :
Si un scripteur coupe un mot parce qu'il est en fin de ligne, on procède de la même manière que pour les corrections orthographiques avec le système de < > mais en signalant que la correction est due à un problème de décodage du segment par le logiciel et non à une erreur d'orthographe (cela, pour éviter que les coupes en fin de ligne ne soient comptées comme des erreurs d'orthographe, qu'elles ne sont pas).
Par exemple :



pourquoi on a jouer au carte arter un vieill-
ard Jack dit vous jouer au carte alors que moi

pourquoi on a <jouer>_<joué> <au>_<aux> <carte>_<cartes> avec un <vieill-£
ard>_<vieillard> Jack dit vous <jouer>_<joué> <au>_<aux> <carte>_<cartes> alors que moi

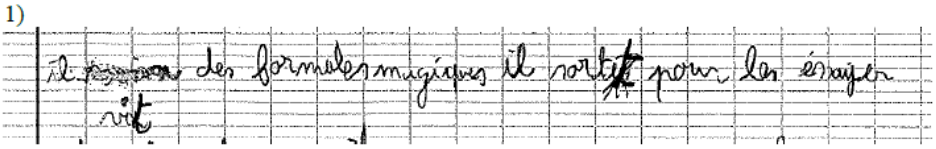
Cas des remplacements :
D'un point de vue strictement procédural, remplacement = suppression + ajout.

EA 7345 *Clesthia*

EA 4671

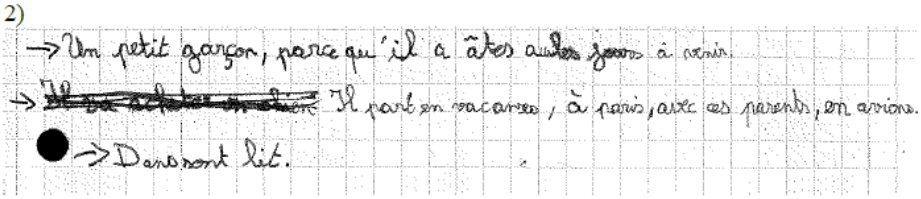
Mais d'un point de vue linguistique, remplacement = substitution de X par Y, les deux occupant une même place syntaxique.
Donc un adjectif peut être remplacé par une proposition relative, un GN par un autre GN, etc.
Exemples de successions suppression + ajouts qui sont, ou pas, des remplacements :

1)



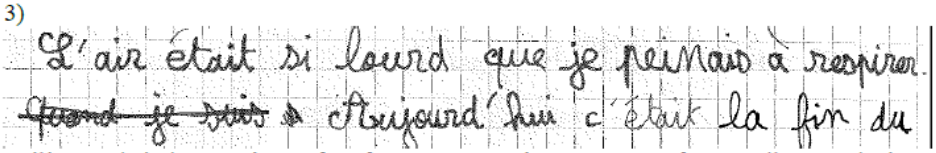
- L'élève avait écrit il *essaia*, il a supprimé *essaia* et a ajouté *vit*. On a bien une équivalence syntaxique (verbe / verbe), donc on peut dire que le scripteur a remplacé *essaia* par *vit*.

2)



Sur la deuxième ligne l'élève avait écrit *Il va acheter un chien*, il l'a supprimé puis a écrit *Il part en vacances, à paris, avec ces parents, en avions*. C'est bien un remplacement (substitution d'une phrase X à une phrase Y).

3)



L'élève avait écrit *L'air était si lourd que je peinais à respirer quand je suis*, il a supprimé *quand je suis* et mis un point après *respirer*. Ce n'est pas un remplacement : c'est une suppression suivie de l'ajout d'un point (pas d'équivalence syntaxique).
Le Trameur n'est pas capable de décider tout seul si la succession suppression+ajout est ou n'est pas un remplacement. Il faut donc le lui signaler. On place un ® au début et à la fin des segments faisant l'objet d'un remplacement. Donc, sur les exemples précédents :

- exemple 1 :

il ®[essaia] //vit//® des formules magiques il sortit pour les éssayer

- exemple 2 :

®[Il va acheter un chien] //Il part en vacances, à paris, avec ses parent, en avions.//®

- exemple 3 :

L'air était si lourd que je peinais à respirer//.//
[Quand je suis] Aujourd'hui c'était la fin du

Cas des remplacements de lettres :

~~m~~Municipal

On transcrit et annote :

©[m]//M//©unicipal

La substitution de lettre est repérée avec © au lieu de ®
Attention : dans ce cas il ne faut pas mettre d'espace entre //M// et unicipal, de manière à ce que le logiciel traite Municipal comme un mot.

Quand un scripteur biffe un terme comportant une erreur d'orthographe :

J'ai vu mon ~~per~~

on transcrit et annote :

J'ai vu mon [<per>_<père>]

Pour un nom ajouté qui est mal orthographié :

J'ai vu mon //vie// père.

sera annoté :

J'ai vu mon //<vie>_<vieux>// père.

EA 7345 *Clesthia*

Ecole, Mutations, Apprentissages
EA 4507

EA 4671

Tableau récapitulatif des règles d'annotation :

Elément à annoter	Proposition de codage
Erreur (ortho)graphique - erreur sur un mot : Les pettit filles - erreur sur un mot composé : rez de chaussée - erreur de segmentation : lape tifil - lettre illisible : dans la for♦t lointaine	<forme erronée>_<forme normée> Les <pettit>_<petites> filles <rez de chaussée>_<rez-de-chaussée> <lape tifil>_<la petite fille> Dans la < for#x#t>_<forêt> lointaine
Ajout Ajout d'une forme erronée : je <lai> vis	//segment ajouté// je //<lai>_<les>// vis
Suppression d'une forme erronée : J'ai vu mon ~~per~~	[segment supprimé] J'ai vu mon [<per>_<père>]
Coupure de mot - en fin de ligne - mot inopinément coupé avec tiret	 on a joué aux <car-ftes>_<cartes> on a joué aux <car-tes>_<cartes>
Remplacement - de mot : il ~~essaia~~ vit - de suite de mots : ~~Il se promène~~ Il part en vacances - de lettres : ~~m~~Municipal	Repérage par ® : il ®[essaia] //vit//® des formules magiques ®[Il se promène] //Il part en vacances//® Repérage par © : ©[m]//M//©unicipal

Toutes les opérations recensées dans le tableau de transcription sont combinables avec l'annotation.

Pour une annotation plus facile, automatiser la transformation des signes d'ajout, < > dans la transcription, en // // (convention choisie pour remplacer, dans le fichier annoté, les chevrons qui deviennent des balises XML).

EA 7345 *Clesthia*

Ecole, Mutations, Apprentissages
EA 4507

EA 4671

4. DENOMINATION DES DOCUMENTS

Le nom doit indiquer :

- le type d'établissement : école (EC), collègue (CO), lycée (LY), université (UN)
- le niveau de classe : CM2 / 6 / 2 / L1
- l'identifiant de la classe (une lettre aléatoire)
- le numéro du devoir en cas de plusieurs écrits provenant de la même classe
- le numéro de l'élève dans la classe

Donc, pour la série « description de chambre qui a déjà reçu une première numérotation :
Exemple des numéros commençant par B : B1, B2 etc.
On les renomme :
EC-CM2-B-1-1, EC-CM2-B-1-2, EC-CM2-B-1-3 etc.
Et puisqu'il y a un second devoir venant de la même classe :
EC-CM2-B-2-1, EC-CM2-B-2-2, EC-CM2-B-2-3 etc.

S'il y a plusieurs versions du même texte (esquisse, brouillon, mise au net, etc.) on ajoute : V1 pour la 1ère version, V2 pour la deuxième. (s'il y a deux brouillons, on met V1 et V1bis pour faire en sorte que V2 soit toujours la version définitive).

Problèmes sur la numérotation :

- il faut indiquer l'année de production sur tous les écrits (pas seulement les L1)
- on a une succession de nombres à la fin, par forcément très pertinent. Eventuellement à changer en précisant dans le codage à quoi correspond le nombre.

Exemple : version 1 du 1er devoir de l'élève n°3, école magonty 2, page 2 :
EC-CM2-2015-MAG2-1-3-V1-2
Ou alors, en étant peut-être plus clair :
EC-CM2-2015-MAG2-D1-E3-V1-P2

Annexe 3 : Livrets de recueil

1. Livret de consigne en classe de CP

Livret produit pour la recherche « Lire – Écrire au CP » et destiné aux enquêteurs en classe de CP.

B – Test collectif

3. Production d'écrit

Test de production d'une histoire sur la base de 4 images montrées par l'évaluateur.

En temps limité – 15 minutes (après les consignes).

a) Conditions de passation et matériel

Matériel évaluateur :

- *Chronomètre.*
- *En annexe A, les 4 vignettes reproduites sur 4 feuilles au format A3 (feuilles séparées)*
- *Aimants ou scotch pour afficher les vignettes au tableau.*
- *Un exemplaire du livret élève pour montrer la page où ils écriront.*

Matériel élève :

- *Un crayon papier et une gomme.*

Dans le cahier, on trouve la page avec les lignes puis, au verso, les 4 vignettes reproduites.

Les élèves sont assis à leur place. L'enseignant est dans la classe ; il aide au bon déroulement de l'épreuve. Le prévenir de ne donner aucune aide aux élèves. On retire les affichages muraux susceptibles d'aider les élèves (le mot *chat*, des listes de verbes, etc.).

Les cahiers seront distribués une fois que l'évaluateur aura montré les 4 images.

b) Consigne

Quand les élèves sont tous attentifs, dire lentement : « **Aujourd'hui vous allez écrire chacun l'histoire d'un petit chat. Je vais vous montrer ce qui arrive à ce petit chat. Regardez bien comment ça commence.** »

Pour l'ordre des images, s'appuyer sur la série représentée dans le livret élève.

Montrer image 1 (format A3).

« Est-ce que vous le voyez, ce petit chat ? Oui, il est ici »

Le montrer du doigt. Puis bien montrer l'image à tous les élèves et l'afficher sur la partie la plus à gauche du tableau.

Si les élèves s'expriment, leur faire signe qu'ils ne doivent rien dire : « **Chut, regardez bien… On va regarder sans rien dire. Vous gardez toutes vos idées dans votre tête…** »

Montrer image 2 (format A3).

« Regardez bien la 2ème image… »

Si les élèves s'expriment, leur faire signe qu'ils ne doivent rien dire : « **Chut, regardez bien…**»

Bien montrer l'image à tous les élèves et l'afficher à droite de la première.

9

Montrer image 3 (format A3).

« Regardez bien la 3ème image… »

Si les élèves s'expriment, leur faire signe qu'ils ne doivent rien dire : **« Chut, regardez bien…»**

Bien montrer l'image à tous les élèves et l'afficher à droite de la deuxième image.

Montrer image 4 (format A3).

« Regardez bien la 4ème image… »

Si les élèves s'expriment, leur faire signe qu'ils ne doivent rien dire : **« Chut, regardez bien…»**

Bien montrer l'image à tous les élèves et l'afficher à droite des trois premières.

Les vignettes resteront au tableau pendant que les élèves écrivent.

« Vous avez bien regardé, vous avez bien dans votre tête l'histoire de ce petit chat ? »

Il est important de leur laisser le temps de « mettre l'histoire dans leur tête ».

Montrer ensuite la feuille du cahier sur laquelle ils vont écrire (lignes Seyes).

« Vous allez écrire cette histoire ici. Si vous avez oublié l'histoire, vous pouvez retourner la feuille pour retrouver les dessins. »

« Vous avez 15 minutes pour ce travail. Vous allez travailler seul ; personne ne vous aidera, par exemple à écrire un mot.

Distribuer les cahiers, bien faire repérer la page d'écriture et déclencher le chronomètre ⏲ quand tous les élèves sont prêts. Avertir les élèves lorsqu'il restera 5 minutes. Ramasser les cahiers au bout de 15 minutes.

Durant le déroulement de cette épreuve, faire les réponses les plus neutres possible, ne pas donner d'aide aux élèves.

Quand les élèves ont fini, leur dire de se reposer, par exemple en posant la tête sur les avant-bras.

- Ne pas autoriser le coloriage des vignettes qui pourrait inciter certains élèves à interrompre leur tâche d'écriture.
- Ne pas non plus autoriser les élèves à se lever et aller lire un livre.

c) Cotation

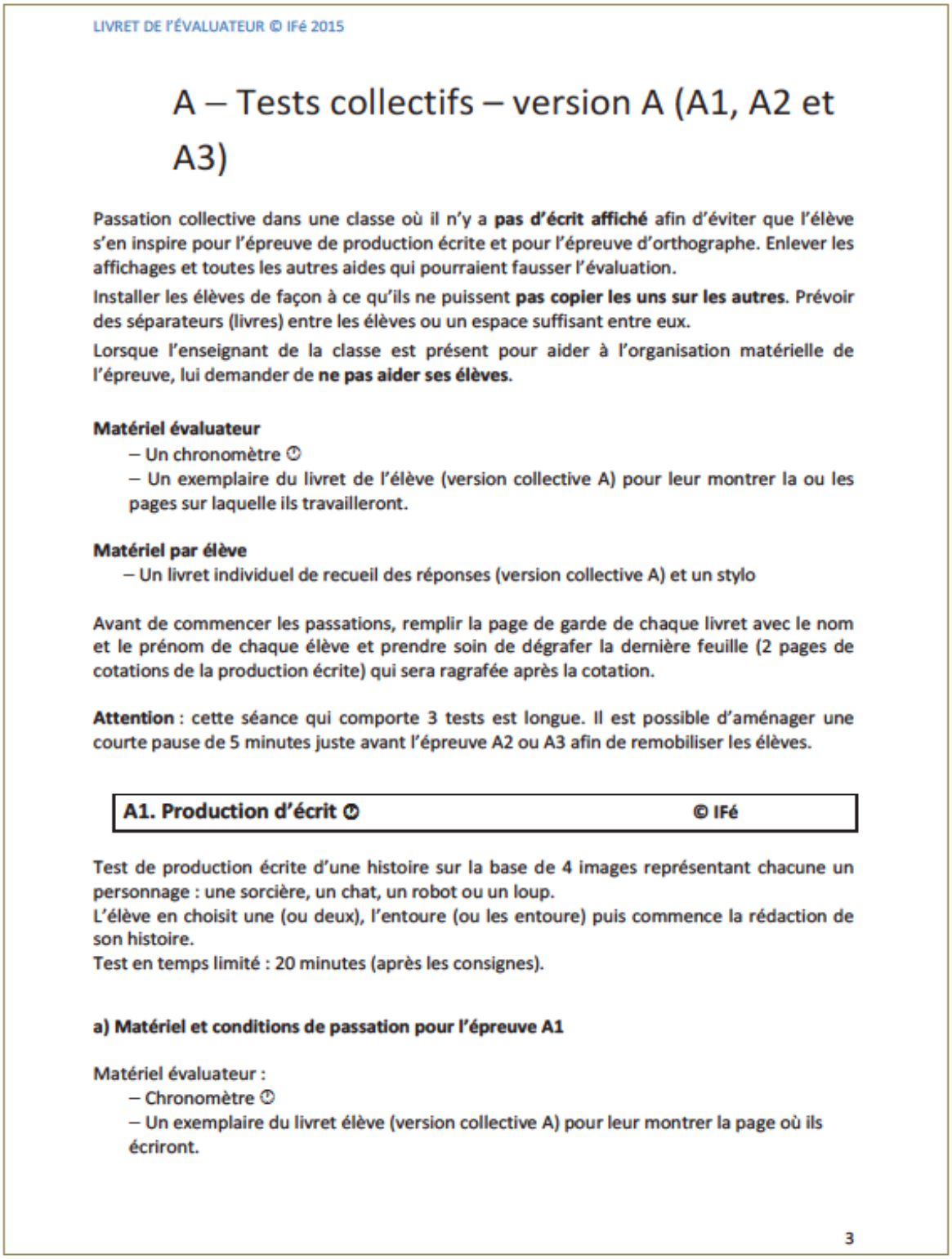
A coter sur le livret élève collectif (page finale).
Six analyseurs : longueur du texte produit – segmentation – lisibilité – séparateurs d'idées – quantité d'informations – traces de narration.

2. Support d'écriture, classe de CP et de CE1

Reproduction du support papier sur lequel les élèves étaient invités à écrire en classe de CP et de CE1 (recherche « Lire – Écrire au CP »).

3. Livret de consigne en classe de CE1

Livret produit pour la recherche « Lire – Écrire au CP » et destiné aux enquêteurs en classe de CE1.



A – Tests collectifs – version A (A1, A2 et A3)

Passation collective dans une classe où il n'y a **pas d'écrit affiché** afin d'éviter que l'élève s'en inspire pour l'épreuve de production écrite et pour l'épreuve d'orthographe. Enlever les affichages et toutes les autres aides qui pourraient fausser l'évaluation.

Installer les élèves de façon à ce qu'ils ne puissent **pas copier les uns sur les autres**. Prévoir des séparateurs (livres) entre les élèves ou un espace suffisant entre eux.

Lorsque l'enseignant de la classe est présent pour aider à l'organisation matérielle de l'épreuve, lui demander de **ne pas aider ses élèves**.

Matériel évaluateur

– Un chronomètre ⏲
– Un exemplaire du livret de l'élève (version collective A) pour leur montrer la ou les pages sur laquelle ils travailleront.

Matériel par élève

– Un livret individuel de recueil des réponses (version collective A) et un stylo

Avant de commencer les passations, remplir la page de garde de chaque livret avec le nom et le prénom de chaque élève et prendre soin de dégrafer la dernière feuille (2 pages de cotations de la production écrite) qui sera ragrafée après la cotation.

Attention : cette séance qui comporte 3 tests est longue. Il est possible d'aménager une courte pause de 5 minutes juste avant l'épreuve A2 ou A3 afin de remobiliser les élèves.

A1. Production d'écrit ⏲ **© IFé**

Test de production écrite d'une histoire sur la base de 4 images représentant chacune un personnage : une sorcière, un chat, un robot ou un loup.
L'élève en choisit une (ou deux), l'entoure (ou les entoure) puis commence la rédaction de son histoire.
Test en temps limité : 20 minutes (après les consignes).

a) Matériel et conditions de passation pour l'épreuve A1

Matériel évaluateur :

– Chronomètre ⏲
– Un exemplaire du livret élève (version collective A) pour leur montrer la page où ils écriront.

Matériel élève :
- Un stylo (noir ou bleu)
- Le livret élève version collective A

Rappel : si l'enseignant est dans la classe et aide au bon déroulement de l'épreuve, lui demander de ne donner aucune aide aux élèves.

Les affichages muraux susceptibles d'aider les élèves (les mots *robot, sorcière, loup* ou *chat* ; des listes de verbes, etc.) ont été retirés.

b) Consigne

Demander aux élèves d'ouvrir leur livret à la page 3, titrée *production écrite.*

Quand tous les élèves sont attentifs, dire lentement : *aujourd'hui, vous allez écrire chacun une histoire avec un ou deux personnages.*

Demander aux élèves de mettre le doigt sur l'image 1 et leur faire nommer le personnage : *montrez-moi la première image. Vous avez tous le doigt sur la première image ? Bien. Vous connaissez ce personnage ? Oui, il s'agit d'une* ***sorcière****.*

Demander ensuite aux élèves de mettre le doigt sur l'image 2 et leur faire nommer le personnage : *mettez le doigt sur la deuxième image. Ça y est ? Bien. Quel est ce personnage ? Oui, il s'agit d'un* ***chat****.*

Procéder de la même manière pour **le robot** (image n° 3) **et le loup** (image n° 4).

Dire : *je vais vous lire la consigne qui est écrite sous les images, écoutez bien la consigne. Voici 4 personnages. Choisis un ou deux personnages et raconte une histoire. Entoure le ou les personnages que tu as choisis.*

Dire aux élèves : *Vous avez bien regardé les images ? Vous avez choisi un ou deux personnages ? Quand vous avez choisi, entourez votre ou vos personnages (un ou deux). Avant de commencer à écrire, mettez bien dans votre tête l'histoire de ce personnage ou de ces deux personnages.*

Vous avez 20 minutes pour écrire cette histoire. Vous allez travailler seul ; personne ne vous aidera, par exemple à écrire un mot.

Déclencher le chronomètre quand tous les élèves sont prêts.

Avertir les élèves lorsqu'il restera 5 minutes. Durant le déroulement de cette épreuve, faire les réponses le plus neutres possible, ne pas donner d'aide aux élèves.

Quand les élèves ont fini, leur dire de se reposer.

Ne pas autoriser le coloriage des vignettes qui pourrait inciter certains élèves à interrompre leur tâche d'écriture.

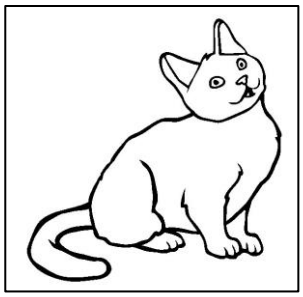
Ne pas non plus autoriser les élèves à se lever et aller lire un livre.

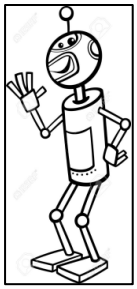
Attention ! Ne pas laisser les élèves commencer à lire le texte « Les enfants et la sorcière » de la page 7 lorsqu'ils auront terminé la production d'écrit. Le début de l'épreuve A2 doit être synchronisé pour tous.

4. Support d'écriture, classe de CE2, CM1 et CM2

Reproduction du support papier sur lequel les élèves étaient invités à écrire en classe de CE2, CM1 et CM2 (recherche *Scoledit*).

Prénom : Nom :





1 2 3 4

Voici 4 personnages. Choisis un ou deux personnages et raconte une histoire. Entoure le ou les personnages que tu as choisis.

Code de l'élève : Code de la classe :

5. Consignes CE2-CM2

Consignes pour le recueil de CM2

(sur la base de consignes données en CE1 et CE2)

Sur chaque feuille, demander aux élèves d'écrire le prénom et le nom et d'inscrire « NuméroDeL'Ecole – Niveau » après *Code de la classe :*, ne rien mettre après *Code de l'élève*.

1. Dictée

a) Consigne pour la dictée de phrases

Dire aux élèves : *sur les lignes, nous allons faire une dictée de phrases. Voici la première phrase. Je vais vous la lire une première fois (vous n'écrivez rien), puis je vous la dicterai. On commence*.

Dicter aux élèves : ***En été, les salades vertes poussent dans les jardins.***

Procéder de la même façon pour : ***Les jeunes canetons picorent le blé avec la poule noire.***

Relire les deux phrases à haute voix. Laisser une minute ou deux aux élèves

b) Consigne pour la dictée de phrases

Dire au élèves : *Je vous lis maintenant un texte de deux phrases (vous n'écrivez rien), puis je vous les dicterai l'une après l'autre*.

Ecrire le titre ***Le corbeau*** sur la première ligne, plus petite.

Le corbeau / perché sur l'antenne d'un bâtiment / tient dans son bec / une souris blessée. / Rendus furieux / par cet oiseau cruel, / des enfants lancent des cailloux / pour l'obliger à s'envoler. /

Dire aux élèves. *Vous avez deux minutes pour relire et pour corriger les erreurs que vous avez pu faire.*
Préciser : *Vérifiez si vous avez bien mis les majuscules et les points. Vérifiez si vous avez bien fait attention à tous les accords.*

2. Production d'écrit

a) Conditions de passation

Test de production écrite d'une histoire sur la base de 4 images représentant chacune un personnage : une sorcière, un chat, un robot ou un loup.
L'élève en choisit une (ou deux), l'entoure (ou les entoure) puis commence la rédaction de son histoire.
Test en temps limité : 30 minutes (après les consignes), 25 min d'écriture et 5 minutes de relecture (pour tous). Leur préciser qu'au bout de 25 minutes vous leur direz qu'il reste 5

minutes et qu'ils doivent maintenant relire. Si l'enseignant est dans la classe (ce qui est souhaitable) et aide au bon déroulement de l'épreuve, lui demander de ne donner aucune aide aux élèves.

b) Consignes

Demander aux élèves d'écrire leur nom et prénom.
Quand tous les élèves sont attentifs, dire lentement :
aujourd'hui, vous allez écrire chacun une histoire avec un ou deux personnages.

Demander à un élève de dire à voix haute quels sont les personnages représentés pour être sûrs que tous les élèves aient les mêmes personnages en tête.

Dire : *je vais vous lire la consigne qui est écrite sous les images, écoutez bien la consigne*.
Voici 4 personnages. Choisis un ou deux personnages et raconte une histoire. Entoure le ou les personnages que tu as choisis.

Dire aux élèves : *Vous avez bien regardé les images ? Vous avez choisi un ou deux personnages ? Quand vous avez choisi, entourez votre ou vos personnages (un ou deux). Avant de commencer à écrire, mettez bien dans votre tête l'histoire de ce personnage ou de ces deux personnages.*

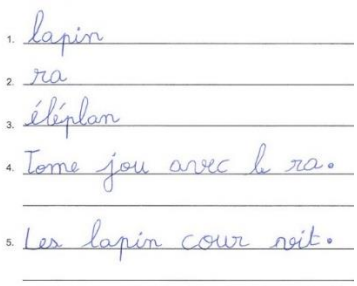
Vous avez 25 minutes pour écrire cette histoire. Vous allez travailler seul ; personne ne vous aidera, par exemple à écrire un mot. Au bout de 20 minutes, je vous demanderai de relire votre histoire et de vérifier l'orthographe.

Déclencher le chronomètre quand tous les élèves sont prêts.
Avertir les élèves lorsqu'il restera 5 minutes. Durant le déroulement de cette épreuve, faire les réponses le plus neutres possible, ne pas donner d'aide aux élèves.

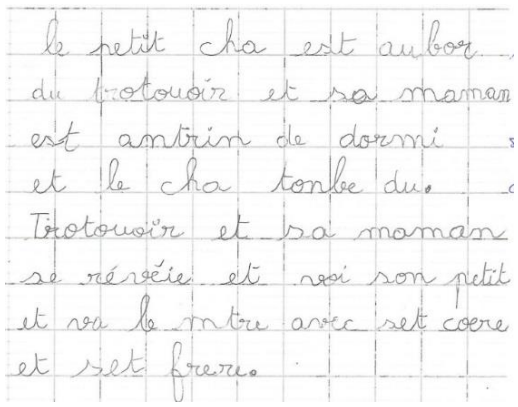
Quand les élèves ont fini, leur dire de se reposer.
Ne pas autoriser le coloriage des vignettes qui pourrait inciter certains élèves à interrompre leur tâche d'écriture. Ne pas non plus autoriser les élèves à se lever et aller lire un livre.

Annexe 4 : Ensemble des productions recueillies pour un élève (élève 96)

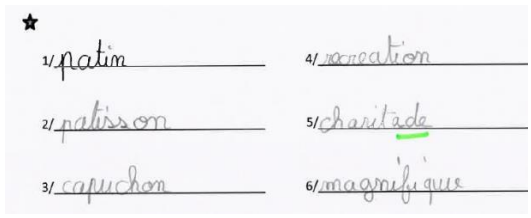
Dictée - début CP

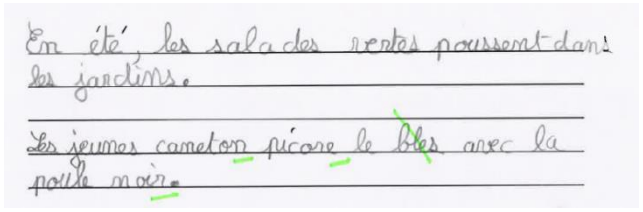


Dictée - fin CP

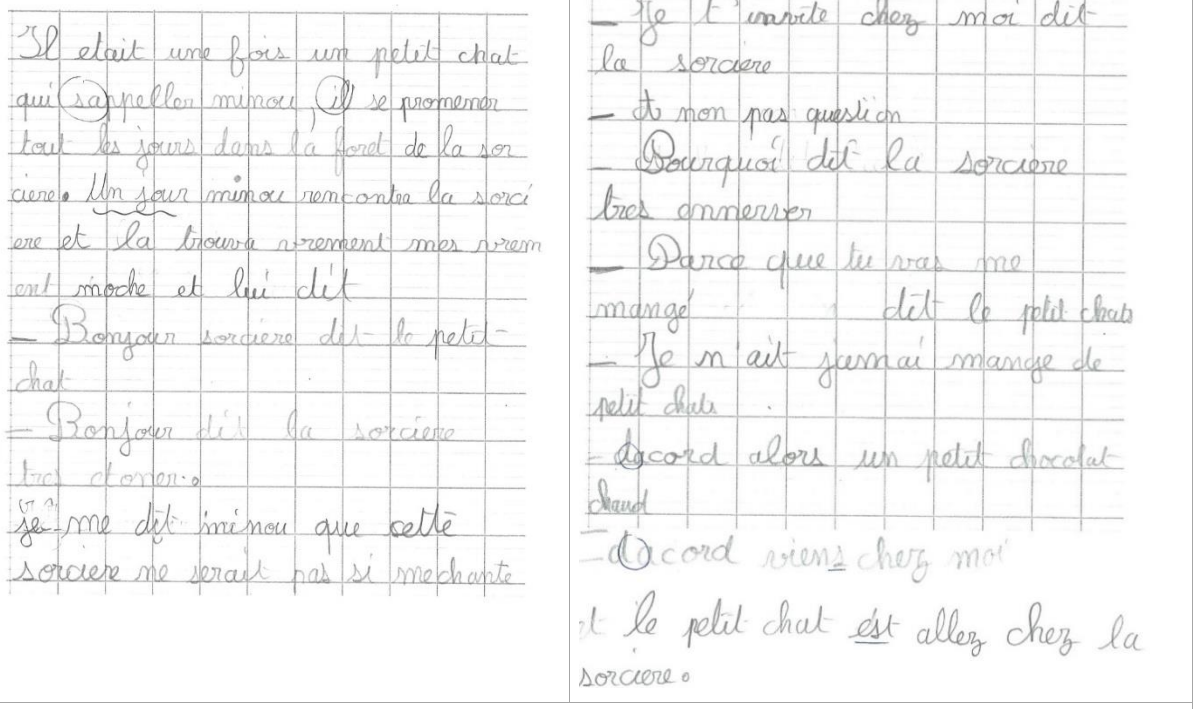


Texte – CP





Dictée - CE1



Il etait une fois un petit chat qui s'appelle minou, il se promenait tout les jours dans la foret de la sorciere. Un jour minou rencontre la sorciere et la trouve vraiment mes vraiment moche et lui dit

— Bonjour sorciere dit le petit chat

— Bonjour dit la sorciere tres etonner.

Se me dit minou que cette sorciere ne serait pas si mechante que sa

— Je t'invite chez moi dit la sorciere

— Je non pas question

— Pourquoi dit la sorciere tres etonner

— Parce que tu vas me manger dit le petit chat

— Je n'ait jamais manger de petit chat.

— D'accord alors un petit chocolat chaud

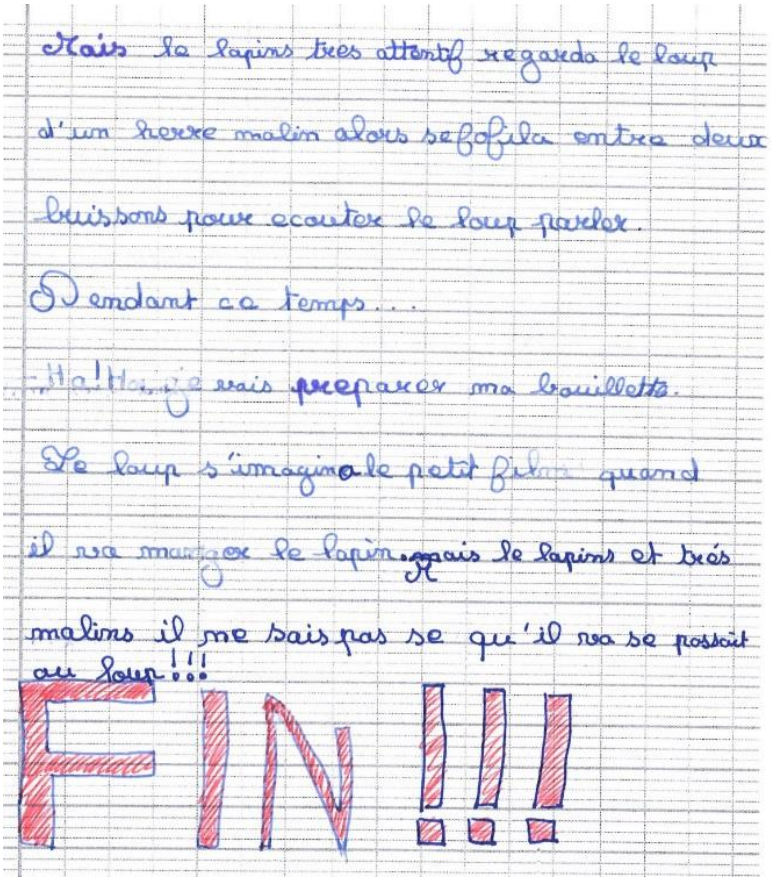
— D'accord viens chez moi

et le petit chat est allez chez la sorciere.

Texte - CE1

Il était une fois un chat tres malin et attentif à se qu'il faisait, mais il y a le loup aussi, lui tout le contraire du chat distrait et pas du tout malin. Un jour le chat se promener mais le loup était toujours à surveiller les chat, les lapins, et il tomba sure le chat.

- Mmm c'est un bon repas que je vais là je vais l'attraper et le manger pour le repas de se soir dira le loup.



Mais le lapins tres attentif regarda le loup
d'un herre malin alors se fofila entre deux
buissons pour ecouter le loup parler.
Pendant ce temps...
Ha! Ha! je vais preparer ma bouillette.
Le loup s'imagina le petit film quand
il va manger le lapin. Mais le lapins et trés
malins il ne sais pas se qu'il va se passait
au loup!!!

FIN!!!

Texte – CE2

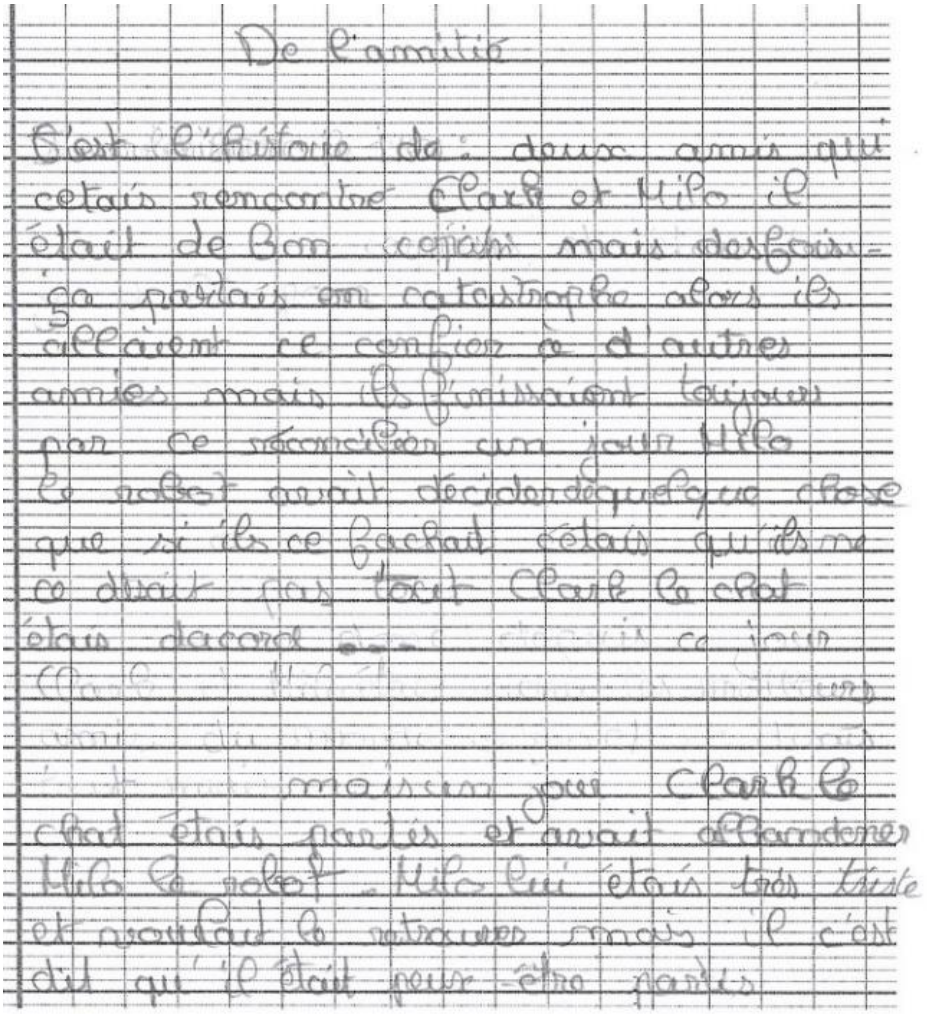
1) patin 4) récréation
2) patisson 5) charitable
3) capuchon 6) magnifique

En été, les salades vertes poussent dans les jardins.
Les jeunes cametons picorent le blé avec la poule noir.

Le corbeau

Le corbeau perché sur l'antaine d'un batiment tiens dans son bec une souris blessé. Rendu furieu par cet oiseau cruel, des enfants lancent des cailloux pour l'obliger à s'envoler.

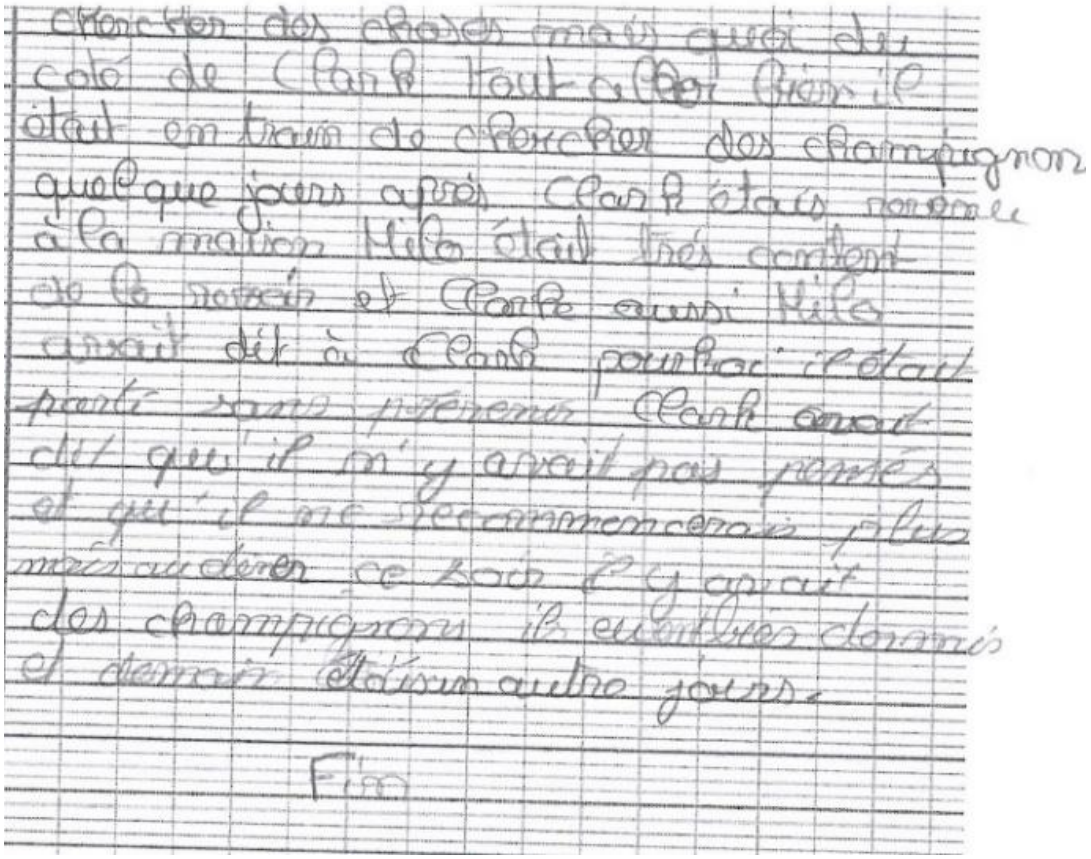
Dictée - CM1



De l'amitié

C'est l'histoire de deux amis qui cetais rencontre Clark et Milo il était de Bon copain mais desfois ça partais en catastrophe alors ils allaient ce confier à d'autres amies mais ils finissaient toujours par ce réconcilier un jour Milo le robot avait décider de quelque chose que si ils ce fachait cétais qu'ils ne ce disait pas tout Clark le chat étais dacord ... ce jour [illegible]

mais un jour Clark le chat étais partis et avait abandoner Milo le robot. Milo lui étais très triste et voulait le retrouver mais il c'est dit qu'il était peut-être partis



chercher des choses mais quoi du
coté de Clark tout aller bien il
était en train de chercher des champignon
quelque jours après Clark étais revenu
à la maison Milo était très content
de le revoir et Clark aussi Milo
avait dit à Clark pourkoi il était
parti sans prévenir Clark avait
dit qu'il n'y avait pas pensés
et qu'il ne recommencerais plus
mais au dîner ce soir il y avait
des champignons ils eurent bien dormis
et demain étaisun autre jours.

Fin

Texte – CM1

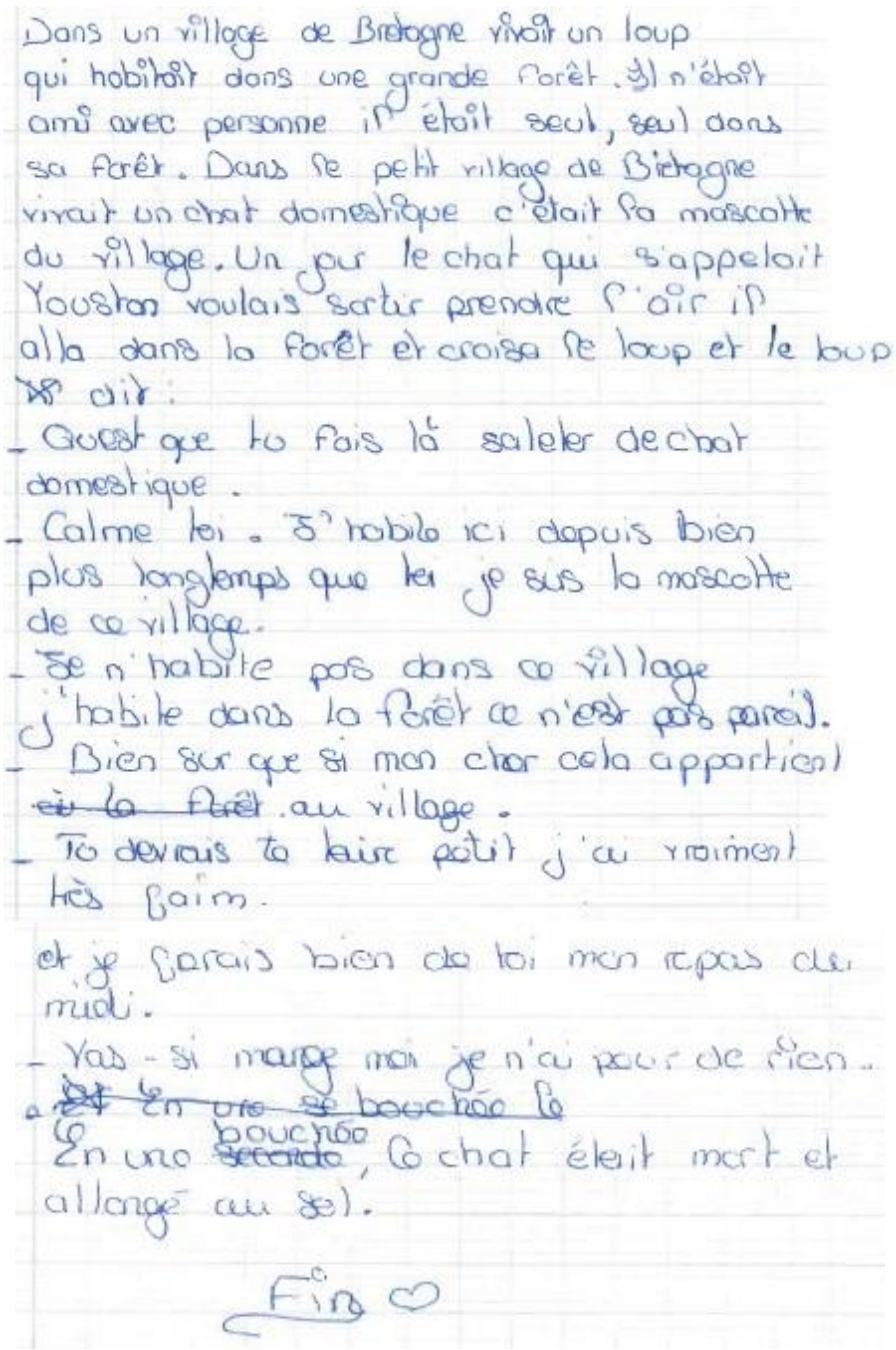
En été, les salades vertes pousses dans les jardins.

Les jeunes canetons picorts le blé avec la poule noir

Le corbau

Le corbau perché sur l'antenne d'un batiment
tiens dans son bec une souris blessé. Rendu
furieux par cette oiseau cruèle, des enfants
lance des cailloux pour l'obliger à sen volé.

Dictée – CM2



Dans un village de Bretagne vivait un loup qui habitait dans une grande forêt. Il n'était ami avec personne il était seul, seul dans sa forêt. Dans le petit village de Bretagne vivait un chat domestique c'était la mascotte du village. Un jour le chat qui s'appelait Youston voulais sortir prendre l'air il alla dans la forêt et croisa le loup et le loup dit :

- Qu'est que tu fais là saleter de chat domestique.
- Calme toi. J'habite ici depuis bien plus longtemps que toi je suis la mascotte de ce village.
- Je n'habite pas dans ce village j'habite dans la forêt ce n'est pas pareil.
- Bien sur que si mon chor cela appartient ~~en la forêt~~ au village.
- Tu devrais te taire petit j'ai vraiment très faim.

et je ferais bien de toi mon repas de midi.

- Vas-y si manges moi je n'ai peur de rien.

~~Et en une se bouchée le~~
En une ~~seconde~~ bouchée, le chat était mort et allongé au sol.

Fin ♡

Texte - CM2

Annexe 5 : Guide de transcription

Rédaction : Claire Wolfarth
Proposition issue du travail collectif : Catherine Brissaud, Claude Ponton, Corinne Totereau, Claire Wolfarth

La transcription est une des étapes du processus de numérisation. L'objectif de cette étape est de reproduire numériquement le texte de l'auteur ou de l'autrice. L'ensemble des traitements et analyses qui seront effectués sur le corpus seront effectués sur ces transcriptions. Cependant, il n'est pas possible de reproduire l'ensemble des informations contenues dans la version manuscrite des productions, des choix doivent être effectués au cours de cette étape en fonction des analyses et traitements prévus. Les choix effectués dans le cadre du projet *Scoledit* sont exposés ci-après.

Transcription des éléments génétiques 295
Transcription des éléments extratextuels 296
Transcription des éléments textuels 297
Transcription de l'incertitude du transcripteur ou de la transcriptrice 299

Transcription des éléments génétiques

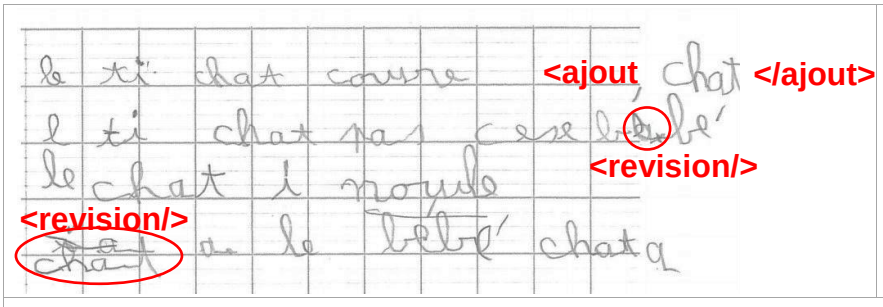
Sont considérés comme éléments génétiques les éléments qui témoignent des étapes successives de correction et de révision du texte. Cela peut se traduire par des ratures, des ajouts ou toute autre trace de modification.

<revision/>

La balise **<revision/>** est utilisée pour signaler des traces de révisions, comme les réécritures, les ratures et les traces de gomme.

<ajout>*ElementAjouté***</ajout>**

La balise **<ajout>***ElementAjouté***</ajout>** est utilisée pour signaler les ajouts, de quelque nature que ce soit.



le ti chat coure // l ti chat pas c ese b<revision/>ébé <ajout>chat</ajout> / le chat i nouile // <revision/>chat a le bébé chat a

Scan et transcription de la production de texte de CP de l'élève 1363

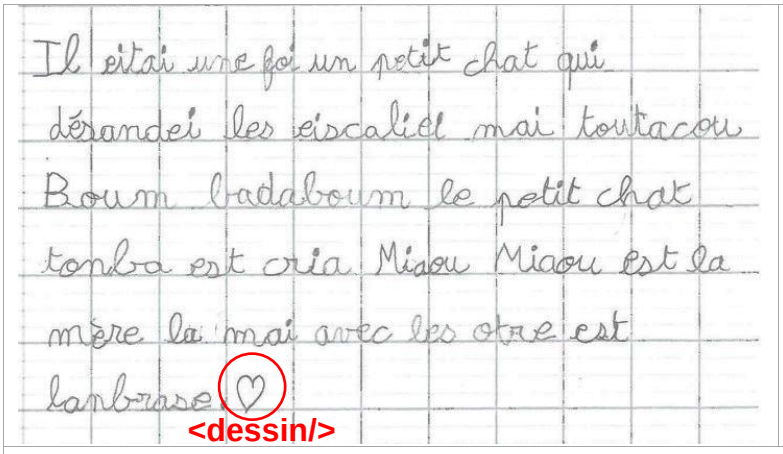
Transcription des éléments extratextuels

Sont considérés comme éléments extratextuels les éléments visuels et méta-textuels.

Sont considérés comme éléments visuels essentiellement les dessins, ajoutés par les apprenant·e·s à l'intérieur de leur texte ou à la suite de leur texte.

<dessin/>

La balise **<dessin/>** est utilisée pour signaler la présence d'un dessin dans la production de texte.



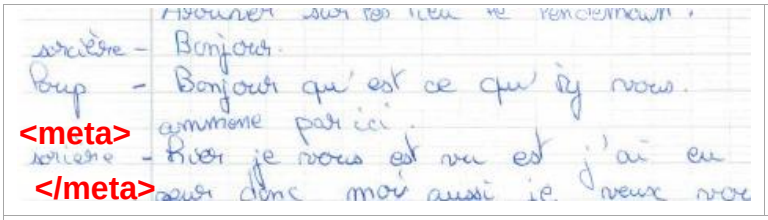
Il eitai une fois un petit chat qui / désandei les eiscaliei mai toutacou / Boum badaboum le petit chat / tomba est cria Miaou Miaou est la / mère la mai avec les otre est / lanbrase. **<dessin/>**

Scan et transcription de la production de texte de CP de l'élève 2413

Sont considérés comme éléments méta-textuels des éléments visibles sur le scan et qu'il parait important de reporter car porteurs d'informations mais qui ne seront pas pris en compte dans l'analyse.

<meta/>

La balise **<meta>***ElementConcerné*</meta>** permet de signaler un élément visible mais qui ne sera pas pris en compte dans le traitement du contenu textuel, mais elle permet également de rendre visible des pans de production ou des données linguistiques qui ne seront pas pris en compte dans notre analyse, mais qui peuvent l'être pour d'autres analyses.



<meta>sorcière</meta> - Bonjour. / <meta>loup </meta> - Bonjour qu'est ce qu'<unsure>iy</unsure> vous / ammene par ici. / <meta>sorcière</meta> - hier

Scan et transcription de la production de texte de CM2 de l'élève 103

Dans cet exemple, le nom des locuteurs est donné dans la marge, afin qu'ils ne soient pas pris en compte dans l'analyse syntaxique, on les entoure de la balise <meta/>.

<empty/>

La balise **<empty/>** est utilisée pour signaler une production vide, c'est-à-dire une production pour laquelle l'apprenant était présent mais n'a rien produit.

Transcription des éléments textuels

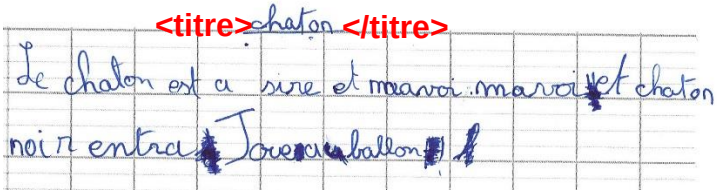
Sont considérés comme éléments textuels les éléments qui traient au contenu textuel et linguistique des productions.

Caractéristiques linguistiques et textuelles

La transcription conserve les choix textuels effectués par l'apprenant tels que l'orthographe, le lexique, la syntaxe, etc.

<Titre>TitreDeLaProduction**</titre>**

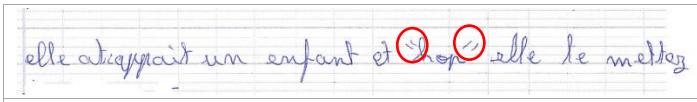
La balise **<titre>***TitreDeLaProduction***</titre>** est utilisée pour signaler la présence d'un titre en encadrant celui-ci.

	**<titre>**chaton**</titre>** Le chaton est a sire et mavoi mavoi <revision/>et chaton / noi r entra <revision/> Joue<revision/> a <revision/> ballon<revision/> <revision/>

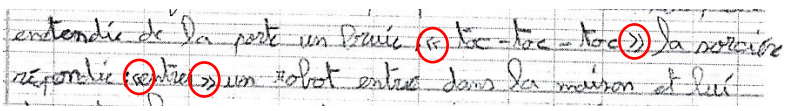
Scan et transcription de la production de texte de CE1 de l'élève 1283

Marque de dialogue

La transcription conserve les choix des marques de dialogue effectués par l'apprenant et admet les guillemets dits *français* « et » et les guillemets dits *anglais* " et ".

	[…] mieux / elle atrappait un enfant et "hop" elle le mettez / […]

Extrait du scan et de la transcription de la production de texte de CE2 de l'élève 855

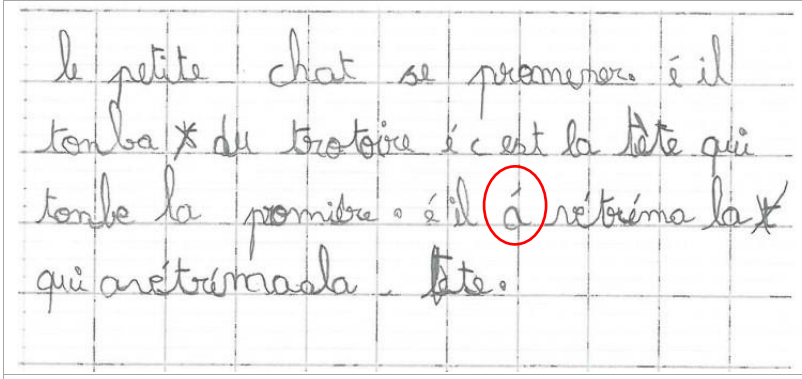
	[…] entendie de la porte un bruie : « toc-toc-toc » la sorciére / répondie : « entrer » un robot entre dans la maison et lui […]

Extrait du scan et de la transcription de la production de texte de CE2 de l'élève 3003

Signes diacritiques

La transcription conserve les choix d'accentuation ou non des majuscules par l'apprenant et admet les majuscules accentuées telles que É ou À.

La transcription conserve les erreurs d'accentuation sur la lettre *a* et admet le caractère *á*.



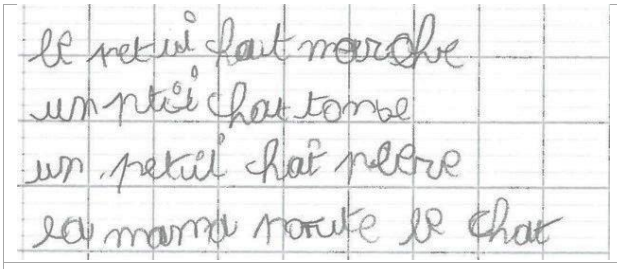
le petite chat se promener. é il / tonba <revision/> du trotoire é c est la tète qui / tonbe la promiére. é il **á** vétréma la <revision/> / qui avétrémaala <revision/>tète.

Scan et transcription de la production de texte de CP de l'élève 1573

Changements de ligne et changements de page

Le symbole **/** est utilisé pour signaler un retour à la ligne induit par la fin du support physique, c'est-à-dire la fin de la feuille.

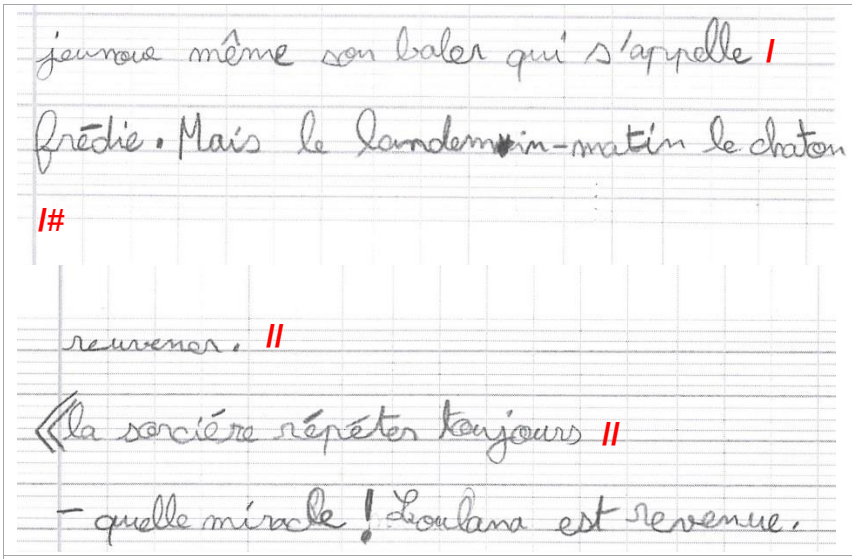
Les symboles **//** sont utilisés pour signaler un retour à la ligne volontaire, produit avant la fin du support physique, c'est-à-dire avant la fin de la feuille.



le petui chat m<revision/>arche // un ptii chat tonbe // un petii chat plere // la mama porite le chat

Scan et transcription de la production de texte de CP de l'élève 218

Le symbole **#** est utilisé pour signaler un changement de page.



[…] jeunoue même son baler qui s'appelle **/** frédie. Mais le landemin-matin le chaton **/#** reuvener. **//** « la sorciére répéter toujours **//** - quelle miracle ! Loulana est revenue.

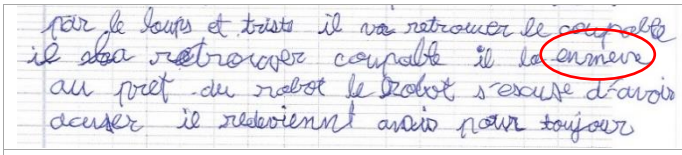
Extrait du scan et de la transcription de la production de texte de CE2 de l'élève 1339

Transcription de l'incertitude du transcripteur ou de la transcriptrice

Sont considérés comme moments d'incertitude les passages jugés illisible par la ou le transcripteur ou les passages pour lesquels la ou le transcripteur n'est pas sûr de sa transcription.

<unsure>*ElémentPeuSûr***<unsure>**

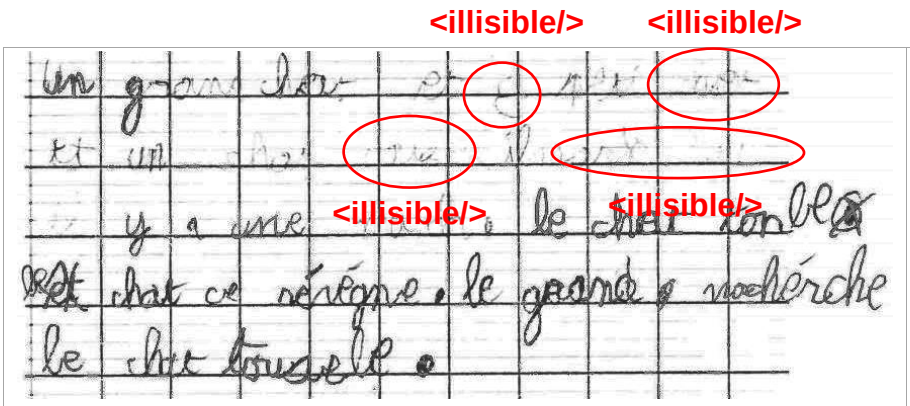
La balise **<unsure>*ElémentPeuSûr*<unsure>** est utilisée pour signaler un passage jugé peu sûr au regard de la fiabilité de la transcription.

	[…] il <revision/>la retrouver coupable il la **<unsure>**enmene**</unsure>** / / […]

Extrait du scan et de la transcription de la production de texte de CE2 de l'élève 239

<illisible/>

La balise **<illisible/>** est utilisée pour signaler un passage jugé illisible par la ou le transcripteur.

	un grand chat et **<illisible/>** petit **<illisible/>** / et un chat **<illisible/>** il **<illisible/>** / y a une march. le chat tonbl<revision/> / et chat ce révégne. le grande vachérche / le chat tousele.

Scan et transcription de la production de texte de CP de l'élève 207

Dans cet exemple, il est possible de deviner certains passages à demi effacés. Néanmoins, certains passages ne peuvent pas être déchiffrés, il convient donc de les relever à l'aide de la balise <illisible/>

Annexe 6 : Guide de normalisation

Guide de normalisation

Rédaction : *Catherine Brissaud*
Proposition issue du travail collectif : Catherine Brissaud, Claude Ponton, Lilia Teruggi, Corinne Totereau, Claire Wolfarth

Éléments retenus
Ne sont retenus dans la normalisation que les éléments textuels produits par l'enfant : les mots, ou groupes de lettres, et la ponctuation.
Dans le cas où le texte est écrit en majuscules, la normalisation se fait en minuscules (la majuscule est un marqueur de phrase).
Le retour à la ligne intentionnel, noté // dans la transcription (qui marque un nouveau paragraphe) est conservé et noté **<p/>** dans la version normalisée.
Éléments non retenus
Sont absentes de la normalisation les balises et autres marques spécifiques à la transcription, tels que : <dessin/> (présence de dessin) <revision/> / (retour à la ligne en fin de ligne) /# (changement de page).
La balise suivante est modifiée lors de la normalisation : <illisible> se transforme en <incomprehensible/>

1. Segmentation en mots, orthographe et accords

1.1. Segmentation en mots

La segmentation en mots est rétablie.
En cas d'hyposegmentation (agglutination d'un ou de plusieurs mots)
- il lepren (563, CP) - [normalisation] il le prend
- le chatpleur (78, CP) - [normalisation] le chat pleure
En cas d'hypersegmentation (séparation d'un mot en plusieurs mots ou unités)

- il et te un foin (1127, CP) - [normalisation] Il était une fois

1.2. Orthographe

L'orthographe est normée.
- Il etait une foie (568, CP) - [normalisation] Il était une fois
L'orthographe des onomatopées et interjections les plus courantes est normée, en minuscules. On conserve le point d'exclamation (un seul) après l'onomatopée seulement dans le cas où il marque la fin d'une phrase.
- comme par surprise BOOM le loup le mangea !!! (825, CM1) - [normalisation] comme par surprise boum le loup le mangea !
En cas de mot concerné par les rectifications orthographiques de 1990, le choix de l'élève est respecté.
- Ses maitres etait trés gentille. (1557, CE1) - [normalisation] Ses maitres étaient très gentils. - c'est son pére quil fait le maître. (2937, CE1) - [normalisation] c'est son père qui fait le maître.

1.3. Les accords

Les accords en genre, en nombre et en personne sont rétablis.
- et les chaton aussi (1143, CP) - [normalisation] et les chatons aussi
- pourquoi tu est venu (1866, CE2) - [normalisation] pourquoi tu es venu

1.4. La morphologie verbale

La morphologie normée est rétablie.
- il entendu un bruit il alla se cacher dans les buisson (94, CE2) - [normalisation] il entendit un bruit, il alla se cacher dans les buissons
En cas de régularisation de verbes irréguliers (notamment au passé simple), la forme normée est rétablie.
- elle prena le petit cha (584, CP) - [normalisation] elle prit le petit chat

2. Cohésion et utilisation des temps

L'utilisation des temps verbaux n'est pas modifiée.
- la Maman se révéilla. Et alait chairchai son petit. (83, CP) - [normalisation] la maman se réveilla et allait chercher son petit.
Le critère de l'oral l'emporte sur celui de l'écriture en cas d'homophonie. Par exemple, dans le cas suivant, on rétablit l'imparfait qui correspond à l'intention de l'auteur et que l'élève a orthographié *er*.
- il tombe et il **pleurer**. (586, CP) - [normalisation] Il tombe et il **pleurait**.

3. Le plan lexical

3.1. Omission d'un mot

S'il est identifiable de manière non ambigüe, le mot manquant est rétabli.
- mai il a pas vus le trautoir (562, CP) - [normalisation] Mais il **n'**a pas vu le trottoir.
- la maman le pran par bouche (200, CP) - [normalisation] la maman le prend par **la** bouche
S'il est non identifiable ou que les possibilités sont multiples, mais que la catégorie du mot absent est identifiable sans ambigüité, il est marqué par une balise **<omission type="verbe"/>**. Les catégories retenues pour cette balise sont les suivantes : **verbe**, **nom**, **pronom**, **adjectif** (qualificatif), **adverbe**, **préposition**, **déterminant**.
- il été t'une foi un petit chat se promené acoté de sa litière (1055, CP) - [normalisation] Il était une fois un petit chat **<omission type="pronom"/>** se promenait à côté de sa litière[111]
- Le petit chat boum et tou ta tou / le petit chat se mi a miau miaue (114, CP) - [normalisation] Le petit chat **<omission type="verbe"/>** boum et tout à coup le petit chat se mit à **<omission type="verbe"/>** miaou miaou
Dans le cas où il manquerait deux mots (par exemple déterminant et nom) on nomme la principale catégorie : le nom dans le cas suivant.

[111] Dans cet exemple, une reprise pronominale est nécessaire pour lier la première proposition « Il était une fois un petit chat » et la deuxième proposition « se promenait à côté de sa litière ». Plusieurs possibilités peuvent être envisagées :

- Il était une fois un petit chat *qui* se promenait à côté de sa litière
- Il était une fois un petit chat*, il* se promenait à côté de sa litière

Pour ne pas avoir à choisir entre ces différentes possibilités, on indique la nécessité de la reprise pronominale par la balise <omission type="pronom"/>.

- sa révéill la / maman vien laidai. (562, CP) - [normalisation] ça réveille la maman <omission type="pronom"/> vient l'aider.

3.2. Répétition ou ajout d'un mot

Lorsqu'un mot est identifié comme surnuméraire de manière non ambigüe et de manière non stylistique, il est supprimé.
- le chat pannicai le chat le chat dit « c'est pour quoi » (63, CE2) - [normalisation] le chat paniquait <segmentation/> le chat dit <dialogue> « c'est pourquoi » </dialogue>
Lorsque la répétition d'un mot ou groupes de mots est vue comme un procédé stylistique d'accentuation, elle est conservée.
- le loup avait faim très très faim (1155, CM1) - [normalisation] le loup avait faim très très faim

3.3. Choix lexicaux

Dans le cas de mots non conventionnels, par exemple repris à la littérature de jeunesse, on respecte le choix de l'élève.
- Quand / il chercher de la nouriture il ne s'avait pas / chasser, du cou il demander à c'est amis qui / l'aider toujour à chasser des petites sourissette / ai des oiseaux. (586, CE1) - [normalisation] Quand il cherchait de la nourriture il ne savait pas chasser <s/> du coup il demandait à ses amis qui l'aidaient toujours à chasser des petites souricettes et des oiseaux.
On conserve aussi le lexique très familier.
- le petit chat se mi a miau miaue / Et il se ro refou en bébé (114,CP) - [normalisation] tout à coup le petit chat se mit à <omission type="verbe"/> miaou miaou et il se refout en bébé.
Lorsqu'un mot n'existe pas, ou est difficile à interpréter, il est signalé par les balises <unsure>.
- Petits chat cour tranquileman pui il tonb parrtre / li plere miaou miaou, et quel-quin vin le / réconcolier. (500, CE1) - [normalisation] Petit chat court tranquillement puis il tombe par terre, il pleure miaou miaou, et quelqu'un vient le <unsure> réconcolier </unsure>.

4. Les constructions issues de l'oral

Les constructions issues de l'oral sont conservées (détachement, non accord du présentatif *c'est*).
- Le cha il mache (61, CP) - [normalisation] Le chat il marche
- c'est leurs maitresses et le maitre. (114, CE1) - [normalisation] C'est leurs maitresses et le maitre[112]

5. Syntaxe et ponctuation

5.1. Organisation des mots dans la proposition

En cas d'inversion de mots ou de construction verbale erronée, l'ordre attendu n'est pas rétabli.

5.2. La ponctuation

La majuscule en début de production et le point en fin de production sont rétablis.
- un peticha mareche il tonbe de / la marche il a male samaman / la mène avec ce peté (213, CP) - [normalisation] Un petit chat marche, il tombe de la marche et il a mal <segmentation/> sa maman l'amène avec ses petits.
Dans le cas où un texte s'ouvre sur un titre, la majuscule est rétablie pour le titre. Elle est rétablie également pour le premier mot qui suit le titre.
- les chaton un chat et des sendu de la maman il ettondé (1336, CP) - [normalisation] <titre>Les chatons</titre> Un chat est descendu de la maman <segmentation/> il est tombé
En cas d'interrogation directe, le point d'interrogation est rétabli (mais pas la majuscule qui suit).
- - bon jour la sorsiere commens allé vous. (1927, CE2) - [normalisation] <dialogue> - Bonjour la sorcière <s/> comment allez-vous ? </dialogue>
Lorsqu'un point de fin de phrase (. ? ou !) est présent sans majuscule, la majuscule est rétablie.
- le cha et tonbe sur le tapi. le cha pler (48, CP) - [normalisation] le chat est tombé sur le tapis. Le chat pleure

[112] Cas relevant des tolérances orthographiques et grammaticales de 1976. http://pernoux.pagesperso-orange.fr/tolerances.pdf

Lorsqu'une ponctuation forte est manquante et que le mot suivant contient une majuscule, le point est rétabli.
- Le petit chat sest fé male Le peti chat a male (1531, CP) - [normalisation] Le petit chat s'est fait mal. Le petit chat a mal
Les points surnuméraires sont supprimés.
- un chat civoulé fér une pet balde. il sentrafre. ése fémale. Est soudin sa maman se révéie. Ee raméne. (204, 1P) - [normalisation] Un chat qui voulait faire une petite balade. Il s'entrave et se fait mal. Et soudain sa maman se réveille et <"omission type "pronom"> ramène.
- et il lui dit. bonjours sorsier, vien chez / mon on va boire du thé. bon dacore /(69 CE1) - [normalisation] et il lui dit <dialogue> bonjour sorcière <s/> viens chez moi <s/> on va boire du thé </dialogue>. <dialogue> Bon d'accord </dialogue> - il était une fois dans le bois. Le petit chaperon rouge. Avait rancontré un robo. (103, CE1) - [normalisation] Il était une fois dans le bois. Le petit chaperon rouge avait rencontré un robot.
En cas de doute, dans les cas discutables, on ne modifie pas la ponctuation (subordonnée isolée).
- La sorsillere et tré en colére à/pré le chat. Parsece le chat à / casé sa boule ma gice. (204, 2P) - [normalisation] La sorcière est très en colère après le chat. Parce que le chat a cassé sa boule magique.

5.3. La segmentation du texte

La balise <segmentation/> est utilisée pour signaler/indiquer un nouveau « fragment textuel » qui correspond à un **changement de thématique/épisode/situation**, repérable notamment dans le changement de détermination (de l'indétermination à la détermination), du sujet grammatical, des temps verbaux.
- la maman chat grdre set 3 petite chat un chat vaver la maman chat (670, CP) - [normalisation] La maman chat garde ses 3 petits chats <segmentation/> un chat va vers la maman chat
a. Changement de détermination : d'un déterminant indéfini à un déterminant défini
- Il y a une corsiere qui abite dans un grand châtaux sonbre elle aver un chat noir et elle se trouve moche (48, CE1) - [normalisation] Il y a une sorcière qui habite dans un grand château sombre <segmentation/> elle avait un chat noir et elle se trouve moche

b. De la présentation d'un procès duratif indéterminé à une action plus ponctuelle
- Il était une foi un petit chat qui marché tranqileman toutaquou ! il trebucha et il se mis a pleurer. la maman chat l'antendi et elle le consola. (493, CP) - [normalisation] Il était une fois un petit chat qui marchait tranquillement <segmentation/> tout à coup il trébucha et il se mit à pleurer. La maman chat l'entendit et elle le consola.
c. Passage d'une présentation au début de la narration
- il este TUN foi un chat qui a vé BOCOU DE chat le chat sonfui a ton D il sai fai mal (1152, CP) - [normalisation] Il était une fois un chat qui avait beaucoup de chats <segmentation/> le chat s'enfuit, est tombé et il s'est fait mal
d. Passage de la présentation de la situation à la présentation des personnages
e. Séparer la narration de la fin de l'histoire ou séparer la narration de la formule de clôture ou du mot « fin ».
- et il sont ariver dans la maisson / et il devinre amis fin. (66, CE1) - [normalisation] et ils sont arrivés dans la maison et ils devinrent amis <segmentation/> fin. - … et le rolacha tonbe / par tr et se fait tretrémal et vois / la fin de li stor. (864, CP) - [normalisation] … et le relâcha <segmentation/> <omission type="pronom"/> tombe par terre et se fait très très mal <segmentation/> et voilà fin de l'histoire.
La balise **<segmentation/>** n'est pas utilisée dans le cas où le sujet grammatical est repris sous forme de pronom ou à l'identique (ou presque) reste le même. Dans ce cas, on le traite comme une énumération et on utilise la virgule ou « et » dans le cas où deux éléments seraient juxtaposés.
- il cur il tonbe partere il pleur il saute sur un notre chat (1173, CP) - [normalisation] il court, il tombe par terre, il pleure et il saute sur un autre chat. - mes mon chat qui été dans la marmite a eu telleman pere que il est sorti de la marmote a monter lechelle et courer dans touts sanse. (2978, CE1) - [normalisation] mais mon chat qui était dans la marmite a eu tellement peur qu'il est sorti de la marmite, a monté l'échelle et courait dans tous les sens.
Les connecteurs valent pour marque de segmentation en propositions, la balise <segmentation/> ne sera donc pas ajoutée quand les connecteurs « et », « puis », « après », « ensuite », « à la fin », etc. sont utilisés.
- Le chat sailoigne de sa maman est des bébé pan dans que sa maman dor ansuit il tonbe est le bébé pleur est sa maman lanten pui sa maman ramas le chton. (1665, CP)

- [normalisation] Le chat s'éloigne de sa maman et des bébés pendant que sa maman dort ensuite il tombe et le bébé pleure et sa maman l'entend puis sa maman ramasse le chaton.
- Le chat coure et aprer il toube et aprer il pleur et la maman seut rerverlle et aprer la maman elle la porte (81-CP)
- [normalisation] Le chat court **et après** il tombe **et après** il pleure **et** la maman se réveille **et après** la maman elle le porte.

Le retour à la ligne intentionnel vaut ponctuation forte. On ne rajoute donc pas de balise <segmentation> dans ce cas mais la balise <p/> qui signifie « retour à la ligne » (mais sans retour à la ligne dans la version normalisée). Cela vaut pour un titre détaché de la suite du texte.

- Le chat eon ilavan // Le chateon et blécé // Le chateon i pler // Le hat itin (58, CP)
- [normalisation] Le chaton il avance. <p/> Le chaton est blessé. <p/> Le chaton il pleure. <p/> Le chat <incomprehensible/>

5.4. 5.4. Le cas de la virgule

L'énumération est le seul cas où une virgule peut être ajoutée. Il peut s'agir d'une énumération de noms, d'adjectifs ou de verbes. Dans tous les autres cas, on n'ajoute pas de virgule.
Quand l'énumération comporte deux éléments, on remplace la virgule par le conjonction « et ». Si elle en comporte plus de deux, on ajoute des virgules et la conjonction « et » entre le dernier et l'avant-dernier élément de l'énumération.

- Le chat désan les éscalié il tonb sa maman / se révéle et la maman lui fai un / calin. (66-CP)
- [normalisation] Le chat descend les escaliers **et** il tombe <segmentation/> sa maman se réveille et la maman lui fait un câlin.

Dans le cas d'une énumération de propositions, le sujet grammatical reste le plus souvent le même, est sous-entendu ou repris sous forme de pronom.

- Le peti cha senva il tond il révége / sa maman sa maman le souoige. (97-CP)
- [normalisation] Le petit chat s'en va, il tombe et il réveille sa maman <segmentation/> sa maman le soigne.
- Le petit chat désen las marche il / tonbe des marche le chat die aei // le chat fais des bébé. (102, CP)
- [normalisation] Le petit chat descend la marche**,** il tombe des marches **et** le chat dit <dialogue> aïe </dialogue> le chat fait des bébés.

- Le petit chat sèloign de sa maman il tonbe il a mal / sa maman <revision/> vin le secourire (502, CP) - [normalisation] Le petit chat s'éloigne de sa maman, il tombe et il a mal <segmentation/> sa maman vient le secourir.
- me elle ne ses pa fair de la magi et elle mai des machin truque des soupe avec des sier de oiseaux et des cou de lesar et des quocille des éscarco (48, CE1) - [normalisation] mais elle ne sait pas faire de la magie et elle met des machins trucs, des soupes avec des serres d'oiseaux et des cous de lézard et des coquilles d'escargots
- le robot étté janti tres janti. (107, CE1) - [normalisation] Le robot était gentil très gentil.
Dans le cas d'un emploi impropre de la virgule, celle-ci est supprimée (par exemple après « il était une fois » ou entre sujet et prédicat).
- il était une fois dans le bois. Le petit chaperon rouge. Avait rancontré un robo. (103, CE1) - [normalisation] Il était une fois dans le bois. Le petit chaperon rouge avait rencontré un robot. - le chat noir c'est, enfuit de sa cage. (70, CE2) - [normalisation] le chat noir s'est enfui de sa cage.
- Il était une fois, une sorciére qui avait un chat noir qui s'appelait Coco. (1826, CE1) - [normalisation] Il était une fois une sorcière qui avait un chat noir qui s'appelait Coco.
Les emplois discutables (« limites ») de la virgule sont conservés.
- et il vaicure, heureux jusque à la fain des tant. (100, CE1) - [normalisation] et ils vécurent heureux jusqu'à la fin des temps. - Il atrapa le chat, et le ramena chez lui. (49, CE1) - [normalisation] Il attrapa le chat, et le ramena chez lui.

5.5. La balise <s/>

On utilise la balise **<s/>** quand on ressent la nécessité d'ajouter un signe de ponctuation tout en hésitant entre plusieurs signes (virgule, deux points ou point ou encore la conjonction « et »).
C'est le cas, par exemple, quand un lien de cause à effet unit deux propositions.
- Il pler sa maman vien le / chré (1125 CP) - [normalisation] Il pleure **<s/>** sa maman vient le chercher.
On l'utilise aussi pour l'introduction d'une explication.

On peut l'utiliser entre deux propositions ayant un sujet grammatical différent, y compris quand ils désignent un même personnage.
- Tagbot ouvri la porte / le chat se reveilla et fi plein / de calin à Tagbot. - [normalisation] Tagbot ouvrit la porte **<s/>** le chat se réveilla et fit plein de câlins à Tagbot.
On l'utilise aussi dans les dialogues pour séparer une réponse à une demande de la proposition qui suit.
- il était une fois dans le bois. Le petit chaperon rouge. Avait rancontré un robo. bonjours je me suis perdu dans le bois ! (103, CE1) - [normalisation] Il était une fois dans le bois. Le petit chaperon rouge avait rencontré un robot. <dialogue> Bonjour <s/> je me suis perdu dans le bois ! </dialogue>

6. Traitement des dialogues

Chaque fois qu'un personnage prend la parole, ses propos sont encadrés par les balises **<dialogue>** (au début) **</dialogue>** (à la fin).
- le chat dit : Bonjour méchante et orible sorciere j'ai apris que tu allais dominer le monde. Oui c'est sa et comment t'appelle tu je m'appelle jean - pomme de terre. Sa te dit con fase équipe ensemble […] (1336, CM1) - [normalisation] le chat dit : <dialogue> Bonjour méchante et horrible sorcière <s/> j'ai appris que tu allais dominer le monde. </dialogue> <dialogue> Oui c'est ça et comment t'appelles-tu ? </dialogue> <dialogue> Je m'appelle Jean-Pomme de terre. </dialogue> <dialogue> Ça te dit qu'on fasse équipe ensemble ? </dialogue> […]
Les ponctuations de dialogue correctes existantes sont conservées ; les autres sont retirées. Les ponctuations manquantes, comme la virgule, le point, le double-point, les guillemets ou encore les tirets, ne sont pas rétablies, sauf lorsqu'une marque double (ex. guillemets) est incomplète ou mal positionnée. Dans ce cas, la deuxième marque est rétablie.
- elle a dit «» je vait le raiqupérait. est anfin elle a réqupérait con chat. (937, CE1) - [normalisation] elle a dit <dialogue> « je vais le récupérer » </dialogue>. Et enfin elle a récupéré son chat.
La proposition incise à l'intérieur d'un dialogue est sortie du dialogue : la balise <dialogue> est fermée devant l'incise puis rouverte après.
- Aléde ! dit le chat. // quesquilia dit le loup. // J'ai peur de toi. - [normalisation] <dialogue> À l'aide ! </dialogue> dit le chat. <p/> <dialogue> qu'est-ce qu'il y a ? </dialogue> dit le loup. <p/><dialogue> J'ai peur de toi. </dialogue>
Les pensées ou idées des personnages sont également encadrées de balises <dialogue>.
- elle eu une idée et si je me transphorma en chat !!! (1111, CM1)

- [normalisation] elle eut une idée <dialogue> et si je me transformais en chat ! </dialogue>
La balise <dialogue> implique une segmentation, la balise <segmentation/> est donc inutile dans le voisinage de la balise <dialogue>.
- il alla voire la sorsier / et il lui dit. bonjours sorsier, vien chez / mon on va boire du thé. bon dacore / et le loup une foit chez lui il / menja la sorsier. // fin (69, CE1) - [normalisation] Il alla voir la sorcière et il lui dit <dialogue> bonjour sorcière <s/> viens chez moi on va boire du thé </dialogue>. <dialogue> Bon d'accord </dialogue> et le loup une fois chez lui il mangea la sorcière. <p/> Fin.

7. Notations spécifiques

Lorsque le texte est clairement non achevé en fin de production, on le marque à l'aide de la balise **<nonfini/>**.
- et di maou maou qui r (1228, CP) - [normalisation] et dit <dialogue> miaou miaou </dialogue> qui r**<nonfini/>**
Les segments dont le sens ne peut être compris sont remplacés par la balise **<incomprehensible/>.**
- l ti chat pas c ese bébe cha (1363, CP) - [normalisation] le petit chat **<incomprehensible/>** bébé chat
Les titres en début de production sont encadrés par la balise **<titre>**.
- Le chat Il était une fois un chat (1845, CE2) - [normalisation] **<titre>** Le chat **</titre>** Il était une fois un chat
En cas de doute du normalisateur, on peut introduire une possible interprétation en l'encadrant par la balise **<unsure> … </unsure>**.
- Petits chat cour tranquileman pui il tonb parrtre / li plere miaou miaou, et quel-quin vin le / réconcolier. (500, CE1) - [normalisation] Petit chat court tranquillement puis il tombe par terre, il pleure miaou miaou, et quelqu'un vient le <unsure> réconcolier </unsure>.

Annexe 7 : French *TreeTagger* Part-of-Speech Tags

Liste des étiquettes de catégories utilisées par *TreeTagger*, d'après Achim Stein, 2003 (https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/ french-tagset.html).

Étiquette	*Signification*
ABR	abreviation
ADJ	adjective
ADV	adverb
DET:ART	article
DET:POS	possessive pronoun (ma, ta, ...)
INT	interjection
KON	conjunction
NAM	proper name
NOM	noun
NUM	numeral
PRO	pronoun
PRO:DEM	demonstrative pronoun
PRO:IND	indefinite pronoun
PRO:PER	personal pronoun
PRO:POS	possessive pronoun (mien, tien, ...)
PRO:REL	relative pronoun
PRP	preposition
PRP:det	preposition plus article (au,du,aux,des)
PUN	punctuation
PUN:cit	punctuation citation
SENT	sentence tag
SYM	symbol
VER:cond	verb conditional
VER:futu	verb futur
VER:impe	verb imperative
VER:impf	verb imperfect
VER:infi	verb infinitive
VER:pper	verb past participle
VER:ppre	verb present participle
VER:pres	verb present
VER:simp	verb simple past
VER:subi	verb subjunctive imperfect
VER:subp	verb subjunctive present

Annexe 8 : Table des correspondances des formats de transcription phonologique

Liste des correspondances entre les formats API, SAMPA et LIA pour les phonèmes du français (utilisés dans l'outil *LIA-PHON*).

Format API	*Format Sampa*	*Format LIA*
i	i	ii
e	e	ei
ɛ	E	ai
a	a	aa
ɔ	O	oo
o	o	au
u	u	ou
y	y	uu
ø	2	eu
œ	9	oe
ə	@	ee
ɛ̃	e~	in
ɑ̃	a~	an
ɔ̃	o~	on
œ̃	9~	un
j	j	yy
w	w	ww
ɥ	H	uy
p	p	pp
t	t	tt
k	k	kk
b	b	bb
d	d	dd
g	g	gg
f	f	ff
s	s	ss
ʃ	S	ch
v	v	vv
z	z	zz
ʒ	Z	jj
l	l	ll
ʁ	R	rr
m	m	mm
n	n	nn
ŋ	N	ng

Annexe 9 : Liste des correspondances graphophonémiques établie par J. Riou et R. Goigoux (2017)

Ce document reproduit la liste des correspondances graphophonémiques (CGP) établie par J. Riou et R. Goigoux (2017). Cette liste est disponible à l'adresse http://anagraph.ens-lyon.fr/docs/Riou-J&Goigoux-R-2016-A-3-4-1-4-Planification-de-l-etude-du-code-et-usage-des-manuels.pdf.

Relevé des CGP étudiées		
Correspondances Graphèmes-Phonèmes		Exemples
a (à, â)	[a] ou [ɑ]	avocat, patte, pâte
ai	[ɛ] ou [e]	aimer, serai
ain (aim)	[ɛ̃]	pain, daim
an (am)	[ã]	enfant, ampoule
au	[o] [ɔ]	autant, autobus
b	[b]	bleu
c (cc)	[k]	couleur, accomplir
c (ç)	[s]	cerise, déçu
ch	[ʃ]	chat
d	[d]	deux
e	[ə] [oe] [Ø]	devenir, lime
e	[e]	dessin, effort
e	[ɛ]	vert, mer
é (e)	[e]	école, été
è (ê)	[ɛ]	espèce, rêve
eau	[o]	bureau
ei (ey)	[e] ou [ɛ]	reine, hockey
ein (eim)	[ɛ]	rein, Reims
en	[ɛ]	rien
en (em)	[ã]	enfant, emporter
er (et)	[e] ou [ɛ]	manger, jouet
es (ez)	[e] ou [ɛ]	tu es, nez
eu	[oe] ou [Ø]	peur, bleu
eu	[Y]	j'ai eu
f (ff)	[f]	fou, effort
g(e)	[ʒ]	plage, plongeon
g (gu)	[g]	goutte, guépard
n	ɲ	mignon
i	[i]	ami

i	[j] (yod)	avion
ill (il)	[j] (yod)	œil, famille
in (im)	[ɛ̃]	sapin
j	[ʒ]	jaune
k	[k]	kilo
l (ll)	[l]	stylo, bulle
m (mm)	[m]	mon, femme
n (nn)	[n]	niche, bonne
o (ô)	[o] [ɔ]	stylo, école
oeu (oe)	[œ] ou [Ø]	cœur, vœu, œil
oo	[o] [u]	zoo, foot
oi	[wa]	roi
oin	[wɛ̃]	coin
on (om)	[õ]	salon, pompe
ou (où)	[u]	fou, où
p (pp)	[p]	père, apporte
ph	[f]	phoque
q (qu)	[k]	coq, quatre
r (rr)	[R]	rouge, arrêt
s (ss)	[s]	soir, assis
s	[z]	usé
t (tt)	[t]	toupie, belette
t	[s]	solution, patient
u	[Y]	usine
u	[ɔ]	album
u(i)	[ɥi]	pluie
un	[oɛ̃] [ɛ̃]	brun
v	[v]	voyage
w	[w]	Web
w	[v]	wagon
x	[gz] ou [ks]	examen, axe
y	[i]	analyse
y	[j] (yod)	voyage
z	[z]	zoo

Annexe 10 : Protocole d'alignement en graphèmes

Protocole d'alignement en graphèmes

Rédaction du protocole : Louise-Amélie Cougnon (MiiL, UCL) et Claire Wolfarth (Lidilem, UGA)

Unités d'alignement

L'aligneur d'alignement en graphèmes est un aligneur dont les unités sont les graphèmes et les ponctuants (signes de ponctuation, marques de dialogue, …).

A. Les graphèmes

Un **graphème** est une lettre ou un groupement de lettres correspondant à un phonème (phonogramme) ou à une absence de phonème (graphèmes muets). Pour ce travail, nous nous baserons sur la liste des graphèmes établie par Riou (2017, pp. 93-94 et lien, pp. 2-3) à partir des graphèmes les plus fréquents relevés dans les manuels scolaires.

Pour discriminer les graphèmes, la priorité est donnée aux sons (critère phonogrammique) plutôt qu'au sens (critère morphogrammique). Nous analysons les critères suivants (en ordre croissant de priorité) :

- *[Critère phonogrammique] Un graphème composé d'un groupement orthographique (une ou plusieurs lettres) correspond à un phonème (Peereman, 2007).*

  1. Lorsqu'une lettre unique correspond à un phonème (son) unique (exemple (1)), il faut la considérer comme un graphème.

     (1) sac → s – a – c ; /sak/ (3 graphèmes correspondant à 3 phonèmes)

  2. Lorsqu'un groupement de lettres correspond à un phonème (son) unique (*ai* et *in* dans l'exemple (2), *er, ez* et *es* dans les exemples (3), (4) et (5) et *ll* dans les exemples (6) et (7)), il faut le considérer comme un graphème unique.

     (2) raisin → r – ai – s – in ; /Rɛzɛ̃/ (4 graphèmes correspondant à 4 phonèmes)

     (3) marcher → m – a – r – ch – er ; /maRʃe/

     (4) nez → n – ez ; /ne/

     (5) est → es – t ; /ɛ/

     (6) bulle → b – u – ll – e ; /byl/

     (7) belle → b – e – ll – e ; /bɛl/

Lorsqu'une lettre qui peut être considérée comme muette dans d'autres contextes est à proximité de ce groupement sans être impliquée dans sa réalisation sonore (*t* dans les exemples (5) et (8)), il faut la considérer comme un graphème distinct. Lorsque cette lettre ne peut pas ou que très occasionnellement être considérée comme une lettre muette dans un autre contexte (*a* dans l'exemple (9) et *e* dans l'exemple (10)), il faut la considérer comme inclue dans le graphème.

(8) fait → f – ai – t ; /fɛ/

(9) pain → p – ain ; pɛ̃/

(10) eau → eau ; /o/

3. Lorsqu'une lettre unique correspond à plusieurs phonèmes (sons) distincts (par ex. *x* et *y* dans les exemples (11) et (12)), il faut la considérer comme un graphème.

(11) examina → e – x – a – m – i – n – a ; /ɛ-gz-amina/

(12) voyage → v – o – y – a – g – e ; /vwajaʒ/

4. Lorsqu'un groupement de lettres correspond à plusieurs phonèmes (sons) distincts (*gu*, *cc* ou *mm* dans les exemples (13) et (15) et (17)), il faut les considérer comme des graphèmes distincts, même s'ils constituent des graphèmes uniques dans d'autres contextes (*gu* ou *cc* dans les exemples (15) et (16)).

(13) aiguille → ai – g – u – i – ll – e (*gu* ; /gy/ : 2 graphèmes)

(14) anguille → an – gu – i – ll – e (*gu* ; /g/ : 1 graphème)

(15) vaccin → v – a – c – c – in (*cc* ; /ks/ : 2 graphèmes)

(16) accrédita → a – cc – r – é – d – i – t – a (*cc* ; /k/ : 1 graphème)

(17) emmena → em – m – e – n – a (e*mm* ; /ɑ̃m/ : 2 graphèmes)

Exception :

Lorsque les groupements de lettres *oi* et *oin* sont associés aux phonèmes (sons) /wa/ (exemple (18)) et /wɛ̃/ (exemple (19)), il faut les considérer comme des graphèmes uniques.

(18) loi → l - oi ; /lwa/

(19) loin → l – oin ; /lwɛ̃/

Lorsque le groupement de lettres *oy* est associé aux phonèmes (sons) /waj/ (exemple (20)), il faut le considérer comme deux graphèmes différents.

(20) voyage → v – o – y – a – g – e ; /vwajaʒ/

- *[Critère morphogrammique] Un graphème composé d'un groupement orthographique (une ou plusieurs lettres) correspond à une seule valeur morphologique.*

5. Lorsqu'une lettre muette est isolée, en début ou fin de mot (*e* dans l'exemple (21)) ou entre deux graphèmes sonores (*e* dans l'exemple (22)), il faut la considérer comme un graphème.

(21) marche → m – a – r – ch – e

(22) mangea → m – an – g – e – a

6. Lorsqu'un groupement de lettres muettes correspond à une même information morphologique, comme la personne (*ent* dans l'exemple (23)), il faut le considérer comme un graphème unique.

(23) vivaient → v – i – v – ai – ent

7. Lorsqu'un groupement de lettres muettes correspond à plusieurs valeurs morphologiques distinctes, comme le genre (*e* dans l'exemple (24)), le nombre (*s* dans l'exemple (24)), le mode, le temps (*e* dans l'exemple (25) ou *d* dans l'exemple (26)) ou la personne (*s* dans l'exemple (26)), il faut les considérer comme des graphèmes distincts.

(24) petites → p – e – t – i – t – e – s

(25) vivent → v – i – v – e – nt

(26) défends → d – é – f – en – d – s

B. Les ponctuants

Un **ponctuant** est un signe de ponctuation, tel que le point, la virgule, le point d'exclamation et le point de ponctuation ; une marque de dialogue, telle que les guillemets doubles, les guillemets français et les tirets ou un caractère de séparation, tel que l'apostrophe et le trait d'union.

1. Il faut considérer un ponctuant ou un groupement de ponctuants comme une unité d'alignement isolée des lettres (exemple (25)).

(27) c'est → c – ' – es – t

2. Lorsqu'un groupement de ponctuants est composé de ponctuants de même fonction (plusieurs signes de ponctuation successifs par exemple), il faut les considérer comme une seule unité d'alignement (exemple (28)).

(28) boum !!! → b – ou – m – !!!

3. Lorsqu'un groupement de ponctuants est composé de signes de ponctuation et de marques de dialogue, il faut les considérer comme des unités d'alignement distinctes (exemple (28)).

(29) !" → ! – "

C. Les balises

La normalisation comporte différentes balises :

- <p/> : cette balise marque le retour à la ligne volontaire, marquant ainsi un nouveau paragraphe. Elle ne doit être alignée avec aucun élément transcrit.

(30)

[normalisation] Un tigre l'a mordu parce qu'il l'a énervé. <p/> Le tigre

- <dialogue> </dialogue> : ces balises délimitent l'insertion d'un discours direct ou d'une parole dans un texte. Les balises sont ouvertes (<dialogue>, début de discours) et fermées (</dialogue>, fin de discours) à chaque tour de parole (exemple (31) et exemple (32)).

(31) elle lui dit "Que fait-tu la mon ami".

[normalisation] et elle lui dit <dialogue>"Que fais-tu là mon ami ?" </dialogue>.

(32) Pour quoi ne pas le fer fuir. quomen lui demande son fraire. Et bein alui faisen peure dis mimi

[normalisation] <dialogue> Pourquoi ne pas le faire fuir? </dialogue> <dialogue> Comment? </dialogue> lui demande son frère. <dialogue> Et bien en lui faisant peur </dialogue> dit Mimi

- <titre> </titre> : ces balises encadrent la présence d'un titre. Ces balises doivent être alignées avec leur équivalent transcrits s'il y a lieu, avec aucun élément sinon.

(33) <titre>le chat et la sorciere.</titre>

[normalisation] <titre>Le chat et la sorcière</titre>

- <segmentation/> et <s/> : dans une construction syntaxique traditionnelle, certaines pauses ou éléments sont marqués de ponctuants (par exemple le point ou la virgule) ou de connecteurs (par exemple *et* et *mais*) qui permettent de séparer et donc d'identifier des propositions distinctes. Lorsque ces ponctuants ne sont pas utilisés (ex. (34)) ou mal utilisés (exemple (35)) par les élèves, ils sont remplacés par les balises <segmentation/> et <s/>. En cas de ponctuation manquante, ces balises ne doivent être alignées avec

aucun élément transcrit. En cas de ponctuation fautive, ces balises doivent être alignées avec la ponctuation incorrecte.

(34) il était une fois une sorcière elle avait un chat magique

[normalisation] Il était une fois une sorcière <segmentation/> elle avait un chat magique

(35) le chat tombe. pleure et sa maman le ramène[113]

[normalisation] le chat tombe <s/> pleure et sa maman le ramène

- <omission type="xxx"/>, où xxx = {verbe, nom, pronom, adjectif, adverbe, préposition, déterminant} : cette balise marque l'absence d'un élément lexical nécessaire à la construction syntaxique ou à la compréhension. Elle ne doit être alignée avec aucun élément transcrit.

(36) sa révéill la maman vien laidai.

[normalisation] ça réveille la maman <omission type="pronom"/> vient l'aider.

- <incomprehensible/> : cette balise marque les passages qui n'ont pas pu être interprétés par la personne qui a normalisé. Cette balise doit être alignée avec le passage transcrit concerné (exemple (37)).

(37) pour qu'il protège les pèsitio

[normalisation] pour qu'il protège les <incomprehensible/>

- <unsure> </unsure> : cette balise encadre les passages pour lesquels la personne qui a normalisé n'est pas sûre. Elle ne doit être alignée avec aucun élément transcrit.

(38) La sorçiere metta une grenouille, des cailloux et une pice de tout en 1.

[normalisation] La sorcière mit une grenouille, des cailloux et une <unsure> pièce </unsure> de tout en 1.

- <nonfini/> : cette balise marque la présence d'une production inanchevée. Elle ne doit être alignée avec aucun élément transcrit.

(39) et di maou maou qui r

[normalisation] et dit <dialogue> miaou miaou </dialogue> qui r<nonfini/>

[113] Exemple construit.

D. Les espaces

Les espaces sont présents pour faciliter la lecture, ils ne sont pas à évaluer.

Résumé

Depuis peu, émerge une réelle dynamique de constitution et de diffusion de corpus d'écrits scolaires, notamment francophones. Ces corpus, qui appuient les travaux en didactique de l'écriture, sont souvent de taille restreinte et peu diffusés. Des corpus longitudinaux, c'est-à-dire réalisant le suivi d'une cohorte d'élèves et permettant de s'intéresser à la progressivité des apprentissages, n'existent pas à ce jour pour le français.

Par ailleurs, bien que le traitement automatique des langues (TAL) ait outillé des corpus de natures très diverses, peu de travaux se sont intéressés aux écrits scolaires. Ce nouveau champ d'application représente un défi pour le TAL en raison des spécificités des écrits scolaires, et particulièrement les nombreux écarts à la norme qui les caractérisent. Les outils proposés à l'heure actuelle ne conviennent donc pas à l'exploitation de ces corpus. Il y a donc un enjeu pour le TAL à développer des méthodes spécifiques.

Cette thèse présente deux apports principaux. D'une part, ce travail a permis la constitution d'un corpus d'écrits scolaires longitudinal (CP-CM2), de grande taille et numérisé, le corpus *Scoledit*. Par « constitution », nous entendons le recueil, la numérisation et la transcription des productions, l'annotation des données linguistiques et la diffusion de la ressource ainsi constituée. D'autre part, ce travail a donné lieu à l'élaboration d'une méthode d'exploitation de ce corpus, appelée *approche par comparaison*, qui s'appuie sur la comparaison entre la transcription des productions et une version normalisée de ces productions pour produire des analyses.

Cette méthode a nécessité le développement d'un aligneur de formes, appelé *AliScol*, qui permet de mettre en correspondance les formes produites par l'élève et les formes normalisées. Cet outil représente un premier niveau d'alignement à partir duquel différentes analyses linguistiques ont été menées (lexicales, morphographiques, graphémiques). La conception d'un aligneur en graphèmes, appelé *AliScol_Graph*, a été nécessaire pour conduire une étude sur les graphèmes.

Mots clés

Linguistique outillée ; Recueil de corpus d'écrits scolaires ; TAL et corpus scolaires ; Alignement automatique ; Apprentissage de l'écriture ; Production de texte

Abstract

In recent years, there has been an actual effort to constitute and promote children's writings corpora especially in French. The first research works on writing acquisition relied on small corpora that were not widely distributed. Longitudinal corpora, monitoring a cohort of children's productions from similar collection conditions from one year to the next, do not exist in French yet.

Moreover, although natural language processing (NLP) has provided tools for a wide variety of corpora, few studies have been conducted on children's writings corpora. This new scope represents a challenge for the NLP field because of children's writings specificities, and particularly their deviation from the written norm. Hence, tools currently available are not suitable for the exploitation of these corpora. There is therefore a challenge for NLP to develop specific methods for these written productions.

This thesis provides two main contributions. On the one hand, this work has led to the creation of a large and digitized longitudinal corpus of children's writings (from 6 to 11 years old) named the *Scoledit* corpus. Its constitution implies the collection, the digitization and the transcription of productions, the annotation of linguistic data and the dissemination of the resource thus constituted. On the other hand, this work enables the development of a method exploiting this corpus, called the comparison approach, which is based on the comparison between the transcription of children's productions and their standardized version.

In order to create a first level of alignment, this method compared transcribed forms to their normalized counterparts, using the aligner *AliScol*. It also made possible the exploration of various linguistic analyses (lexical, morphographic, graphical). And finally, in order to analyse graphemes, an aligner of transcribed and normalized graphemes, called Ali*Scol_Graph* was created.

Keywords

Corpus linguistics; Children's writings corpus collection; NLP and children's writings corpus; Automatic alignment; Writing acquisition; Writing