Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

Impact des modalités induites par les outils d'annotation manuelle : exemple de la détection des erreurs de français

Anaëlle Baledent
Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France
anaelle.baledent@unicaen.fr

RÉSUMÉ

Certains choix effectués lors de la construction d'une campagne d'annotation peuvent avoir des conséquences sur les annotations produites. En menant une campagne sur la détection des erreurs de français, aux paramètres maîtrisés, nous évaluons notamment l'effet de la fonctionnalité de retour arrière. Au moyen de paires d'énoncés presque identiques, nous mettons en exergue une tendance des annotateurs à tenir compte de l'un pour annoter l'autre.

ABSTRACT

Impact of modalities induced by manual annotation tools : example of French error detection

Some choices made during the construction of an annotation campaign can have consequences on the resulting annotations. By conducting a campaign on French error detection, with controlled parameters, we mainly evaluate the effect of the backtracking feature. Using pairs of nearly identical utterances, we highlight a tendency of annotators to take one into account when annotating the other.

MOTS-CLÉS : annotation manuelle, constitution de corpus, corpus d'erreurs de français, fonctionnalité de retour arrière, ordre de présentation.

KEYWORDS: manual annotation, corpus building, French errors corpus, backtracking feature, order of presentation.

1 Introduction

Pour le Traitement Automatique des Langues (T.A.L.), il est nécessaire d'avoir des corpus de référence, qui ne s'obtiennent communément qu'à partir d'annotations manuelles (Fort, 2016). Pour obtenir ces corpus de référence, nous menons des campagnes d'annotations où plusieurs annotateurs annotent le même corpus, et nous comparons ensuite leurs annotations pour tenter d'établir une référence (Artstein & Poesio, 2008; Bayerl & Paul, 2011). La qualité de ces annotations est critique, car c'est à partir de ces corpus que de nombreux autres outils ou traitements sont produits ou réalisés.

La préparation de ces campagnes commence, entre autre, par le choix de l'outil d'annotation. Plusieurs paramètres doivent rentrer en considération lorsque nous décidons de privilégier tel ou tel outil [1] : la possibilité de satisfaire les contraintes d'annotation du phénomène à annoter, une interface accessible pour les annotateurs peu familiers avec l'informatique, ou encore le caractère *open source* de l'outil. Une autre étape importante est la construction du corpus, qui doit être le plus représentatif et adapté au phénomène étudié, de même que l'écriture du guide d'annotation.

1. Le site gouvernemental ÉTALAB a d'ailleurs mis en ligne une page dédiée aux critères de sélection de l'outil.

Dans cet article, nous nous interrogeons sur les phénomènes qui peuvent influencer le travail de l'annotateur et introduits dès cette phase de construction d'une campagne. Plus particulièrement, nous nous intéressons à trois aspects, finalement peu étudiés :

— le lien entre le contenu des items à annoter (partie 4.4) ;
— la modalité d'interaction ou de saisie (partie 4.1) ;
— la modalité de présentation.

Si le premier point concerne l'objet annoté, les deux derniers se rapportent à la manière dont nous construisons une campagne et les modalités imposées par l'environnement d'annotation. Pour la modalité de présentation, nous ne l'étudions que partiellement dans certaines analyses et cela mériterait un approfondissement ultérieur.

Pour approfondir ces trois aspects d'une campagne d'annotation, cet article suit le plan suivant. Nous présentons dans la partie 2 la tâche d'annotation choisie et l'état de l'art des corpus disponibles. En section 3, nous explicitons nos choix pour la campagne d'annotation. Nous procédons ensuite à l'analyse des résultats dans la partie 4. Enfin, dans la section 5, nous terminerons par quelques mots de conclusion et de perspectives.

2 Choix de la tâche d'annotation et état de l'art

Nous souhaitons observer et analyser les éventuels impacts de certains phénomènes sur les annotations. Pour ce faire, nous avons choisi de mener une campagne d'annotation sur du texte et avec une annotation catégorielle. La tâche d'annotation retenue est la catégorisation d'énoncés selon le fait qu'ils contiennent ou non une erreur de français. À nos yeux, il s'agit d'une tâche intéressante. Bien qu'en linguistique, une différence importante existe entre la *norme* — ce qui est prescrit par l'Académie Française, par exemple —, et l'*usage* — ce qui est effectivement observé —, les locuteurs se rapportent généralement à la norme et leurs rapports à celle-ci diffèrent selon les personnes. De plus, les annotateurs peuvent se tromper sur l'application des règles ou ne connaissent pas toutes les règles, *a fortiori* rares ou complexes.

2.1 Typologie des erreurs

Dans un premier temps, afin de mieux appréhender cette tâche, nous avons effectué des recherches concernant la typologie des erreurs. Ho-Dac *et al.* (2016), eux-mêmes inspirés de (RO & Ledegen, 2012; Roubaud, 2014; Anxionnaz, 2015), ont montré que certains types d'erreurs reviennent plus souvent et peuvent être classifiés ainsi :

Erreur orthographique : L'énoncé présente une erreur ou un problème d'homophonie. Cela regroupe des erreurs telles que l'oubli ou le rajout de consonnes, de lettres finales muettes ou d'accents, une confusion entre deux sons proches ou deux écritures (*i* ou *y*, *f* ou *ph*). Par exemple : *proffesseur* au lieu de *professeur*, *toujour_* au lieu de *toujours*.

Erreur grammaticale : L'énoncé présente une erreur d'accord (verbal ou au sein du groupe nominal), ou de conjugaison. Ainsi, nous retrouvons dans cette catégorie des erreurs comme un participe passé suivi d'un infinitif, un indicatif futur au lieu d'un conditionnel, un oubli d'accord de l'adjectif qualificatif, etc. Par exemple : *La jeune fille est tombé_* au lieu de *La jeune fille est tombée*, *S'est à toi de voir* au lieu de *C'est à toi de voir*.

Erreur syntaxique : L'énoncé présente une inversion de l'ordre des mots, une mauvaise préposi-

tion employée, etc. Par exemple : *Tu te rappelles du film d'hier* au lieu de *Tu te rappelles le film d'hier*.

D'autres types d'erreurs peuvent être distingués, comme des erreurs d'usage (utilisation de *du coup*, niveau de langue ou de registre inapproprié, etc.) ou des erreurs sémantiques. Ces catégories peuvent se recouvrir : une erreur de grammaire peut découler d'une erreur d'orthographe ou lexicale. Par ailleurs, il est parfois compliqué de savoir précisément pourquoi un locuteur se trompe : commet-il une erreur de grammaire parce qu'il ne connaît pas la règle ou parce qu'il se trompe entre différentes erreurs ?

2.2 Corpus d'erreurs de français disponibles

Nous souhaitons à présent chercher un corpus regroupant des erreurs de français. Cette recherche a pour but principal de trouver un *support* pour construire notre expérience, plutôt que d'analyser comment la campagne d'annotation pour établir ce corpus a été réalisée. Ce support doit idéalement posséder des phrases courtes contenant une erreur, proposant une certaine variété dans les erreurs et avec une difficulté variable [2].

Il existe déjà des corpus regroupant des erreurs de français. Nous pouvons notamment citer le corpus WICOPACO [3], construit et mis à disposition par Wisniewski *et al.* (2010). Cette ressource regroupe des révisions (erreurs et corrections) de pages WIKIPÉDIA en langue française, allant de simples erreurs orthographiques ou grammaticales à des reformulations. Le corpus TRACE [4] (Yvon & Segal, 2012) rassemble quant à lui des segments, avec une ou plusieurs erreurs, provenant de sources diverses : blogs, extraits de textes provenant d'un correcteur de français ou d'un traducteur automatique, ou encore fragments d'un intranet d'une société.

Les corpus regroupant les productions d'élèves de l'enseignement primaire et secondaire et d'apprenants de langue française étrangère – ou FLE – constituent aussi des candidats potentiels. La plupart de ces corpus sont encore sous forme manuscrite, les créations « sur papier » des élèves et apprenants étant seulement numérisées en format image, et sont donc inappropriés pour un traitement informatique. Néanmoins, quelques-uns de ces corpus sont disponibles dans un format textuel utilisable dans le cadre d'un traitement automatique, comme c'est le cas des corpus É-CALM [5] (Ho-Dac *et al.*, 2020), EMA [6] (Boré & Elalouf, 2017), DIRE AUTREMENT [7] (Hamel & Milicevic, 2007) ou encore CEFLE [8] (Ågren, 2008).

Toutefois, ces corpus ne correspondent pas à ce que nous souhaitons pour notre expérience. Dans les cas de WICOPACO et de TRACE, les erreurs sont assez répétitives et simples. Certaines phrases – généralement trop longues – possèdent souvent plusieurs erreurs. Ces remarques valent aussi pour les corpus d'écrits scolaires. De plus, la compréhension de certaines phrases extraites de ces corpus n'est pas toujours aisée. Enfin, s'il s'agit d'une version transcrite, la transcription rajoute souvent des annotations complexes qui sont difficiles à exploiter.

2. Pour ce dernier point, il nous semblait important d'avoir certaines erreurs difficiles à repérer, voire ambiguës, justement pour avoir une variabilité dans les réponses.
3. https://wicopaco.limsi.fr/
4. https://anrtrace.limsi.fr/
5. http://e-calm.huma-num.fr/
6. https://www.ortolang.fr/market/corpora/ema-ecrits-scolaires-1
7. http://web5.uottawa.ca/direautrement/index.html
8. https://projekt.ht.lu.se/cefle

3 Description de l'expérience

Comme nous l'avons vu, des corpus d'erreurs français sont disponibles, toutefois aucun ne nous convenait pour l'expérience. Dans cette partie, nous abordons, au prisme des trois critères évoqués en 1, la construction du corpus que nous avons décidé de créer, ainsi que la manière dont nous avons conçu l'expérience. Pour rappel, ces trois aspects sont :

— la constitution du corpus et les relations entre les items proches par leur contenu ;
— la modalité d'interaction ou de saisie : pour notre expérience, nous nous interrogeons si la possibilité du retour arrière influence ou non la qualité des annotations ;
— la modalité de présentation : en l'occurrence, il s'agit de l'ordre des items.

3.1 Objet annoté et liens entre les items

La tâche d'annotation choisie, le repérage d'erreurs de français, peut avoir trois significations différentes : (i) localiser l'erreur (ii) repérer si une phrase contient une erreur ou non (iii) catégoriser le type d'erreur. Ces trois manières ne sont pas sans rapport avec la différence qui existe entre l'*unitizing*, tel que défini par (Krippendorff, 1995), et la catégorisation : dans le premier cas il s'agit de déterminer où se situe l'occurrence du phénomène annoté dans un flux textuel, et dans les deuxième et troisième cas il faut associer l'occurrence à une catégorie précise — soit binaire (pas d'erreur/avec erreur), soit nominale (orthographe/grammaire/etc.). Généralement, les annotateurs ne s'acquittent que de la catégorisation ; pour certaines tâches d'annotation, en revanche, les annotateurs doivent faire ces deux opérations en même temps.

La question de la forme que devait prendre la tâche d'annotation s'est donc posée. Pour notre expérience, nous ne souhaitons pas complexifier la tâche en intégrant la phase *unitizing*, qui rend l'annotation plus fastidieuse et qui fait intervenir une annotation plus délicate et plus difficilement comparable. Cela nous paraît aussi en marge de nos intentions premières, notamment en multipliant les sources de désaccord possibles. Ensuite, si notre premier réflexe a été de proposer une tâche en deux temps (repérage sans/avec erreur puis catégorisation du type d'erreur), nous nous sommes vite aperçue que la tâche pourrait être décourageante pour les participants — qui répondraient sur la base du volontariat. En effet, la distinction des catégories d'erreurs demande une compétence ou d'instructions particulières que tous les annotateurs n'ont pas forcément, et certaines erreurs peuvent être ambiguës et risquent de perturber l'annotateur.

Pour cette expérience, nous avons donc choisi de restreindre la tâche à une annotation catégorielle *Sans erreur* ou *Avec erreur* au niveau de l'énoncé, sans demander à l'annotateur de préciser le type et la localisation de l'erreur. L'avantage de cette approche est que les consignes d'annotation sont immédiatement claires et compréhensibles par tous, et ne nécessitent pas d'entraînement supplémentaire pour les annotateurs.

Ce point éclairci, il nous restait aussi à décider de la répartition du nombre d'énoncés pour chaque catégorie. Nous supposons en effet que les deux catégories relèvent d'une charge cognitive et d'une complexité égales. Nous avons décidé de produire cent énoncés, répartis équitablement entre les deux catégories (cinquante énoncés *Sans erreur* et cinquante énoncés *Avec erreur*), pour éviter un biais de prévalence comme étudié dans (Di Eugenio & Glass, 2004).

Enfin, nous désirons aussi analyser le comportement des annotateurs lorsqu'ils rencontrent des occurrences presque identiques à annoter : ont-ils tendance à tenir compte d'une pour annoter l'autre,

ou au contraire, à annoter sans préjugé ? Pour tenter d'observer ce qu'il se produit dans ce cas, nous intégrons au corpus ce que nous nommons « paires d'énoncés », c'est-à-dire un même énoncé décliné dans deux versions : une *Sans erreur* et une *Avec erreur*. Ci-dessous un exemple de paire tiré du corpus :

SE Le festival est censé se dérouler au printemps.
AE Le festival est sensé se dérouler au printemps.

Toutefois, nous ne voulions pas que la tâche d'annotation soit réduite à la question de trouver la meilleure proposition entre deux. Pour éviter ce glissement de tâche d'annotation, nous avons limité le nombre de paires d'énoncés ; au total, nous avons treize paires de phrases.

Nous avons produit nous-même le corpus. Pour ce faire, nous avons créé des énoncés, d'une longueur d'une phrase. Ces énoncés contiennent soit aucune erreur, soit une erreur. L'acceptabilité de l'énoncé a été jugée en accord avec la norme de l'Académie Française.

3.2 Modalité d'interaction et de saisie : le retour arrière

Comme évoqué en 1, le choix de l'outil d'annotation n'est pas à sous-estimer, et il convient de prendre la décision la plus éclairée possible pour que la campagne d'annotation se passe au mieux et que les annotations soient de qualité suffisante. Si une partie du choix repose sur les besoins et les contraintes intrinsèques du phénomène annoté, une autre partie dépend des fonctionnalités propres à l'outil (gestion du schéma d'annotation, intégration de ressources externes, prise en charge de l'aspect collaboratif d'une campagne, etc.). Toutefois, toutes les fonctionnalités et leur impact sur l'annotation n'ont pas fait l'objet d'analyses systématiques.

Plus spécifiquement, nous nous intéressons dans cet article à la possibilité de voir et modifier ses annotations précédentes. Une grande majorité des outils d'annotation propose un retour arrière, ou du moins le fait de revenir sur ses anciennes annotations : en effet, lorsque nous annotons un flux textuel, l'annotateur a accès à l'ensemble du texte sur lequel il travaille. Cependant, cette fonctionnalité entraîne parfois des problèmes au niveau de l'ergonomie de l'outil, par exemple si le texte à annoter est trop long ou nous ne pouvons pas naviguer commodément entre les textes. De plus, certains outils ne le proposent simplement pas, comme dans le cas de (Poesio *et al.*, 2013; Fort *et al.*, 2014).

Il convient alors de s'interroger si la disponibilité ou la facilité d'utilisation de cette fonctionnalité entraîne ou non des modifications d'annotations, notamment des annotations moins fiables lorsque le retour arrière n'est pas possible. Ainsi, nous pensons qu'il est intéressant de proposer deux types de scénarios aux annotateurs : des scénarios avec retour arrière, et d'autres sans retour arrière. Cette méthode nous permettra d'avoir une approche contrastive et d'observer, peut-être, un potentiel impact de ce paramètre sur les annotations.

3.3 Modalité de présentation : ordre de présentation

La modalité de présentation, c'est-à-dire l'ordre de présentation des items à annoter, constitue un élément central de notre réflexion et méritait d'être examinée en détail. Une de nos principales préoccupations concerne la prévalence d'une catégorie à un niveau local. Si la surreprésentation d'une catégorie à l'échelle du corpus a déjà été traitée (Mathet & Widlöcher, 2016; Fort *et al.*, 2012), il n'y a pas, à notre connaissance, d'étude sur une distribution inégale présente seulement dans un segment du corpus. C'est pour cette raison que nous souhaitons présenter à certains annotateurs des séries

d'énoncés regroupés selon leur catégorie, c'est-à-dire uniquement des énoncés *Sans erreur*, puis des énoncés *Avec erreur*. Nous espérons, entre autre, pouvoir analyser un potentiel phénomène de diminution de la vigilance chez l'annotateur, une attention moindre pour détecter un changement de catégorie ou pour repérer un « intrus » d'une autre catégorie. Bien sûr, nous pourrions aussi observer le phénomène inverse, c'est-à-dire une augmentation de la vigilance.

En supplément de cette première condition, le traitement des paires d'énoncés est à examiner. Deux aspects concernant ce point sont à considérer :

— la distance entre les paires : une première observation envisageable repose sur le fait que les paires doivent être directement contiguës ou non ; il convient aussi d'approfondir le cas où les énoncés sont séparés par d'autres énoncés, notamment pour savoir si, plus les phrases d'une paire sont espacées, plus cela a un impact fort sur les annotations ;

— l'ordre interne des paires : nous souhaitons aussi regarder si le fait de présenter les énoncés des paires *Sans erreur*/*Avec erreur* ou *Avec erreur*/*Sans erreur* modifiait les annotations.

Finalement, nous nous sommes arrêtée sur quatre scénarios différents [9], dont la description est listée ci-dessous. Nous rappelons que chaque scénario est proposé dans deux versions, une sans retour arrière possible, une autre avec retour arrière disponible [10] :

Scénario 1 : Tous les énoncés *Sans erreur*, puis tous les énoncés *Avec erreur*.

Scénario 2 : Il s'agit d'une variante du scénario 1, mais où les paires sont rapprochées. L'ordre de présentation *Sans erreur* et *Avec erreur* est préservé, mais les paires sont contiguës. Pour **sept** paires de phrases, les phrases *Sans erreur* sont immédiatement suivies de la phrase *Avec erreur* ; pour les six autres paires de phrases, les phrases *Avec erreur* sont immédiatement suivies de la phrase *Sans erreur*.

Scénario 3 : Les énoncés sont présentés dans un ordre aléatoire, les paires de phrases sont contiguës et l'ordre de présentation de ces dernières (*Sans erreur*/*Avec erreur* ou *Avec erreur*/*Sans erreur*) varie selon les cas.

Scénario 4 : Les énoncés sont présentés dans un ordre aléatoire (différent du **Scénario 3**), les paires de phrases sont séparées par 33 énoncés [11] et l'ordre de présentation de ces dernières (*Sans erreur*/*Avec erreur* ou *Avec erreur*/*Sans erreur*) varie selon les cas.

3.4 Déroulement de la campagne

Afin que les résultats obtenus soient significatifs, nous devons avoir un nombre suffisant de participants répartis équitablement entre les deux types de questionnaires et entre les différents scénarios. Dans ce sens, nous avons partagé le lien de la campagne avec plusieurs groupes de participants potentiels : des étudiants de licence (de la L1 à la L3), des collègues d'un laboratoire et des doctorants.

Il y a donc une multiplicité d'origines des annotateurs. En ce sens, des questions obligatoires ont été prévues dans le questionnaire afin de déterminer, à gros grains, l'origine de l'annotateur, si cette connaissance devait nous servir lors de nos expériences. Les trois questions concernent : (i) le niveau d'études ; (ii) l'auto-évaluation du niveau de français, sur une échelle de 1 (Très mauvais) à 5 (Excellent) ; (iii) si le français est la langue maternelle.

9. Nous avons réfléchi à d'autres scénarios, notamment pour observer plus finement le lien entre les objets. Nous avons néanmoins fait le choix de nous restreindre à seulement huit scénarios : augmenter le nombre de scénarios réduisait le nombre d'annotateurs dans chaque, et les résultats risquaient de ne pas être suffisamment significatifs.

10. Du point de vue de l'implémentation logicielle, les énoncés étaient, dans ce cas-là, affichés sur une seul et même page.

11. Il est à noter que l'éloignement de 33 phrases est arbitraire.

Par ailleurs, afin d'inciter plus volontiers les personnes à participer à la campagne, nous avons aussi voulu rendre l'expérience davantage pédagogique. Cette volonté se traduit d'une part par l'affichage de score après avoir complété et envoyé les réponses, et d'autre part par la réalisation et la mise à disposition des annotateurs d'un document explicatif des réponses et des règles abordées durant l'expérience.

Après des vérifications et des corrections, la campagne est prête à être lancée en novembre 2021. Le lien est d'abord envoyé aux étudiants de L1 littéraire, puis ensuite à tout le laboratoire d'Informatique et aux doctorants d'une école doctorale scientifique.

	S1	**S2**	**S3**	**S4**	**Total**
ARA	5	10	8	4	27
SRA	2	10	7	4	23

TABLE 1: Nombre d'annotateurs par scénario. **ARA** signifie *Avec retour arrière*, **SRA** signifie *Sans retour arrière*.

4 Résultats et discussion

Dans un premier temps, il convient d'avoir une approche simple des annotations obtenues pour observer les principales tendances. Pour ce faire, nous calculons la dispersion des scores (le pourcentage d'exactitude), présentée sur le graphique de la figure 1, ainsi que la distribution effective des catégories *Sans erreur* et *Avec erreur*, présentée en dans la Figure 2.

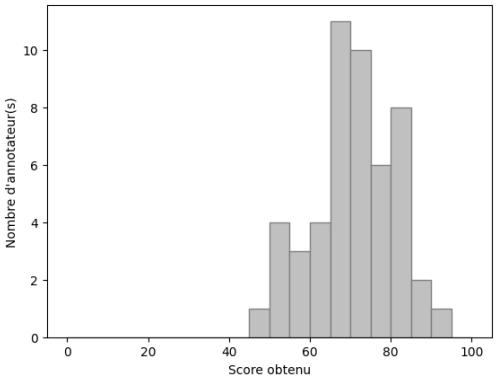

FIGURE 1: Dispersion des scores.

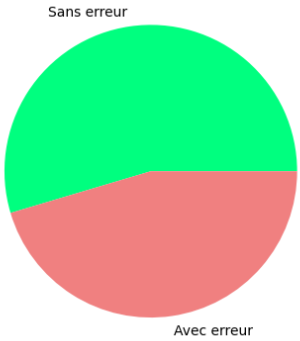

FIGURE 2: Distribution des catégories *Sans erreur* et *Avec erreur*.

Concernant les scores individuels, ils sont assez disparates : les annotateurs ont généralement un score aux alentours de 70, mais plusieurs ont plutôt des scores hétérogènes entre 55 et 85. Pour la distribution des catégories, les énoncés étant répartis équitablement entre les deux catégories, nous aurions pu nous attendre à retrouver une distribution équilibrée entre *Sans erreur* et *Avec erreur* ; ce n'est pas le cas. En effet, nous remarquons une légère tendance des annotateurs à ne pas repérer d'erreur, car il y a plus d'annotations *Sans erreur*. Concrètement, sur les 5000 annotations récoltées, il y a 2730 (54,6%) annotations *Sans erreur* pour 2270 (45,4%) annotations *Avec erreur*. Les annotateurs, lorsqu'ils ne détectent pas d'erreurs, répondraient par défaut que la phrase est correcte.

4.1 Le retour arrière influence-t-il les annotations ?

La possibilité, pour les annotateurs, de retourner sur leurs anciennes annotations et de les modifier se relève, à nos yeux, un paramètre important à prendre en compte quant aux fonctionnalités proposées par un logiciel d'annotation. Comme nous l'avons vu en 3.2, tous les logiciels ne permettent pas de regarder les anciennes annotations et de les changer. Pour observer un éventuel impact de ce paramètre sur les annotations, nous comparons donc les moyennes des scores obtenus par scénario et type de campagne. Les résultats sont présentés dans le tableau 2.

	ARA	**SRA**	**ARA + SRA**
S1	70 (± 14)	70 (± 0)	70 (± 8,57)
S2	69,3 (± 5,78)	72,1 (± 13,4)	70,7 (± 10,42)
S3	70,12 (± 4,99)	71,29 (± 11,61)	70,67 (± 8,75)
S4	73,5 (± 14,17)	63 (± 9,03)	68,25 (± 12,99)
S1 à S4	70,3 (± 8,39)	70,09 (± 12,01)	70,2 (± 10,22)

TABLE 2: Moyenne et écart-types des scores pour chaque scénario.

Nous ne notons pas de différences notables entre les moyennes des scores, tant au niveau des scénarios qu'au niveau du retour arrière possible : toutes les moyennes de scores tournent autour de 70. La seule étrangeté observée est le scénario S4–SRA, où la moyenne baisse à 63. Grâce aux écart-types importants, nous voyons aussi que les scores obtenus sont disparates au sein des scénarios.

D'un point de vue *performance*, la possibilité du retour arrière ne semble donc pas avoir un impact sur les scores globaux et généraux. D'un point de vue *pratique*, la plateforme utilisée ne nous permet pas de vérifier si les annotateurs ayant eu le retour arrière ont effectivement utilisé cette fonctionnalité. Nous regarderons, dans la partie 4.4, si le retour arrière peut plutôt avoir un impact au niveau local, notamment grâce aux paires d'énoncés.

4.2 Niveau d'expertise attribué par les annotateurs

Au début du questionnaire, nous avons posé trois questions préliminaires aux annotateurs. Une question concernait notamment l'auto-évaluation du niveau français du participant. Ils devaient aussi indiquer leur niveau d'études — cette question nous permettrait surtout de retrouver sommairement la cohorte d'origine de l'annotateur.

Cette auto-évaluation du niveau de français rejoint la pratique courante, en annotation, de demander leur indice de confiance aux annotateurs. En effet, une présupposition courante est de juger plus fiables les annotations des annotateurs s'attribuant un bon indice de confiance. Nous nous demandons alors si le niveau d'expertise, que des annotateurs non experts du domaine s'auto-attribuent, est corrélé à des annotations de qualité plus fiable. Pour vérifier cette hypothèse, nous avons calculé la moyenne des scores obtenus par les annotateurs selon leur niveau de français qu'ils s'étaient attribué, et leur niveau d'études ; le tableau 3 expose ces moyennes.

Nous remarquons déjà qu'il y a une minorité de participants qui ont estimé leur niveau de français entre 1 et 2 : la majorité s'est auto-attribué un niveau de 4 sur 5. La principale observation que nous pouvons tirer du tableau porte sur la croissance des moyennes au fur et à mesure que le niveau estimé par l'annotateur augmente. Il semblerait donc qu'il y ait une corrélation forte entre le niveau estimé

Études \ Français	1	2	3	4	5	1 à 5
Collège	∅	∅	∅	75 (1)	∅	75
Lycée	∅	∅	∅	68 (2)	∅	68
De Bac +1 à Bac +3	46 (1)	∅	65,29 (7)	68,8 (5)	∅	65,15
De Bac +4 à Bac +5	∅	62 (1)	64 (3)	71,83 (6)	90 (2)	72,08
Au-delà de Bac +6	∅	58 (1)	73,67 (6)	73,2 (10)	71 (5)	72,14
Tous niveaux	46	60	68,19	71,58	76,43	

TABLE 3: Moyennes des scores selon le niveau d'étude et le niveau de français estimé. Les nombres entre parenthèses correspondent au nombre d'annotateur(s) dans ce sous-ensemble.

de français et le score effectivement obtenu. Quant au niveau d'étude des participants, nous ne voyons pas de corrélation entre le niveau d'étude et un score typique de tel ou tel niveau.

La prochaine étape de cette expérience est d'observer à un niveau plus local, au niveau de la question, s'il est notamment possible d'identifier les points sur lesquels les annotateurs ont des difficultés et les lier à leur estimation personnelle de leur niveau. Pour ce faire, à la manière de l'expérience suivante 4.3, nous pouvons calculer le taux de réponses correctes à telle ou telle question ou la tendance à repérer ou non une erreur, selon le niveau estimé de français de l'annotateur.

4.3 Taux de réponses correctes par énoncé

Comme vu dans le tableau 2, le score moyen tourne autour de 70. Une question intéressante est de regarder si le taux de réponses (jugées) correctes est uniforme sur toutes les questions ou si nous pouvons identifier des phénomènes expliquant une disparité de réussite à certaines questions.

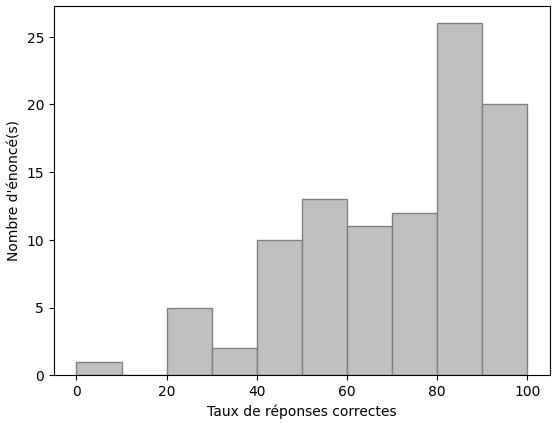

FIGURE 3: Taux de réponses correctes pour chaque énoncé.

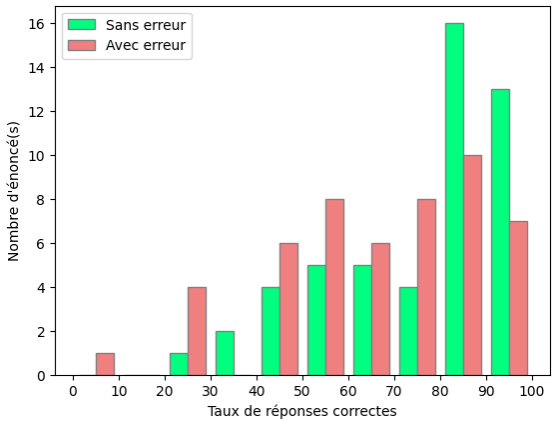

FIGURE 4: Taux de réponses correctes pour chaque énoncé, selon la catégorie.

Une première approche est de classer les questions en fonction de leur taux de réponses correctes. Nous obtenons alors l'histogramme de la figure 3. Sur cet histogramme, nous observons que presque la moitié des questions a un taux de réponses correctes d'au moins 80%. Nous remarquons aussi un plateau entre 40% et 80% de réponses correctes. Enfin, nous remarquons quelques questions atypiques avec moins de 40%.

Le premier groupe de questions repéré (plus de 80% de réponses correctes) correspond à des questions considérée comme faciles. Le deuxième regroupe des questions, utilisant des règles de français ou du vocabulaire plus complexes ou plus soutenues. Le troisième groupe rassemble plutôt des questions « pièges » ou un énoncé pour lequel l'usage ne correspond pas à la norme. Nous avons mis quelques exemples d'énoncés significatifs dans le tableau 4.

Catégorie	**Énoncé**	**Taux de bonnes réponses correctes**
AE	Tu es sûr que c'est bien de Pierre dont tu parles ?	8
SE	Au vu de leurs notes, elles ont l'air sérieux comme candidates à une bourse de thèse.	32
AE	Les cahiers oranges appartiennent à mon frère.	46
AE	Les poésies qu'il a entendues chanter en Grèce lui ont donné le sens de la prosodie.	58
SE	Nous pourrons discuter de ce point quand la suite aura été traitée.	88
AE	Il faut finaliser le travail commencer en classe sur les interprétations.	94

TABLE 4: Exemples d'énoncés, avec leur catégorie et le taux de bonnes réponses. **AE** correspond à *Avec erreur*, **SE** à *Sans erreur*. Les parties soulignées correspondent aux erreurs.

Nous souhaitons à présent regarder plus finement le taux, selon le type d'énoncé. Pour cela, nous avons raffiné l'histogramme précédent en distinguant les énoncés qui avaient une erreur de ceux n'en ayant pas. Nous obtenons alors le graphique de la figure 4. Les énoncés *Sans erreur* ont généralement un taux de réponses correctes plus élevé que les énoncés *Avec erreur*. Les annotateurs ont plutôt tendance à manquer une erreur plus que d'en inventer une. Cette observation se retrouve aussi dans les tableaux de la partie 4.4, où nous voyons que les annotateurs ont une inclinaison à indiquer deux énoncés de paires *Sans erreur*, plutôt qu'à les considérer tous les deux *Avec erreur*. Ce type de biais, interne à l'objet annoté, mériterait une étude plus poussée, notamment lorsque nous utilisons la majorité pour établir une annotation de référence, par exemple en modifiant le seuil de la majorité.

4.4 Paires d'énoncés

Comme vu dans la partie 4.1, le retour arrière ne semble pas être un paramètre préjugeant de « meilleures » annotations (au sens « plus approchant » de la référence). Nous nous demandons, alors, si l'impact n'a pas plutôt lieu au niveau local, et notamment de la stabilité des annotations fournies. Dans cette partie, pour vérifier (ou non) cette hypothèse, nous nous concentrons sur les paires d'énoncés. En repérant les paires, les annotateurs auront peut-être tendance à répondre deux réponses différentes pour chaque énoncé d'une paire. Nous pouvons imaginer qu'un annotateur ait d'abord répondu *Sans erreur* au premier énoncé d'une paire de phrase ; en voyant la seconde phrase, dans laquelle il ne repère pas non plus de erreurs, il peut vouloir reconsidérer sa précédente annotation.

Pour tester cette hypothèse, nous calculons, pour chaque annotateur d'un sous-groupe considéré, les combinaisons de réponses possibles aux paires. Il y a quatre cas possibles :

1. l'annotateur répond *Avec* pour l'énoncé *Sans erreur* (incorrect) et *Avec* pour l'énoncé *Avec erreur* (correct) ;

2. il répond *Sans* pour l'énoncé *Sans erreur* (correct) et *Avec* pour l'énoncé *Avec erreur* (correct) ;
3. il répond *Avec* pour l'énoncé *Sans erreur* (incorrect) et *Sans* pour l'énoncé *Avec erreur* (incorrect) ;
4. il répond *Sans* pour l'énoncé *Sans erreur* (correct) et *Sans* pour l'énoncé *Avec erreur* (incorrect).

Les premiers résultats, sur l'ensemble des annotateurs, sont visibles dans le tableau 5. Par exemple, la case contenant $4{,}15\%$ se réfère au cas **1**, et la case grisée correspond au cas **2** : les annotateurs ont eu des réponses correctes à chaque énoncé de la paire. En général, les annotateurs ont tendance à répondre deux réponses différentes ($53{,}08\%$ et 28%). Nous remarquons aussi une propension des annotateurs à ne pas repérer d'erreurs.

Sans erreur / **Avec erreur**	**Avec (incorrect)**	**Sans (correct)**
Avec (correct)	4,15	55,08
Sans (incorrect)	28	12,77

TABLE 5: Comparaison des réponses pour chaque paire d'énoncés (en pourcentage).

Nous raffinons ensuite le tableau précédent en distinguant les deux types de campagnes : ARA en 6a et SRA en 6b. Nous observons une plus forte tendance à annoter distinctement les paires chez les annotateurs avec un questionnaire où le retour arrière était possible. La principale observation est une augmentation forte des couples ayant deux réponses différentes au détriment de ceux ayant répondu *Sans erreur* aux deux énoncés d'une paire. Nous supposons que ce phénomène se produit, lorsque après avoir mis un *Sans erreur* au premier énoncé d'une paire, l'annotateur repère le deuxième énoncé de la paire qui lui fait reconsidérer sa première réponse. Pour cela, il faudra encore raffiner le tableau en prenant en compte l'ordre dans lequel les paires arrivaient.

Sans erreur / **Avec erreur**	**Avec**	**Sans**
Avec	3,7	58,12
Sans	29,91	8,26

(a) ARA

Sans erreur / **Avec erreur**	**Avec**	**Sans**
Avec	4,68	51,51
Sans	25,75	18,06

(b) SRA

TABLE 6: Comparaison des réponses pour chaque paire d'énoncés (en pourcentage) selon le type de campagne.

Un autre angle d'attaque est de regarder l'influence de la distance entre chaque énoncé des paires. Pour cela, nous pouvons nous appuyer sur les différents scénarios : dans les scénarios S2 et S3 les énoncés des paires se suivent, alors que dans le scénario S1 les énoncés sont séparés de 50 questions, et dans le S4 de 33 phrases. Nous obtenons les tableaux en 7 ; il est à noter que ces tableaux ne considèrent pas la possibilité ou non du retour arrière. Nous constatons effectivement que dans les scénarios S2 et S3, les annotateurs ont une forte tendance à annoter distinctement les paires ; sans toutefois donner toujours les bonnes réponses (ils se trompent respectivement pour $26{,}92\%$ et $33{,}33\%$ des paires). À l'inverse, si les paires sont trop distantes, comme c'est le cas pour les scénarios S1 et S4, les annotateurs ne semblent pas respecter de logique entre les paires.

Lorsque nous voulons comparer, à l'intérieur d'une campagne, deux éléments semblables, il est important de bien faire attention à leur espacement au sein de la campagne. En effet, leur proximité entraîne une réaction différente de l'annotateur, qui peut être souhaitable ou à éviter. Si la distance entre deux occurrences est trop importante, nous pouvons supposer que les annotateurs ont déjà oublié la première occurrence d'une paire lorsque la seconde se présente.

Avec erreur \ Sans erreur	Avec	Sans
Avec	8,79	51,65
Sans	23,08	16,48

(a) S1

Avec erreur \ Sans erreur	Avec	Sans
Avec	2,31	57,69
Sans	26,92	13,08

(b) S2

Avec erreur \ Sans erreur	Avec	Sans
Avec	3,08	56,92
Sans	33,33	6,67

(c) S3

Avec erreur \ Sans erreur	Avec	Sans
Avec	6,73	48,08
Sans	25	20,19

(d) S4

TABLE 7: Comparaison des réponses pour chaque paire d'énoncés (en pourcentage) selon le scénario.

5 Conclusions

Dans cet article, nous avons analysé les premières données obtenues. Si l'analyse de la possibilité du retour arrière pour l'annotateur n'a pas donné de résultats tangibles au niveau global, nous avons pu repérer des différences au niveau local, grâce aux paires d'énoncés. Les autres expériences présentées montrent aussi des résultats encourageants. Nous pensons donc que cette expérience pourra amener des analyses intéressantes et mérite d'être poursuivie, notamment pour obtenir une base d'annotations plus large ; une deuxième vague est actuellement en cours pour récolter de nouvelles annotations et pouvoir corroborer — ou non — les premiers résultats obtenus. Nous mettrons dès que possible le corpus créé et les annotations récoltées en accès libre.

Cependant, la principale poursuite de ce travail concerne l'élaboration d'autres métriques pour détecter d'autres phénomènes impactant les annotations. Il conviendra aussi, et surtout, de réfléchir à des techniques pour corriger les biais repérés, notamment celui d'une prévalence localement déséquilibrée des catégories qui perturberaient les annotations de manière plus sporadique qu'une prévalence déséquilibrée au niveau global.

Enfin, à la fin du questionnaire, nous avons laissé un champ d'expression libre pour permettre aux annotateurs de faire des remarques sur leur expérience. Leurs retours sont intéressants à plusieurs égards. Certains annotateurs ont noté la présence des paires d'énoncés et cela a amené de nombreuses questions (en particulier la pertinence). D'autres participants ont exprimé leur frustration de pas pouvoir corriger leurs réponses, que ce soit parce qu'ils ont répondu trop vite ou parce qu'ils se sont aperçus, plus tard, de leur erreur. Les annotateurs ayant répondu à un scénario où les énoncés étaient regroupés par catégorie s'en sont aussi aperçu et en ont fait part. Enfin, certains retours nous ont aussi permis de corriger certains énoncés pour lesquels nous avions introduit involontairement une erreur ou une approximation d'un autre type (par exemple, oubli d'un point à la fin d'une phrase).

Au final, cette campagne d'annotation s'est révélée fructueuse. Au départ conçue pour mesurer le biais introduit par la fonctionnalité du retour arrière, elle nous a permis via les paires d'énoncés de mettre en évidence un phénomène intéressant : le biais que lorsque nous sommes confrontés à deux énoncés presque identiques, nous cherchons instinctivement à les différencier dans une catégorisation binaire [12].

12. Ceci est très bien souligné par le commentaire suivant : « Beaucoup d'interrogations concernant des séries de deux phrases similaires où il y a forcément une seule erreur ».

Références

ÅGREN M. (2008). *À la recherche de la morphologie silencieuse : sur le développement du pluriel en francais L2 écrit*. Thèse de doctorat.

ANXIONNAZ S. (2015). Le barème graduel : l'évaluation dela dictée au service des apprentissages. *Glottopol*, **26**, 135–157.

ARTSTEIN R. & POESIO M. (2008). Inter-coder agreement for computational linguistics. *Comput. Linguist.*, **34**(4), 555–596. DOI : 10.1162/coli.07-034-R2.

BAYERL P. S. & PAUL K. I. (2011). What determines inter-coder agreement in manual annotations ? a meta-analytic investigation. *Computational Linguistics*, **37**(4), 699–725. DOI : 10.1162/COLI_a_00074.

BORÉ C. & ELALOUF M.-L. (2017). Deux étapes dans la construction de corpus scolaires : problèmes récurrents et perspectives nouvelles. *Corpus*, **16**, 31–63. DOI : 10.4000/corpus.2731, HAL : hal-02980396.

DI EUGENIO B. & GLASS M. (2004). Squibs and discussions : The kappa statistic : A second look. *Computational Linguistics*, **30**(1), 95–101. DOI : 10.1162/089120104773633402.

FORT K. (2016). *Collaborative Annotation for Reliable Natural Language Processing : Technical and Sociological Aspects*. Wiley-ISTE. HAL : hal-01324322.

FORT K., FRANÇOIS C., GALIBERT O. & GHRIBI M. (2012). Analyzing the Impact of Prevalence on the Evaluation of a Manual Annotation Campaign. In *International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey. HAL : hal-00709174.

FORT K., GUILLAUME B. & CHASTANT H. (2014). Creating Zombilingo, a Game With A Purpose for dependency syntax annotation. In *Gamification for Information Retrieval (GamifIR'14) Workshop*, Amsterdam, Netherlands. HAL : hal-00969157.

HAMEL M.-J. & MILICEVIC J. (2007). Analyse d'erreurs lexicales d'apprenants du fls : démarche empirique pour l'élaboration d'un dictionnaire d'apprentissage. *Canadian Journal of Applied Linguistics*, **10**(1), 25–45.

HO-DAC L.-M., FLEURY S. & PONTON C. (2020). E :calm resource : a resource for studying texts produced by French pupils and students. In *Proceedings of the 12th Language Resources and Evaluation Conference*, p. 4327–4332, Marseille, France : European Language Resources Association.

HO-DAC L.-M., MULLER S. & DELBAR V. (2016). L'anticorrecteur : outil d'évaluation positive de l'orthographe et de la grammaire. In *Conférence conjointe JEP-TALN-RECITAL*, volume 2 de *Actes de la conférence conjointe JEP-TALN-RECITAL*, p. 333–341, Paris, France. HAL : hal-01378351.

KRIPPENDORFF K. (1995). On the Reliability of Unitizing Continuous Data. *Sociological Methodology*, **25**, 47. DOI : 10.2307/271061.

MATHET Y. & WIDLÖCHER A. (2016). Évaluation des annotations : ses principes et ses pièges. *Revue TAL*, **57**(2), 73–98. HAL : hal-01712282.

POESIO M., CHAMBERLAIN J., KRUSCHWITZ U., ROBALDO L. & DUCCESCHI L. (2013). Phrase detectives : Utilizing collective intelligence for internet-scale language resource creation. *ACM Trans. Interact. Intell. Syst.*, **3**(1). DOI : 10.1145/2448116.2448119.

RO G. & LEDEGEN G. (2012). Orthographe : ce qui est jugé difficile. *Glottopol*, **19**, 17–36. HAL : hal-01114713.

ROUBAUD M.-N. (2014). *De la description de la langue à son enseignement*. Habilitation à diriger des recherches, UNIVERSITÉ STENDHAL – GRENOBLE 3. HAL : tel-01102494.

WISNIEWSKI G., MAX A. & YVON F. (2010). Recueil et analyse d'un corpus écologique de corrections orthographiques extrait des révisions de wikipédia. In *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, p. 121–130, Montréal, Canada : ATALA.

YVON F. & SEGAL N. (2012). *Des corpus d'erreurs pour TRACE*. Rapport interne, LIMSI.