

Les acquisitions tardives en français écrit : une base de données sur les erreurs et maladresses à un niveau avancé

Françoise Boch, Fanny Rinck, Julie Sorba

▶ **To cite this version:**

Françoise Boch, Fanny Rinck, Julie Sorba. Les acquisitions tardives en français écrit : une base de données sur les erreurs et maladresses à un niveau avancé. 7e Congrès Mondial de Linguistique Française, 78, pp.06005, 2020, 10.1051/shsconf/20207806005 . hal-02931262

HAL Id: hal-02931262
https://hal.science/hal-02931262

Submitted on 5 Sep 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.

Les acquisitions tardives en français écrit : une base de données sur les erreurs et maladresses à un niveau avancé

Françoise Boch[1 A], *Fanny* Rinck[1], et *Julie* Sorba[1]

[1]Laboratoire Lidilem, Université Grenoble Alpes

Résumé. Inscrite dans le cadre des littéracies avancées, cette contribution présente les choix théoriques et méthodologiques à l'œuvre dans l'élaboration en cours d'une base de données sur les acquisitions tardives en français écrit. Le corpus est constitué d'extraits d'écrits d'étudiants comportant un dysfonctionnement linguistique (erreur ou maladresse orthographique, lexicale, syntaxique, etc.) identifié comme tel par les enseignants qui les ont déposés en ligne. Ces extraits sont accompagnés de commentaires (émanant de l'enseignant-déposant et de linguistes et didacticiens de notre équipe) visant à décrire ce dysfonctionnement. L'objectif didactique de cette base de données (comportant en l'état 2550 extraits, issus d'écrits d'étudiants de tous niveaux et de toutes disciplines) est de fournir aux enseignants ou formateurs désireux de faire progresser leurs étudiants à l'écrit des supports pédagogiques, sous la forme d'exempliers comportant un ou plusieurs types de dysfonctionnement. Parallèlement, dans une perspective variationniste, l'analyse des commentaires descriptifs du dysfonctionnement est exploitée pour questionner les normes à l'œuvre dans l'activité des enseignants-correcteurs. Un exemple de traitement des données récoltées, portant sur les pronoms relatifs, est proposé en fin d'article.

Abstract. Late Acquisitions in Advanced Written French: A Database of Student's Errors and Inappropriate Uses. As part of advanced writing (or literacy), this contribution presents the theoretical and methodological choices at work in the ongoing development of a database on late acquisitions in written French. The corpus is made up of excerpts from student writings with a linguistic inappropriate use (spelling, lexicon, syntax, etc.) identified as such by the teachers who submitted them online. These excerpts are accompanied by comments (from the submitting teacher and from linguists and teachers in our team) describing this inappropriate use. The pedagogical objective of this database (which contains 2550 extracts from the writings of students at all levels and in all disciplines) is to help teachers or trainers who want to improve their students' writing skills by providing examples of one or more types of inappropriate use. At the same time, from a variationism perspective, the analysis of descriptive comments on inappropriate use questions the standards at work in the activity of teacher-correctors. An example of the processing of the data collected, relating to relative pronouns, is proposed at the end of the article.

[A] Corresponding author : francoise.boch@univ-grenoble-alpes.fr

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

Introduction

Inscrite dans le projet national *écri+*[i] visant à améliorer les compétences rédactionnelles des étudiants via la production et la mutualisation d'outils de formation et d'évaluation, notre contribution expose et discute les choix théoriques et méthodologiques présidant à l'élaboration d'une base de données. Notre collecte a pour but de recueillir et de classer des extraits de productions écrites d'étudiants identifiés par des enseignants comme comportant un ou plusieurs dysfonctionnements.

Les objectifs de ce vaste recueil d'énoncés (intitulé dorénavant *corpus écri+*) considérés – de manière consensuelle ou plus flottante – comme erronés ou maladroits sont doubles :
(1) dans une perspective didactique, identifier les objets linguistiques qui semblent les plus résistants en termes d'acquisition de l'écrit à un niveau avancé et productifs en termes de leviers de développement des compétences rédactionnelles d'un public étudiant ;
(2) dans une perspective variationniste, mieux cerner et questionner les normes qui régissent notre regard d'enseignant-correcteur, en prenant en compte différentes descriptions de ces dysfonctionnements : en premier lieu, celles (parfois lapidaires, parfois plus étoffées) indiquées par les enseignants-déposants, et dans un second temps, celles proposées par des ouvrages de grammaire et/ou des travaux de linguistique.

Dans les deux premières parties, l'article pose les jalons théoriques (littéracie avancée ; erreur vs maladresse) qui sous-tendent notre travail. La partie 3 détaille les choix méthodologiques à l'œuvre dans le recueil et le traitement de nos données. À titre illustratif, nous présenterons enfin (partie 4) un exemple de traitement de ces données par l'entrée linguistique des pronoms relatifs.

1 Littératie avancée

La littératie s'est imposée depuis quelques années comme cadre pour penser les pratiques de l'écrit dans leur diversité et les compétences qui leur sont associées (Fraenkel & Mbodj, 2010). Elle permet d'envisager le lire et l'écrire comme des acquisitions tout au long de la vie, et, dans une perspective socio-culturelle et cognitive, comme condition d'accès aux sociétés de la connaissance : la visée n'est pas seulement le savoir lire et écrire, mais le fait de penser et d'agir à travers l'écrit.

Nous adoptons le terme de « littératie avancée » sur le modèle de ce qui se nomme « advanced writing » en contexte anglo-saxon, pour désigner un niveau avancé de compétences rédactionnelles. Le public visé est un public adulte et diplômé, en particulier le public étudiant, et le modèle de référence est celui de l'expertise rédactionnelle, étudiée par la rédactologie, ou science de la rédaction experte (Beaudet & Rey, 2015).

La littéracie avancée a partie liée avec le champ des littératies universitaires, qui trouve son origine dans un double constat, celui de difficultés à l'écrit chez le public étudiant et de l'importance capitale du lire-écrire à l'université. Le champ des littératies universitaires a pour objet « la description des pratiques et des genres de l'écrit en contexte universitaire » car « les apprentissages de l'écrit (en réception et production) ne se limitent pas aux premiers apprentissages, ni aux apprentissages fonctionnels mais se déroulent dans un continuum [...] depuis les premiers contacts avec l'écrit avant les apprentissages scolaires [...] jusqu'aux usages épistémiques de l'écrit pour non seulement diffuser, mais transformer l'expérience ou les connaissances » (Delcambre & Lahanier-Reuter, 2010 : 9).

Dans le cadre de la littératie avancée, la question se pose de la formation universitaire à l'écriture, au-delà de la formation à l'écriture universitaire. Les écrits des étudiants représentent un lieu d'observation privilégié mais les besoins de formation concernent également des professionnels très diplômés comme les cadres par exemple : les difficultés observées appellent une attention générale à l'écrit.

On s'intéresse au développement de compétences rédactionnelles tout au long du cursus scolaire, de l'enseignement primaire à l'enseignement supérieur, mais aussi aux écritures professionnelles et expertes. Par rapport au champ de l'acquisition du langage, de nouveaux défis apparaissent : d'abord, parce qu'on a coutume de qualifier d'acquisitions « tardives » celles qui se jouent vers 11 ou 12 ans ; d'autre part, parce que les recherches sur l'acquisition en langue seconde ou étrangère appellent à s'orienter vers « une approche basée non pas sur tout ce qui oppose les différents locuteurs des langues (qu'ils soient natifs, pseudo-natifs, quasi-natifs, néo-natifs, natifs tardifs, etc.) mais sur ce qu'ils ont en commun » (De Cock & Tyne, 2014 : 157). Le fait de viser un public diplômé dans la langue cible présente cet intérêt. L'enjeu premier est d'identifier les acquis et ce qui résiste à un niveau avancé de maitrise de l'écrit, éventuellement pour faire apparaitre ensuite ce qui est commun à tous et spécifique à tel ou tel public. Les écrits recueillis et analysés peuvent alors également servir au développement de ressources pour la formation à l'écrit.

2 Erreurs et maladresses dans les écrits d'étudiants

Le constat de difficultés rédactionnelles à un niveau avancé est aujourd'hui un acquis et les offres de formation prolifèrent. Cependant, il importe de s'interroger sur les exigences mises en avant et les recommandations qui leur sont associées. Les discours de formation à l'écrit des étudiants et des professionnels ont fait l'objet de plusieurs analyses qui pointent des réductions et des projections diverses (Rinck & Sitri, 2012), en particulier :
(1) la tendance à donner une importance considérable à la composante orthographique des écrits, voire à s'y cantonner ;
(2) les références vagues à ce qui est désigné comme le « respect des règles de grammaire », face aux « tournures fautives », « erreurs fréquentes à éviter » et autres « subtilités de la langue » ;
(3) l'impératif de la phrase simple, auquel s'associent des considérations d'ordre stylistique sur la clarté : on en appelle, par exemple, à « des phrases courtes, rythmées, immédiatement compréhensibles par le lecteur »[ii]. Les critères de « qualité » rédactionnelle mis en avant perpétuent une vision normative et se révèlent peu fondés voire contre-productifs en regard du fonctionnement de la langue et des mécanismes entrant dans la compréhension des textes (Beaudet, 2001 ; Béguelin, 2000). Nous montrerons plus loin l'intérêt de passer d'une approche normative à une approche fonctionnelle pour clarifier les attentes et mieux guider les scripteurs.

Notre objectif est de cerner les difficultés effectives du public étudiant et de questionner les normes enseignées et à enseigner. L'hypothèse qui fonde notre démarche est que la linguistique (au sens large, en incluant les apports des analyses de textes et de discours) peut être mise à profit pour identifier les besoins des étudiants. La description des usages est essentielle mais se heurte à des obstacles.

D'abord, la maitrise du français écrit n'opère que dans des genres. L'enjeu est alors d'analyser les difficultés dans des écrits et des disciplines variés, pour mettre en évidence des besoins transversaux, tout autant que la manière dont ils se spécifient en fonction des genres et des contextes. Cet objectif se démarque des discours sur la maitrise du français écrit en général[iii].

Par ailleurs, l'enseignant confronté aux textes de ses élèves y voit des difficultés, mais l'intuition linguistique ne saurait tenir lieu d'analyse de ces écrits non normés. C'est ici la posture du linguiste-didacticien qui est en jeu et la réflexion sur ces questions se rattache aux travaux déjà anciens mais fondateurs de [Reichler]-Béguelin, dans la lignée de la *Grammaire des fautes* de Frei (1929). En premier lieu, les difficultés observables n'ont pas toutes le statut d'erreurs mais sont diversement désignées comme maladresses, anomalies, dysfonctionnements, malformations, etc. Dans le cadre des recherches sur l'acquisition en

L2, c'est notamment la question du statut de l'erreur qui a conduit à renoncer à « l'analyse d'erreurs » (Perdue, 1980) au profit d'un nouveau paradigme, celui des « learner corpora » ou corpus d'interlangues (Granger, 2008). Cependant, les typologies et les annotations développées dans ce champ très productif semblent y aller de soi : elles se justifient par la comparaison avec la langue cible mais au prix peut-être d'un focus sur certaines caractéristiques au détriment d'autres, comme la fluence ou les collocations (De Cock & Tyne, 2015).

Or, trois problèmes se posent, que ce soit pour l'annotation de corpus ou pour la description d'extraits de textes dans une base de données :

(1) Le passage du global au local : face à l'effet global de maladresse dans tel ou tel passage, il faut diagnostiquer précisément ce qui est en jeu en localisant le problème dans la chaine de caractères, pour l'annotation. Le seul fait de collecter des extraits demande à se questionner sur leur délimitation. La difficulté peut également se formuler en termes de marques de surface et de structure profonde, en particulier pour ce qui a trait à la construction des phrases et à l'organisation thématique.

(2) Le passage de l'intuition à la catégorisation : à partir des commentaires comme « lourd », « mal dit », « problème de construction » ou « syntaxe à revoir », quels référents linguistiques identifier ? Quelle théorie linguistique mobiliser ? Les écrits non normés interpellent la grammaire de texte (Péry-Woodley, 1993 ; Charolles, 2005), de phrase (Chanfrault-Duchet, 2001 ; Masseron, 2003), la phraséologie et les grammaires de construction (Taous 2018). Il faut donc envisager des annotations à même d'évoluer et de trouver des éclairages complémentaires en fonction des cadres de référence. Ainsi, les annotations peuvent servir à « rassembler des occurrences d'un phénomène au contour encore flou pour permettre dans un second temps une analyse plus fine » comme le notent Garcia-Debanc et al. (2017 : 8) qui travaillent sur l'annotation discursive de textes d'élèves. En retour, comme le précisent les auteurs, l'annotation de textes non normés pose question aux modèles linguistiques et peut contribuer à les faire évoluer.

(3) Le passage d'une objectivité illusoire à une intersubjectivité assumée : l'orthographe permet de trancher entre le juste et le faux, mais les linguistes et les grammairiens savent que le verdict de l'acceptabilité est problématique y compris face à des énoncés fabriqués. La tendance actuelle est de partir de ce que les enseignants jugent spontanément comme maladroit : les données associent aux écrits des élèves les biffures et commentaires des lecteurs-correcteurs de ces écrits (Doquet & al., 2017). Cependant, il faut tenir compte du fait que les jugements varient d'un correcteur à un autre. Objectiver les maladresses – autrement dit ce qui pose un problème d'acceptabilité, parfois chez certains mais pas chez d'autres – appelle alors à adopter une approche variationniste (Gadet, 2007). D'un point de vue méthodologique, les options possibles sont de recourir à des tests d'acceptabilité, et, en corpus, à l'annotation multiple et à des procédures d'évaluation de l'accord inter-annotateurs.

Outre le problème de l'évaluation des variantes, c'est aussi celui des potentialités du système et du changement diachronique qui est en jeu. Pour illustrer les difficultés d'analyse, on peut évoquer d'abord la fluctuation des normes. Quelques exemples en sont emblématiques : le rejet par certains enseignants de « et » ou « mais » en tête de phrase, comme des figures d'ajout après le point (un point avant des phrases graphiques sans verbe conjugué, sous forme de participiales, ou de relatives en « qui » et « dont » (Combettes, 2007)). D'une part, l'acceptabilité de ces usages varie selon les lecteurs-correcteurs, et, d'autre part, ce sont des usages attestés dans des écrits journalistiques en particulier, donc conformes aux normes en usage, mais qui appellent à tenir compte de l'évolution des contours de la phrase et de sa définition canonique dans la tradition scolaire. De fait, dans d'autres cas, il semble que le lecteur-correcteur des écrits des étudiants incarne une norme académique qui restreint les possibilités laissées ouvertes par les recommandations relatives au bon usage, comme dans l'extrait 1 tiré de notre corpus *écri+* :

(1) j'ai pu participer au Salon du livre de Barr en Alsace, où trois auteurs et illustrateurs étaient présents sur le stand, à savoir Madame Florence JENNER-METZ, Monsieur Yannick LEFRANÇOIS et Monsieur Alexandre ROANE, qui ont tous trois réalisé des kamishibaï pour les Éditions Callicéphale **comme** *À l'heure du déjeuner* et *Le Fil* (Florence JENNER-METZ), *Le lapin de printemps* (Yannick LEFRANÇOIS) ou encore *Le cadeau de Caro et Cocorico !* (Alexandre ROANE). (M2 Métiers de l'Édition, extrait 1781)

Ici, la correction proposée par l'enseignant-correcteur est de remplacer « comme » par deux-points. Elle pointe un problème dans le texte donné à lire (trop long ? peu lisible ?), et actualise une alternative, comme pour inviter l'étudiant à mobiliser des ressources linguistiques plus variées. Les difficultés des étudiants ne sont donc pas seulement dans ce qui est écrit et que biffe l'enseignant, mais dans un déficit par rapport à des possibilités de la langue qu'il juge sous-exploitées.

En somme, le fait de prendre comme point de départ les jugements des enseignants permet d'identifier les attentes académiques à l'égard des étudiants, quitte à les questionner avec les étudiants eux-mêmes ainsi que dans le cadre de la formation des formateurs. Les données ne sont donc pas seulement les écrits des étudiants mais la lecture qui en est faite. Une approche variationniste est nécessaire pour spécifier des degrés de consensus et de flottement quant à l'acceptabilité des énoncés, en fonction aussi des genres et des disciplines concernées. Recenser l'ensemble de ce qui peut être épinglé comme erreur ou maladresse représente une première étape pour examiner ensuite les occurrences d'un point de vue linguistique, au besoin à travers plusieurs plans d'analyse et modèles théoriques. Enfin, la description linguistique permet d'étayer le jugement sur ce qui pose problème et de guider les possibilités de correction ou de reformulation. D'un point de vue didactique, la collecte entreprise dans notre projet permet déjà de structurer les interventions, par exemple autour de la distinction entre le caractère arbitraire et fonctionnel des attentes en matière de qualité rédactionnelle : à titre d'exemple, l'interrogative indirecte avec point d'interrogation mobilise une simple correction typographique. En revanche, l'usage massif de l'anaphore floue et englobante *cela* par les étudiants (là où les rédacteurs experts utiliseraient plus volontiers un substitut lexical, cf. Boch & Rinck, 2015) appelle un travail de réécriture en profondeur jouant sur la conceptualisation qui s'opère dans le texte. L'approche normative et l'approche fonctionnelle qu'adoptent les lecteurs-correcteurs seront illustrées de manière détaillée à travers l'analyse des pronoms relatifs (partie 4).

3 Méthodologie

Dans cette section, nous présenterons la collecte réalisée dans le cadre de notre projet, ses enjeux ainsi que les choix méthodologiques mis en œuvre à cette fin. Nous expliquerons ensuite comment, à partir de cette collecte, nous avons modélisé la typologie d'erreurs et de maladresses qui en est issue[iv].

3.1 Une collecte des dysfonctionnements dans les écrits d'étudiants

3.1.1 Pourquoi une collecte ?

Dans le contexte actuel de recherche sur les écrits académiques, de nombreuses études se fondent sur la constitution de corpus textuels numériques et leur exploitation à des fins didactiques[v]. En effet, les corpus en linguistique ont marqué un tournant décisif en didactique des langues en permettant leur usage didactique comme lieu privilégié d'observation des usages effectifs (en contexte français, cf. Cnesco, 2018 ; Elalouf & Boré, 2007). Ces données linguistiques « assemblées dans l'objectif de faire une analyse de leur

contenu langagier à l'aide d'outils informatiques » (Cavalla & Hartwell, 2018) permettent d'élaborer aussi bien des supports à l'observation dans la classe de langue (cf. par exemple Kübler & Hamilton, 2018 pour l'anglais de spécialité ; Yan, Tutin & Tran, 2018 pour le français langue étrangère) que des ressources destinées aux étudiants et aux formateurs (cf. par exemple, Tran & Falaise, 2018). De plus, l'authenticité de ces données qui permettent d'accéder à un état des usages langagiers en phase avec celui des apprenants rend leur utilisation pertinente dans le cadre pédagogique (Boulton & Tyne, 2014).

Notre collecte respecte ce principe de recueil de données authentiques puisqu'il prend sa source dans des écrits d'étudiants. Mais, à la différence de la démarche de constitution de corpus de textes intégraux, notre protocole demande à des enseignants volontaires de l'enseignement supérieur francophone de saisir des extraits comportant un dysfonctionnement linguistique. Ces extraits d'au moins une phrase sont déposés sur un formulaire en ligne, pourvus de métadonnées et accompagnés d'une identification du dysfonctionnement repéré par l'attribution d'une étiquette prédéfinie et éventuellement par un commentaire libre[vi].

Ce choix méthodologique est fondé sur une double perspective plutôt qualitative que quantitative. En effet, notre collecte permet de récolter deux types de données : d'une part, des données linguistiques pour continuer à documenter les usages « déviants » qui ont désormais toute leur place dans la description linguistique ([Reichler]-Beguelin 1994) ; d'autre part, et c'est là toute son originalité, des données relatives aux attentes des enseignants et à leur seuil d'acceptabilité quand ils rencontrent un dysfonctionnement linguistique. Dans le premier cas, la collecte est un outil complémentaire aux corpus déjà existants car elle permet de faire ressortir les difficultés saillantes. Un va-et-vient avec les corpus permet ensuite de quantifier l'ampleur des phénomènes identifiés. Dans le second cas, notre objectif est de décrypter les attentes souvent implicites des enseignants afin de mieux les outiller pour travailler ces points de blocage avec leurs étudiants. D'un point de vue scientifique, ce travail permettra de questionner les normes en adoptant le point de vue de l'enseignant-correcteur et en s'appuyant sur cette première lecture pour aller ensuite vers des analyses linguistiques portant sur la variabilité des jugements d'acceptabilité. Enfin, comme la collecte s'étend sur une dizaine d'années dans le cadre du projet *écri+,* nous envisageons également d'observer les évolutions de ces attentes.

3.1.2 Format de la collecte

Le formulaire de recueil de données, créé sous *Sphinx*[vii], a été construit pour être accessible à un public large et varié d'enseignants de l'enseignement supérieur. Dans cette optique, nous présentons succinctement les objectifs et modalités de la collecte sur la page d'accueil (cf. figure 1).



Fig. 1. Présentation des objectifs et des modalités de la collecte.

Nous avons diffusé le lien vers ce formulaire au sein de nos réseaux et auprès de tous les partenaires du projet *écri+*. La première campagne de récolte a été lancée début 2018. Nous diffusons régulièrement depuis des appels à collecte. Afin d'optimiser nos chances de recueillir des données, nous avons organisé la collecte sur une seule page en quatre rubriques. Cette architecture nous permet d'obtenir un extrait contenant un ou plusieurs dysfonctionnements repérés par l'enseignant-correcteur (rubrique 1), éventuellement commentés par ce dernier (rubriques 2 et 3) et pourvu de ses métadonnées (rubrique 4). Si les rubriques 2 et 3 ne sont pas renseignées à ce stade, le travail de repérage et de commentaires est effectué ultérieurement par notre équipe de linguistes.

Après avoir recopié le texte de l'extrait dans la rubrique 1, l'enseignant est invité à identifier le dysfonctionnement repéré. Cette identification est présentée de manière optionnelle (cf. figure 2).

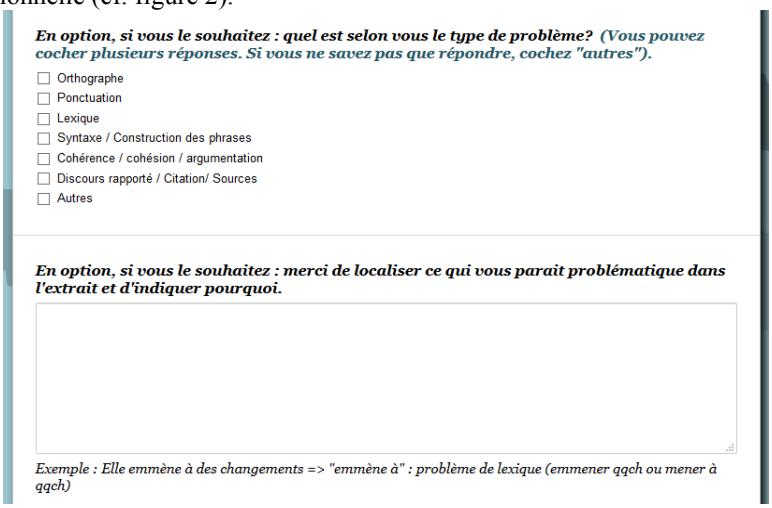


Fig. 2. Intitulés des rubriques d'identification du dysfonctionnement.

Comme nous adressons ce formulaire à des enseignants de tous les champs disciplinaires, nous avons dû tenir compte de leur absence de familiarité avec la terminologie utilisée par les linguistes. C'est pourquoi nous leur proposons d'identifier le problème qu'ils ont repéré à l'aide d'une étiquette prédéfinie ou d'un commentaire libre, les deux options pouvant se cumuler. Les 7 étiquettes prédéfinies ou macro-catégories que nous avons choisies (Orthographe ; Ponctuation ; Lexique ; Syntaxe / Construction des phrases ; Cohérence / cohésion / argumentation ; Discours rapporté / Citation / Sources ; Autres) nous permettent, d'une part, de couvrir un ensemble large de composantes de l'écrit et, d'autre part, d'être compréhensibles par l'enseignant non spécialiste comme par le linguiste. Nous guidons l'enseignant par un exemple qui illustre la manière dont celui-ci peut formuler ce qui lui parait problématique et l'autorise explicitement à formuler de manière synthétique son commentaire (voir figure 2). Dans ces rubriques, les termes *problème* et *problématique* sont utilisés dans leur acception la plus courante pour ne pas entraver le processus de collecte tout en nous permettant de recueillir des données sur les attentes des enseignants.

Dans la dernière rubrique (rubrique 4), l'enseignant est invité à fournir, de manière libre, des indications sur l'extrait saisi. Celles-ci concernent le niveau d'étude (Licence1, Master2, etc.), le contexte dans lequel a eu lieu l'enseignement qui a conduit à l'évaluation d'où provient l'extrait (IUT, Info-Comm, Master MEEF-PE etc.) et le genre textuel auquel appartient celui-ci (dissertation, mémoire, devoir d'analyse type CRPE etc.). Ces informations sont affectées comme métadonnées à chaque extrait. Elles nous permettront par la suite de procéder à d'autres analyses, notamment en reliant certaines attentes des enseignants à un sous-genre académique et à une filière disciplinaire spécifiques.

3.1.3 État de la collecte

La présente étude s'appuie sur l'état de la collecte au 14 décembre 2019. Celle-ci se compose de 2550 extraits. Le tableau 1 ci-dessous présente les résultats bruts par macro-catégorie.

Tableau 1. Le nombre d'extraits recueillis classés par macro-catégorie.

Macro-catégorie	**Nombre d'extraits**	
Orthographe	1338	52%
Ponctuation	768	30%
Lexique	629	25%
Syntaxe / Construction des phrases	1316	51%
Cohérence / cohésion / argumentation	405	16%
Discours rapporté / citation / sources	177	7%

Parmi les 2550 extraits recueillis, plus de la moitié d'entre eux sont affectés aux macro-catégories *Orthographe* et *Syntaxe / Construction des phrases*. Les principaux autres dysfonctionnements relevés concernent les macro-catégories *Ponctuation* (30%) et *Lexique*

(25%). Les phénomènes étiquetés *Cohérence / cohésion / argumentation* (16%) et *Discours rapporté / citation / sources* (7%) sont plus marginaux dans ce premier état de la collecte. La catégorie *Autres* n'apparait pas dans ce tableau car les dysfonctionnements relevant de cette catégorie ont été post-traités par notre équipe soit par suppression (comme par exemple, les cas d'oubli de mots dans une phrase) soit par affectation à une autre macro-catégorie.

Grâce aux informations fournies par les enseignants dans la rubrique 4, nous pouvons caractériser plus finement cette première collecte. Le graphique ci-dessous (Fig.3) récapitule les niveaux des étudiants ayant produit ces extraits.

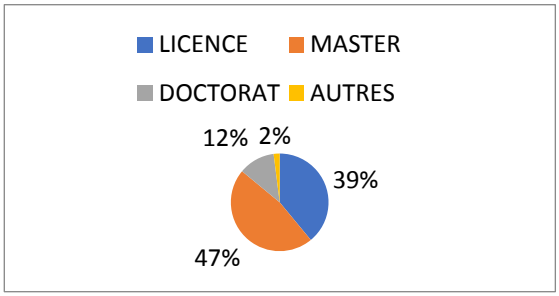


Fig. 3. Répartition des extraits par niveau de formation.

La collecte des métadonnées concernant le niveau est très satisfaisante car tous les extraits sont pourvus de cette information. Dans la Figure 3, la rubrique *Autres* regroupe les filières BTS, IUT et écoles d'ingénieur. De même, plus de 81% des extraits sont étiquetés en fonction d'une filière disciplinaire (Histoire, Sciences du Langage, Lettres, Droit, Philosophie pour les Licences ; MEEF, Didactique des langues, Linguistique, Métiers de l'édition pour les Masters). En revanche, nous avons rencontré une plus grande difficulté pour obtenir des données concernant le genre textuel qui sont le plus souvent lacunaires (mémoire, rapport de stage, résumé, synthèse, examen écrit).

3.2 Une typologie des dysfonctionnements dans les écrits d'étudiants

3.2.1 Pourquoi une typologie ?

Nous avons élaboré cette typologie pour modéliser les données brutes dans l'optique de leur traitement automatique. En effet, notre objectif à moyen terme est de créer une interface pour interroger la base de données ainsi constituée. Il s'agit donc d'une typologie fondée sur une description linguistique des erreurs et maladresses relevées comme telles par des enseignants de l'enseignement supérieur. Comme la collecte court sur une dizaine d'années, celle-ci a pour vocation à s'affiner progressivement en fonction des extraits recueillis.

Cet outil est destiné à la fois au public des enseignants de l'enseignement supérieur mais aussi aux chercheurs en linguistique, en didactique du français et en sociolinguistique. Dans le premier cas, nous prévoyons la possibilité d'interroger la base de données en sélectionnant une macro-catégorie. À partir de cette extraction, l'enseignant pourra confectionner un exemplier avec des extraits qu'il aura choisis afin de conduire son travail en classe à partir de dysfonctionnements attestés dans des productions authentiques. La réflexion est en cours sur la manière d'organiser l'interrogation de l'interface concernant le second volet de notre projet, à savoir les attentes des enseignants selon les disciplines, les genres textuels et les filières. À travers cette tâche de typologisation, nous cherchons à interroger les normes mobilisées par les enseignants à la lumière des analyses linguistiques.

3.2.2 Présentation des macro-catégories

Nous présentons ici chacune des macro-catégories retenues pour le classement de nos extraits en les illustrant avec des commentaires d'enseignants-correcteurs placés entre guillemets. Certains extraits relèvent de l'erreur et d'autres plutôt de la maladresse.

Dans la macro-catégorie *Orthographe*, nous avons regroupé les dysfonctionnements concernant l'orthographe lexicale et l'orthographe grammaticale comme les règles d'accord (« Accord du participe passé avec auxiliaire *avoir* », L2 Sciences du Langage, extrait 2099) et la conjugaison (« Terminaison erronée P3 passé simple L1-2 Histoire, extrait 693).

Dans la macro-catégorie *Ponctuation*, sont répertoriés les extraits présentant des signes de ponctuation dysfonctionnants, qu'ils soient mal placés (« virgule erronée entre le sujet et le verbe », Formation doctorale, extrait 1451) ou absents (« virgule absente après l'incise, M1 Métiers de l'Édition, extrait 1173) mais aussi des erreurs typographiques (par ex., espace manquant, usage abusif de l'italique etc.).

La macro-catégorie *Lexique* recense les dysfonctionnements relatifs aux registres de langue (« trop populaire », L3 Lettres, extrait 693), aux idiomatismes divers tels l'emploi erroné de prépositions après un verbe ou les erreurs de collocation (« Un objet est "simple de réalisation". Une démarche est "simple à réaliser". », L1 Sciences du langage, extrait 156).

Dans la macro-catégorie *Syntaxe / Construction des phrases* ont été classés les dysfonctionnements structurels concernant les locutions prépositionnelles (« "tant qu'à" au lieu de "tandis que" », 3^{e} année école d'ingénieurs, extrait 259), la gestion des anaphores (« Pb avec l'anaphorique "elle" qui ne reprend aucun autre mot précédent », L2 Lettres, extrait 38) et de la co-référence, les pronoms relatifs, l'emploi des temps verbaux (« Emploi erroné du futur simple dans un récit au passé », L1-L2 Histoire, extrait 815), les ruptures de constructions dont les phrases averbales (« Phrase averbale au participe présent "son rôle étant" => virgule plutôt que point avant la proposition, ou "son rôle est de..." », Master MEEF-PE, extrait 211).

La macro-catégorie *Cohérence / cohésion / argumentation* recueille les erreurs et maladresses concernant l'emploi des connecteurs (« Le connecteur "en effet" relie la seconde phrase non pas à la première phrase, mais à une simple proposition relative enchâssée dans l'un des arguments de cette première phrase. », L1 Droit, extrait 157), l'interprétation erronée d'un fait posé, des confusions de concepts, des incohérences qui rendent l'énoncé incompréhensible (« phrase obscure », M1 MEEF, extrait 943) ou encore l'emploi de l'anaphorique neutre (« *cela* à reformuler (lexicalement) pour plus de clarté. », M1 Didactique des langues, extrait 144).

Enfin, dans la macro-catégorie *Discours rapporté / Citation / Sources*, sont classés les dysfonctionnements concernant les conventions ou les normes de citations (« Citation : conventions de citation APA non respectées (initiale du prénom superflue) », M2 mémoire, extrait 1088), l'absence de mention de la source mais aussi l'introduction de la source (« Verbe "exprime" peu adapté », M2 mémoire, extrait 168).

Lors du traitement des données, nous avons constaté l'existence de désaccords entre les différents annotateurs sur l'acceptabilité d'un phénomène linguistique (par exemple, l'absence d'une virgule après un circonstant initial est considérée par certains comme un dysfonctionnement). Notre collecte a donc pour objectif de recenser des extraits et les jugements portés sur ceux-ci. Par ailleurs, certains phénomènes identiques sont parfois affectés à plusieurs macro-catégories. En effet, dans l'exemple 2, l'extrait est affecté aux deux macro-catégories *Syntaxe / construction de phrase* et *Ponctuation*.

(2) Globalement, 14 élèves ont donné plus de mots appartenant au vocabulaire du genre policier après avoir réalisé la séquence d'apprentissage. **Ce qui représente plus de la moitié des élèves**. (Master MEEF-PE, extrait 1121).

Le commentaire de l'enseignant « Phrase sans proposition principale (découpage erroné en deux phrases) » mentionne le phénomène syntaxique de la phrase averbale et le point final comme ponctuant erroné. Nous avons décidé de conserver le pluri-étiquetage pour ces extraits.

3.2.3 Vers une caractérisation plurielle des extraits

Si ces macro-catégories permettent une recherche par grand type de dysfonctionnements, elles demeurent très larges. Or, il faut prévoir la possibilité pour l'utilisateur de mieux cibler les extraits qu'il veut réunir, à des fins d'analyse ou de constitution d'un corpus d'observation pour ses apprenants. C'est pourquoi nous proposons, dans la base de données, une caractérisation plurielle des dysfonctionnements : à chaque extrait sont associés une macro-catégorie (deux, le cas échéant) et un descriptif plus précis du dysfonctionnement observé. L'intérêt d'une base de données est de dépasser les limites de l'annotation d'erreurs et de maladresses en corpus : même avec un jeu d'étiquettes très détaillé, basé sur des catégories et des sous-catégories, on se heurte au problème de la catégorisation des dysfonctionnements. Pour obtenir une annotation fine, on risque de multiplier les catégories et de complexifier leur hiérarchisation, alors qu'on ne connait pas les usages potentiels du corpus et qu'il faut laisser ouvertes des possibilités d'exploitation multiples. Les enrichissements apportés aux corpus, couteux en temps, se révèlent parfois peu utilisés, car trop rigides.

Une base de données permet à la fois de fournir une description des extraits et de s'interroger sur cette description. En effet, un même extrait peut donner lieu à plusieurs descriptions, comme une première description produite par un enseignant, une description produite par un linguiste, puis par un autre. On peut également envisager de soumettre les descriptions produites par les uns au regard des autres, de sorte qu'une description pourra alors être accompagnée d'un commentaire sur la description. On se situe ainsi dans la perspective d'un travail collaboratif de caractérisation des dysfonctionnements. L'idée est de permettre une description évolutive et dynamique, dans le sens où elle peut intégrer des précisions sur le caractère non consensuel du jugement d'erreur ou de maladresse et des vues multiples sur un même phénomène.

4 Illustration à travers un exemple : les pronoms relatifs

Cette partie fournit un exemple de traitement d'un objet linguistique considéré généralement par les didacticiens comme « résistant » à l'écrit, même chez les publics avancés francophones (Beaulieu-Handfield, 2018) : le pronom relatif. Selon Laparra (1995), les nombreuses difficultés que son emploi suscite sont liées à des causes diverses ; nous en retenons une ici, susceptible d'éclairer l'analyse de certaines formes déviantes présentes dans le corpus *écri+* : l'utilisation du pronom nécessite de gérer simultanément plusieurs opérations, en particulier le fait « de sélectionner le pronom avant même d'avoir produit le verbe dont il dépend » (*ibid.* : 69).

En croisant la requête « pronom », « relatif » et « pronom relatif », nous avons réuni 43 séquences que nous avons ensuite classées[viii] selon le type de dysfonctionnement indiqué. Nous traitons ici le cas du choix erroné du pronom relatif[ix]. Un examen plus attentif des extraits nous permet d'affiner ce premier diagnostic : la mauvaise sélection du pronom relatif s'opère dans des considérations diverses qu'il nous semble productif, au plan didactique, de distinguer.

Dans les extraits (3) à (5), l'erreur est liée au non-respect de la construction prépositionnelle du verbe de la relative :

(3) D'une part, il existe des facteurs naturels, incontrôlables **dont** le producteur n'a pas prise. (*sur lesquels* attendu) (M1 MEEF, extrait 931)
(4) Pour les élèves en difficultés en lecture, j'ai observé des progrès **dont** je ne m'attendais pas. (*auxquels* attendu) (M2 MEEF, extrait 1199)
(5) Cette expérience m'a permis d'en apprendre plus sur le monde de l'édition et de découvrir des domaines **pour lesquels** je ne m'intéressais que peu, puisque je pensais ne pas être capable d'y réaliser autant de choses différentes. (*auxquels* attendu) (M1 Métiers de l'édition, extrait 1802)

L'erreur de sélection se manifeste ici avec des verbes dont la construction est à priori connue des étudiants qui les utilisent *(avoir prise sur, s'attendre à, s'intéresser à)*. Toutefois, les relatifs en question relèvent de cas dits obliques, dont la fréquence est faible à l'oral et qui sont l'objet de multiples erreurs (cf. par ex. Laparra, 1995). Ainsi, ces formes peu habituelles à l'oreille des étudiants doivent être produites via un raisonnement grammatical, loin de leur usage spontané de la langue. Les mettre en présence de ces formes erronées et leur demander de les corriger peut constituer une stratégie efficace pour renforcer leur conscience du lien indéfectible qui unit la préposition constitutive du verbe de la proposition relative et la forme du pronom relatif qui lui correspond.

Dans l'extrait (6), si le pronom est erroné, la cause en revient davantage à un problème de collocation verbale.

(6) Nous ne savons pas la cause de son jugement **auxquels il ne se présente pas**. (L1-2 Histoire, extrait 828)

En effet, la collocation *se présenter à un jugement* fonctionne mal au plan sémantique. Avec une autre unité lexicale telle que *procès* par exemple, l'énoncé devient correct : *Nous ne savons pas la cause de son procès auquel il ne se présente pas* (nous corrigeons aussi l'erreur orthographique sur *auxquels*, fréquente pour les relatifs composés dans notre corpus).

Le cas d'erreur est plus classique en (7) car il reflète une tendance forte d'usage oral de la construction verbale *avoir besoin que*, répertoriée par Gadet (1996) comme usage populaire.

(7) Il est aussi possible de retirer le répertoire de mots et de demander aux élèves de chercher les mots **qu'**ils ont besoin seulement dans l'épisode. (*dont* attendu) (Master MEEF, extrait 1197)

On peut supposer que le pouvoir d'attraction de certains usages oraux, qui substituent au pronom relatif *dont* un *que* (Müller, 2006), peut influencer l'écrit des étudiants. Une autre source d'influence potentielle peut être l'existence du patron syntaxique *(QUE* + Sujet + forme verbale conjuguée de *avoir besoin),* correct en français lorsque l'antécédent du relatif n'est plus COI (et donc repris par *dont*) mais COD (et donc repris par *que*). C'est le cas lorsque *avoir besoin* est lui-même suivi d'un COI appelant l'antécédent-COD, par exemple sous la forme de *DE* + verbe transitif à l'infinitif (*Les élèves doivent chercher les mots qu'ils ont besoin de comprendre*) ou encore sous la forme d'une complétive (*Voici les informations que nous avons besoin que vous classiez*), exemple pédagogiquement intéressant au plan de l'analyse (même s'il est susceptible d'être jugé comme stylistiquement assez lourd) en ce qu'il comporte un *que* relatif suivi d'un *que* conjonctif. Ici encore, l'observation guidée d'énoncés nous semble nécessaire pour que les étudiants comprennent en quoi deux patrons syntaxiques en apparence proches imposent la sélection d'un relatif spécifique (*que* ou *dont*).

Dans les exemples qui précèdent, comme pour d'autres dysfonctionnements des pronoms relatifs (ou démonstratifs), ce qui est en jeu est moins l'interprétation de l'énoncé (il est possible d'en reconstruire la cohérence) que ce que Beaulieu-Handfield, s'appuyant sur la théorie de la pertinence de Sperber et Wilson (1989) et sur les travaux de [Reichler-]Beguelin (1994), appelle « le coût de traitement considéré comme trop exigeant par l'interprétant » (Beaulieu-Handfield, 2018 : 36). Tout dysfonctionnement (que celle-ci

nomme indifféremment *anomalie* ou *rupture textuelle*) est défini comme tel si le lecteur juge l'effort d'interprétation excessif par rapport à ce qu'il pourrait être. Cette définition pragmatique du dysfonctionnement, qui permet d'aller au-delà d'une approche se contentant de le biffer, nous semble pédagogiquement productive, en ce qu'elle exige du scripteur de se mettre à la place du lecteur-interprétant. L'observation guidée des extraits qui précèdent peut favoriser chez les étudiants des prises de conscience durables sur l'intérêt de faciliter au maximum la compréhension de son lecteur, y compris à travers leurs choix syntaxiques.

Les extraits (8) à (10) posent la question de la pertinence de leur correction. Si les enseignants qui les ont déposés les ont jugés à priori dysfonctionnant (la sélection du pronom relatif étant toujours en cause), leur analyse – et la correction qu'on peut en faire – nous invite à les considérer comme des cas-limites du point de vue de leur acceptabilité.

(8) Ainsi, j'ai découvert sur le quartier des usages et enjeux (culturels, cultuels, politiques, sociaux, financiers...) **pour lesquels** je n'étais pas préparé (candidature M1 « interventions sociale », extrait 975)
(9) Je demande aux élèves de se situer dans le temps en m'indiquant le mois **dans lequel** nous sommes. (mémoire M2 MEEF, extrait 1233)
(10) Cette hypothèse rejoint l'idée de Jacques David **où** il explique que l'élève se retrouve vite en surcharge cognitive lorsqu'il réfléchit sur la langue. (M1 MEEF, extrait 1278)

Dans l'extrait (8), l'enseignant propose la correction *auxquels*, fondée par l'existence de la construction *être préparé à quelque chose*. Mais le patron syntaxique *être préparé pour quelque chose* semble également recevable, même s'il est moins fréquent dans l'écrit scientifique[x], d'autant que dans l'extrait (8), les deux constructions semblent quasi-synonymes. L'extrait (9) a fait l'objet de la correction suivante : *auquel ??*, le double points d'interrogation témoignant du doute de l'enseignant, et peut-être de sa conscience floue du caractère hypercorrectif de sa proposition (motivée sans doute par le déterminant prépositionnel *au* introduisant l'antécédent *mois* dans la phrase sous-jacente à la structure relative *nous sommes au mois…*). En effet, *Le mois auquel nous sommes* ne fonctionne pas à nos yeux (ou plus mal en tout cas que *le mois dans lequel nous sommes*, sémantiquement plus immédiatement interprétable). Quoi qu'il en soit, le référent *le mois* se prête mal à la reprise par un relatif composé. *Le mois où nous sommes* serait sans doute ici préférable (en exploitant le caractère tout autant spatial que temporel du relatif *où*), si toutefois il fallait apporter une correction à (9).

L'extrait (10) reflète une autre potentialité du pronom relatif *où*, bien décrite en linguistique, qui consiste à renvoyer à un « repérage spatial plus abstrait » (Hadermann, 2009 : 124), ici visible à travers plusieurs indices : le sémantisme abstrait de l'antécédent, le rapport contenant/contenu (une idée peut contenir une explication), la préposition locative sous-jacente introduisant l'antécédent *(il explique dans cette idée que*). La correction en *quand* proposée par l'enseignant nous semble ainsi peu pertinente. Plutôt que la sélection du pronom relatif, c'est au fond le caractère superflu de la relative *où/quand il explique* qui semble en jeu ici. Un lien direct entre l'antécédent *idée* et son contenu sémantique semble plus judicieux (par exemple à travers la reformulation : *Cette hypothèse rejoint l'idée de Jacques David* ***selon laquelle*** *l'élève se retrouve vite en surcharge cognitive […]*).

Conclusion

L'analyse des quelques extraits présentée ici nous permet d'illustrer l'une des richesses du *corpus écri+* : cette collecte autorise, en premier lieu, la constitution d'un exemplier d'énoncés comportant à priori un même type de dysfonctionnement. L'analyse linguistique plus poussée de ces exemples permet, dans un second temps, d'aiguiser notre regard sur ce type de dysfonctionnement, en distinguant mieux ce qui relève strictement de la norme (ce

qui s'écrit ou pas dans tel genre scriptural, du point de vue des ouvrages de référence, des normes disciplinaires, mais aussi, de manière plus floue, des attentes de chacun) d'une approche plus fonctionnelle (ce qui peut être interprété avec plus ou de moins de facilité). Dans le cadre de l'intervention didactique, l'approche fonctionnelle consistant à penser l'écrit en fonction de son lecteur peut être travaillée via la relecture commentée des écrits d'étudiants par leurs pairs. Les travaux de Lejot (2017) montrent que ces ateliers de relectures commentées (dans son cas auprès d'un public de doctorants) favorisent chez les participants une « méta-réflexion commune sur les normes académiques rédactionnelles de leur discipline » (*ibid.*, §36), et des effets réels sur leurs pratiques d'écriture, se traduisant notamment par le fait que les étudiants retravaillent spontanément leur texte en fonction des commentaires des séances précédentes avant de les remettre à leurs pairs pour relecture (*ibid*., §35) : ils deviennent ainsi de meilleurs lecteurs de leurs propres écrits.

On peut faire l'hypothèse que ce type d'ateliers serait à même de développer une attention à l'écrit auprès d'étudiants de tous niveaux. Il serait intéressant d'observer, par exemple, si les étudiants ayant bénéficié d'un tel dispositif mobilisent davantage de métalangage dans leurs commentaires (dans le cas des relatifs, sont-ils capables de nommer le pronom et d'expliquer que la difficulté d'interprétation du pronom est liée à son référent ?) et si l'argumentation qu'ils développent pour justifier des choix linguistiques qu'ils opèrent à l'écrit est plus précise au plan linguistique. Kondo et Takatsuka (2009) ont d'ailleurs montré qu'à l'issue d'un tel travail, les étudiants (ici japonais dans un atelier d'écriture en FLE) ont plus conscience de ce qui est clair et ce qui l'est moins pour le lecteur, et ont tendance à améliorer leurs propres écrits en répondant à des commentaires formulés par leurs pairs.

Ainsi, les enjeux que comporte la mise à disposition de la base de données *écri+,* en construction, sont double. À terme, il s'agira d'une part de montrer les points de résistance d'un public de niveau avancé dans son apprentissage du français écrit. D'autre part, et très concrètement, via les descriptions plurielles des dysfonctionnements identifiés, la base de données fournira des indications potentiellement utiles aux enseignants-correcteurs pour étayer les annotations qu'ils portent sur les écrits de leurs étudiants – en particulier concernant des énoncés corrects mais parfois jugés peu acceptables – guidant ainsi leur travail de réécriture.

Références bibliographiques

Beaudet, C. (éd.) (2001). *Recherches en rédaction professionnelle,* 1(1).

Beaudet, C. & Rey, V. (2015). *Les écritures expertes*. Aix-en-Provence : Presses de l'Université de Provence.

Beaulieu-Handfield, E. (2018). Ruptures de cohérence dans les écrits d'étudiants universitaires. Mémoire de maitrise, UQAM, Montréal. En ligne : https://archipel.uqam.ca/11871/1/M15704.pdf

Béguelin, M.-J. (2000). Diagnostic des erreurs dans un corpus d'écrits techniques. In B. Denis (éd.), *La rédaction technique* (p.105-119). Bruxelles : De Boeck Supérieur.

Boch, F. & Rinck, F. (2015). Anaphores démonstratives dans les écrits d'étudiants de Master. Comparaison avec les pratiques expertes. *Linx*, *72*. doi : 10.4000/linx.1631

Borzeix, A. & Fraenkel, B. (éd.) (2001). *Langage et Travail, Communication, Cognition, Action*. Paris : CNRS.

Boulton, A. & Tyne, H. (2014). *Des documents authentiques aux corpus : démarches pour l'apprentissage des langues*. Paris : Didier.

Chanfrault-Duchet, M. (2001). La phrase au lycée : enjeux didactiques. *Le français aujourd'hui*, *135*(4), 52-63. doi :10.3917/lfa.135.0052.

Charolles, M. (2005). Analyse de discours, grammaire de texte et approche grammaticale des faits de textualité. *Le français aujourd'hui*, *148*(1), 33-45. doi :10.3917/lfa.148.0033.

Cnesco (2018). Écrire et rédiger : comment guider les élèves dans leurs apprentissages. Notes des experts. En ligne : https://www.cnesco.fr/fr/ecrire-et-rediger

Cock de, S. & Tyne, H. (2014). Corpus d'apprenants et acquisition des langues. *Recherches en Didactique des Langues et des Cultures, 11*(1), 137-168.

Combettes, B. (2007). Les ajouts après le point : aspects syntaxiques et textuels. In M. Charolles, N. Fournier, C. Fuchs & F. Lefeuvre (éd.), *Parcours de la phrase : mélanges offerts à Pierre Le Goffic* (p.119-131). Paris : Ophrys.

Doquet, C., Enoiu, V., Fleury, S. & Maziotti, S. (2017). Problèmes posés par la transcription et l'annotation d'écrits d'élèves. *Corpus.* 16. En ligne : http://journals.openedition.org/corpus/2776

Elalouf, M.-L. & Boré, C. (2007). Construction et exploitation de corpus d'écrits scolaires. *Revue française de linguistique appliquée*, *XII* (1), 53-70.

Eshkol-Taravella, I. (2015). *La définition des annotations linguistiques selon les corpus : de l'écrit journalistique à l'oral*. Habilitation à Diriger des Recherches en Linguistique. Université d'Orléans.

Fraenkel, B. & Mbodj, A. (2010). Les New Literacy studies, jalons historiques et perspectives actuelles. *Langage et Société*, *133*, 7-24.

Frei, H. (1929). *Grammaire des fautes*. Paris : P. Geuthner (réédititon en 2007, Ennoïa).

Gadet, F. (1996). Niveaux de langue et variation intrinsèque. *Palimpsestes, 10*, 17-40.

Gadet, F. (2007). *La Variation sociale en français* (2e éd.). Paris : Ophrys.

Garcia-Debanc, C., Ho-Dac, L.-M., Bras, M. & Rebeyrolle, J. (2017). Vers l'annotation discursive de textes d'élèves. *Corpus*, *16*. En ligne : http://journals.openedition.org/corpus/2783

Granger, S. (2008). Learner corpora. In A. Lüdeling & M. Kytö (éd.) *Corpus Linguistics. An International Handbook*. Volume 1 (p.259-275). Berlin & New York: Walter de Gruyter.

Hadermann, P. (2009). Le relatif *où* et ses principaux concurrents : variation morpho-syntaxique et neutralisation entre synchronie et diachronie. *Travaux de linguistique,* 59, 123-146.

Kondo, Y. & Takatsuka, S. (2009). Revision by Electronic Peer Feedback in Japanese College Students' English Writing. *International Journal of Curriculum Development and Practice*, *11*, 1-11.

Lappara, M. (1995). Quelques réflexions didactiques sur l'apprentissage des relatives, *Pratiques, 87*, 59-91.

Lejot, È. (2017). La relecture entre pairs en formation doctorale : de l'analyse des commentaires à l'élaboration d'une grille d'accompagnement. *Lidil*, 55. doi : 10.4000/lidil.4255

Masseron, C. (2003). Le déficit syntaxique dans les copies argumentatives. Hypothèses et propositions de travail. *Le français aujourd'hui, 141*(2), 83-97. doi : 10.3917/lfa.141.0083.

Muller, C. (2006). Sur les propriétés des relatives. *Cahiers de grammaire, 30*, 319-337. En ligne : http://w3.erss.univ-tlse2.fr/publications/CDG/30/CG30-24-Muller.pdf

Oudart, A. C. (2001). *Les chargés de relation clientèle face à la lettre de réclamation, Pratiques, difficultés, Apprentissages*. Paris : Presses Universitaires du Septentrion.

Perdue, C. (1980). L'analyse des erreurs : un bilan pratique. *Langages*, *57*, 87-94.

Pery-Woodley, M.-P. (1993). *Les écrits dans l'apprentissage*. Paris : Hachette.

[Reichler-]Béguelin, M.-J. (1992). L'approche des "anomalies" argumentatives. *Pratiques, 73*, 51-78.

[Reichler-]Béguelin, M.-J. (1994). Encodage du texte écrit : normes et déviances dans les processus référentiels et dans le marquage de la cohésion. In L. Verhoeven & A. Teberosky (éd.), *Understanding early literacy in a developmental and cross-linguistic approach* (p.175-204). Strasbourg : European Science Foundation.

Rinck, F. & Sitri, F. (2012). Pour une formation linguistique aux écrits professionnels. *Pratiques*, *153-154*. doi :10.4000/pratiques.1937.

Sperber, D. & Wilson, D. (1989). *La pertinence : communication et cognition*. Paris : Éditions de Minuit.

Taous, T. (2018). Les zones insoupçonnées de la relecture. Autour des 'formules croisées'. *Le français aujourd'hui, 203*(4), 51-62. doi:10.3917/lfa.203.0051.

[i] Ce projet bénéficie de l'aide de l'ANR PIA3 écri+ (http://ecriplus.fr/).

[ii] Selon http://scriptura-formation.fr/formation-professionnelle/.

[iii] Même l'orthographe se spécifie selon les genres : par exemple, les finales verbales homophones hétérographes augmentent en présence de verbes à l'imparfait, ce que l'on trouve dans les écrits de certains genres et/ou certaines disciplines, mais pas dans tous.

[iv] Ce travail a fait l'objet d'un stage long réalisé par Ivan Romaniuk, étudiant en Licence Sciences du Langage à l'Université Grenoble Alpes.

[v] Voir par exemple le corpus *Littéracie avancée* constitué à partir d'écrits d'étudiants de langue maternelle française. Ce corpus est librement téléchargeable dans l'entrepôt Ortolang https://www.ortolang.fr/market/corpora/litteracieavancee.

[vi] Ce formulaire est accessible à l'adresse suivante https://enquetes.univ-grenoble-alpes.fr/v4/s/tvtvo6 et la collecte se poursuit.

[vii] Pour plus d'information sur ce logiciel mis à notre disposition par l'Université Grenoble Alpes, voir https://www.lesphinx-developpement.fr/.

[viii] Les extraits produits par des étudiants non-natifs ont été écartés de ce sous-corpus, en ce qu'ils présentaient des erreurs caractéristiques qui mériteraient une analyse spécifique.

[ix] D'autres anomalies répertoriées qui mériteraient un examen attentif : présence d'un relatif et d'un pronom complément renvoyant au même référent ; relatif superflu ou manquant ; phrases comportant un grand nombre de relatives.

[x] Dans le corpus Scientext français de 4,8 millions de mots (https://scientext.hypotheses.org/), qui rassemble des écrits scientifiques dans de multiples disciplines, les deux configurations sont rares mais elles co-existent (13 occurrences de « être préparé à » pour 2 occurrences de « être préparé pour »).