De la complexité de l'annotation manuelle : méthodologie, biais et recommandations

Anaelle Baledent

Normandie Université

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité INFORMATIQUE

Préparée au sein de l'Université de Caen Normandie

De la complexité de l'annotation manuelle : méthodologie, biais et recommandations

Présentée et soutenue par
ANAELLE BALEDENT

Thèse soutenue le 01/12/2022 devant le jury composé de		
M. FREDERIC LANDRAGIN	Directeur de recherche au CNRS, LATTICE, ENS, Montrouge	Rapporteur du jury
MME SOPHIE ROSSET	Directeur de recherche, Université Paris Saclay	Rapporteur du jury
MME IRIS ESHKOL-TARAVELLA	Professeur des universités, Université Paris-Nanterre	Membre du jury
MME KARËN FORT	Maître de conférences, Sorbonne Université	Membre du jury
MME LYDIA-MAI HO-DAC	Maître de conférences, Université Toulouse Jean Jaurès maison de la Recherche	Membre du jury
M. ANTOINE WIDLÖCHER	Maître de conférences, Université de Caen Normandie	Membre du jury Co-encadrant
M. JEAN YVES ANTOINE	Professeur des universités, UNIVERSITE DE TOURS	Président du jury

Thèse dirigée par YANN MATHET (Groupe de recherche en informatique, image, automatique et instrumentation)

UNIVERSITÉ CAEN NORMANDIE

MIIS

GREYC
Laboratoire de recherche en sciences du numérique

À Lafayette

Remerciements

Les premiers remerciements s'adressent à Yann Mathet et Antoine Widlöcher, pour m'avoir accompagnée tout au long de cette thèse et pour tout ce qu'ils m'ont apporté durant ces trois années.

Je remercie mes rapporteurs, Frédéric Landragin et Sophie Rosset pour leurs précieux retours et commentaires. Un grand merci aussi à Jean-Yves Antoine, Iris Eshkol-Taravella, Karën Fort et Lydia-Mai Ho-Dac pour avoir accepté de siéger dans mon jury de thèse.

Je garderai avec moi le souvenir de tous les enseignants qui ont contribué à enrichir mon goût pour les études. Tout particulièrement, je tenais à remercier Gaël Lejeune, qui a permis à l'étudiante de Master que j'étais alors de découvrir le monde de la recherche. Je remercie aussi Karine Abiven et Alice Millour, pour leurs échanges et leurs conseils.

J'ai pu profiter de l'accueil chaleureux des membres du couloir du Bâtiment S3. Merci à Codag, Amacc, Mad, ainsi qu'au personnel administratif (à Sophie, à Agnès) et aux administrateurs systèmes (à Davy, à Renaud). J'ai aussi une pensée pour le service DDA.

J'adresse plus spécialement des remerciements aux membres de #LGDLSPB : Alexis, Céline, Josselin, Justine, Lauréline, Matthieu, Nadjet, Pierre, Sébastien et Virginie. La fin de thèse aurait été encore plus difficile sans vous.

Un grand merci à mes ami·e·s Diego, Gwennola, Luce et Marion d'avoir été présent·e·s à mes côtés durant mes études et pour tous les bons moments passés ensembles. J'espère qu'ils seront encore nombreux.

Enfin, mes plus sincères et chaleureux remerciements s'adressent à ma famille, qui m'a toujours soutenue, malgré les moments de doutes et de stress :

- ❀ à ma Maman d'amour chéri ;
- ✯ à mon Papa d'amour chéri ;

❀ à ma Tatie d'amour chéri ;

✯ à ma Mamie d'amour chéri ;

❀ à mon Papi d'amour chéri ;

✯ à Maximus, Scarlett et Thalie.

Un merci tout particulier va à mon Gaétan.

Table des matières

Introduction **2**

Méthodologie des campagnes d'annotation 5

Biais d'annotation et recommandations 6

Présentation du plan 6

I État de l'art **9**

1 Mener une campagne d'annotation **11**

1.1 Appréhender le phénomène à annoter 14

1.1.1 Perception de l'objet étudié 14

1.1.2 Modélisation du phénomène : préparer une première version du guide d'annotation 14

1.1.3 Prendre la complexité en compte 17

1.2 Constituer le corpus de textes à annoter 18

1.2.1 Sélection des textes 19

1.2.2 Sources possibles des textes et corpus déjà disponibles 20

1.3 Choisir l'outil d'annotation 22

1.3.1 Outils existants 22

1.3.2 Critères à considérer pour le choix de l'outil d'annotation 29

1.4 Choisir et accompagner les annotateurs 31

1.4.1 Quelle expertise ? 31

1.4.2 Nombre d'annotateurs 32
1.4.3 Accompagner les annotateurs 33
1.4.4 Rester à l'écoute des annotateurs 34
1.5 Évaluer les annotations 35
1.5.1 Mesurer l'accord inter-annotateurs 36
1.5.2 Mieux appréhender l'accord inter-annotateurs 41
1.6 Établir une référence 43
1.6.1 Méthodes pour établir une référence 43
1.6.2 Problèmes liés à ces méthodes 44
1.7 Diffuser le corpus 45
1.7.1 Formats des annotations 45
1.7.2 Mettre à disposition le corpus 49
1.8 Animer une campagne 50
1.8.1 Des projets parfois de (très) longue haleine 51
1.8.2 L'annotation manuelle, à quel prix ? 51
1.8.3 Conjuguer éthique, réglementation et recherche 52
1.9 Conclusion 53

2 Typologie des différents types d'annotation 55

2.1 Comment aborder une tâche complexe ? 56
2.1.1 Tout annoter simultanément 56
2.1.2 Décomposition en plusieurs tâches d'annotation 57
2.2 Illustration pour chaque type d'ancrage 57
2.2.1 Unités déjà définies 58
2.2.2 Ancrage avec position minimale 62

2.2.3 Segmentation . . . 63
2.2.4 Unitizing . . . 64
2.2.5 Mise en relation . . . 66
2.3 Deux exemples ciblés . . . 68
2.3.1 Entités nommées . . . 69
2.3.2 Coréférence . . . 73
2.3.3 Un manque d'harmonisation . . . 77
2.4 Conclusion . . . 78

II Étude des biais et expérimentations 79

3 Biais d'annotation 81
3.1 Vers une première classification des biais . . . 82
3.1.1 Classification thématique . . . 83
3.1.2 Classification temporelle . . . 90
3.2 Méthodologie des campagnes d'annotation « Portraits » et « Erreurs » . . . 92
3.2.1 Comment étudier un biais d'annotation ? . . . 92
3.2.2 Nos hypothèses et nos attentes initiales vis-à-vis des expériences . . . 93
3.2.3 Un outil transversal : la consensualité . . . 94
3.3 Conclusion . . . 100

4 Campagne d'annotations « Portraits » 101
4.1 Présentation de la campagne . . . 103
4.1.1 Constitution du corpus . . . 103
4.1.2 Biais concernant l'estimation de l'âge . . . 104
4.1.3 Scénarios . . . 105

4.1.4 Déroulement de la campagne d'annotation . . . 106
4.1.5 Une première approche des annotations récoltées : comparaison avec la référence . . . 108
4.2 Analyse des consensualités . . . 109
4.2.1 Rang de consensualité *versus* rang de performance . . . 110
4.2.2 Retirer les annotateurs les moins consensuels . . . 112
4.2.3 Distinguer les consensualités initiale et dynamique . . . 114
4.2.4 Tester l'homogénéité de la consensualité . . . 116
4.3 Influence de l'ordre des items . . . 126
4.3.1 Avec un accès à la référence . . . 127
4.3.2 Détecter un biais sans un accès à la référence . . . 130
4.4 Résultats complémentaires . . . 135
4.5 Conclusion . . . 138

5 Campagne d'annotation « Erreurs » 141

5.1 Typologie des erreurs et corpus disponibles . . . 143
5.1.1 Typologie des erreurs . . . 143
5.1.2 Corpus d'erreurs de français disponibles . . . 144
5.2 Présentation de la campagne . . . 145
5.2.1 Objet annoté et liens entre les items . . . 146
5.2.2 Modalité d'interaction et de saisie : le retour arrière . . . 148
5.2.3 Modalité de présentation : ordre de présentation . . . 149
5.2.4 Déroulement de la campagne . . . 150
5.2.5 Première approche des résultats . . . 153
5.3 Traiter deux cohortes hétérogènes ? . . . 154

5.3.1 Étude des scores . . . 154
5.3.2 Comment traiter une telle disparité ? . . . 157
5.3.3 Réflexions et discussions : motivation et volition des annotateurs . 158
5.4 Utiliser la consensualité pour une annotation catégorielle binaire . . . 159
5.4.1 Adaptation des formules . . . 159
5.4.2 Étude globale de la consensualité et de l'imperfection . . . 160
5.4.3 Consensualité des phrases . . . 164
5.5 Retour arrière possible et paires . . . 167
5.5.1 La possibilité du retour arrière influence-t-elle les annotations ? . . 167
5.5.2 Paires d'énoncés . . . 168
5.6 Résultats complémentaires . . . 172
5.6.1 Niveau d'expertise attribué par les annotateurs . . . 172
5.6.2 Taux de réponses correctes par énoncé . . . 174
5.6.3 Outil non adapté pour l'analyse des biais ? . . . 176
5.7 Conclusion . . . 177

Synthèse **180**

Conclusions et perspectives . . . 180
Le biais : ce qui peut influencer les annotateurs . . . 180
Situer un annotateur par rapport à un groupe . . . 182
Recommandations . . . 183
Choisir . . . 184
Être attentif . . . 186
Documenter . . . 187
Il n'y a pas de référence absolue . . . 188

Se donner les moyens d'observer 190

Bibliographie **193**

Annexes **217**

A Campagne d'annotations « Portraits » **218**

A.1 Texte d'appel à participation transmis aux étudiants de la licence HUMANITÉ 218

A.2 Corpus Portraits 219

B Campagne d'annotations « Fautes » **224**

B.1 Texte d'appel à participation 224

B.2 Guide d'annotation 224

B.3 Corpus Erreur 225

B.4 Document explicatif des réponses et des règles 230

Table des figures

1.1 Étapes d'une campagne d'annotation manuelle. 12

1.2 Cycles MATTER et MAMA repris de Pustejovsky et Stubbs (2012). 13

1.3 Illustration des différents types d'ancrage. 15

1.4 Synthèse des six dimensions de complexité définies par Fort, Nazarenko et al. (2012) (version française du diagramme reprise de la thèse de Fort (2012)). 18

1.5 Exemple d'annotation via l'interface d'un traitement de textes WYSIWYG. Texte provenant de l'épilogue des « Trois mousquetaires » d'Alexandre Dumas. 24

1.6 Exemple d'étiquetage morpho-syntaxique annoté via un tableur. 24

1.7 Exemple de corpus annoté via l'interface brat. 26

1.8 Exemple de corpus annoté via l'interface WebAnno. 27

1.9 Interface du jeu JeuxDeMots. 28

1.10 Exemple d'annotation sur le jeu Zombiludik, jeu frère de ZombiLingo. . . 28

1.11 Problématiques de l'*unitizing*, figure reprise de Mathet et al. (2015). . . . 39

1.12 Seuils de fiabilité de l'accord inter-annotateur (pour la mesure κ) (Bregeon et al., 2019 ; Fort, 2022). 42

2.1 Schéma de la coréférence, repris de Delaborde (2020). 73

3.1 Schématisation des deux notions. 95

3.2 Vue d'ensemble du corpus. 98

3.3 Onglet Annotations, regroupant toutes les annotations. 99

3.4 Onglet Annotateurs, permettant de voir l'imperfection individuelle. 99

4.1 Âge moyen trouvé par les annotateurs pour chaque photographie (ensemble des annotations). 109

4.2 Âge moyen trouvé par les annotateurs pour chaque annotateur selon les vagues. 110

4.3 Rangs des annotateurs selon leur performance et leur consensualité (*consensualité dynamique*) 111

4.4 Imperfection de l'annotateur en retirant l'annotateur le moins consensuel à chaque fois 113

4.5 Imperfection de groupe en retirant l'annotateur le moins consensuel à chaque fois (consensualité dynamique). 114

4.6 Moyenne des imperfections pour les annotateurs les plus consensuels 115

4.7 Boîtes à moustache représentant la distribution des rangs pour chaque annotateur, basée sur un échantillon de 100 sous-corpus de 50 photographies aléatoires. 117

4.8 Exemple de diagrammes en violon représentant la distribution des rangs pour six annotateurs de distribution de la consensualité pour 100 sous-corpus de 50 photographies aléatoires. 118

4.9 Boîtes à moustache représentant la distribution des degrés de consensualité pour chaque annotateur, basée sur un échantillon de 100 sous-corpus de 50 photographies aléatoires. 120

4.10 Évolution de la variance des moyennes des rangs selon la taille des sous-corpus, en fonction des consensualités initiale et dynamique. 122

4.11 Exemples d'évolution de la distribution des rangs de consensualité pour six annotateurs choisis. 123

4.12 Évolution de la performance de groupe pour les n% annotateurs les plus consensuels (en moyenne). 125

4.13 Distribution des moyennes pour 6 parmi 52 groupes (erreur individuelle). . 128

4.14 Différences des moyennes attribuées par les annotateurs des scénarios **S1** et S4. 132

4.15 Régression linéaire synthétisant l'évolution des écarts par rapport à la référence 136

4.16 Erreurs globales pour chaque combinaison de $p \in [20; 40]$ et $f \in [0{,}1; 0{,}30]$. 137

5.1 Exemples tirés du corpus WICOPACO 145

5.2 **Avec** retour arrière possible 151

5.3 **Sans** retour arrière 151

5.4 Retour arrière : différence de présentation. 151

5.5 Dispersion des scores. 153

5.6 Moyenne des scores selon les niveaux d'études des participants. 155

5.7 Dispersion des scores. 156

5.8 Rangs des annotateurs selon leur performance et leur consensualité (consensualité dynamique). 161

5.9 Rangs des annotateurs selon leur performance et leur consensualité (consensualité dynamique), sans ou seulement avec les collégiens. 162

5.10 Imperfection de groupe en fonction les n% annotateurs les plus consensuels, pour les deux types de consensualité. 163

5.11 Variance pour chaque énoncé, selon la cohorte considérée. 165

5.12 Taux de réponses correctes pour chaque énoncé. 174

5.13 Taux de réponses correctes pour chaque énoncé, selon la catégorie. 176

5.14 Exemple de grille de comparaison, extraite de HO-DAC et POUDAT (2021). 185

Liste des tableaux

1.1 Grille de catégorisation externe : liste hiérarchique des paramètres situationnels. Adaptation de l'anglais depuis Biber (1993). 20

2.1 Notions liées à la coréférence, repris de Delaborde (2020). 74

2.2 Évaluation de la coréférence, repris de Lion-Bouton et al. (2020). Les images sont reprises de cet article. 76

3.1 Une autre classification des biais, selon les dimensions Collectif/Individuel. 91

4.1 Répartition des photographies selon les tranches d'âge. 104

4.2 Nombre d'annotateurs par scénario (vague 1) 107

4.3 Moyennes des erreurs absolues en fonction du scénario (en année). 128

4.4 Moyennes des erreurs absolues sur les tranches inversées, selon le scénario (en année). 130

4.5 Moyennes des erreurs absolues par rapport à l'annotation moyenne en fonction du scénario (en année). 131

4.6 Moyennes des erreurs absolues par rapport à l'annotation moyenne sur les tranches inversées, selon le scénario (en année). 131

4.7 Écart-type moyen selon le scénario. 134

4.8 Écart-type moyen sur les tranches inversées 134

5.1 Exemples d'énoncés extraits du corpus. 147

5.2 Nombre d'annotateurs par scénario à l'issue du premier appel à participation. **ARA** signifie *Avec retour arrière*, **SRA** signifie *Sans retour arrière*. . 152

5.3 Énoncés corrigés entre les deux vagues. Les correction apparaissent en **gras**. Pour les énoncés SF030, SF043 et AF032, un point final a été ajouté. . . . 152

5.4 Nombre d'annotateurs par scénario pour la vague 2. 153

5.5 Distribution des catégories *Sans erreur* et *Avec erreur*, selon les deux vagues. 154

5.6 Nombre de participants par vague et par niveau d'études. 155

5.7 Nombre d'annotateurs par scénario pour la cohorte des Non Collégiens. . . 157

5.8 Exemples de remarques écrites par les collégiens de la vague 2. 158

5.9 Énoncés dont les réponses sont les plus consensuelles (partie haut du tableau) et les moins consensuelles (partie basse). Les lignes sur fond gris correspondent à la cohorte des non collégiens. 166

5.10 Moyenne et écarts-types des scores pour chaque scénario, pour la cohorte des non collégiens. 168

5.11 Comparaison des réponses pour chaque paire d'énoncés (en pourcentage). 169

5.12 Comparaison des réponses pour chaque paire d'énoncés (en pourcentage) selon le type de campagne. 169

5.13 Comparaison des réponses pour chaque paire d'énoncés (en pourcentage) selon le scénario. 170

5.14 Comparaison des réponses pour chaque paire d'énoncés (en pourcentage) pour les scénarios S3 et S4 selon la modalité. 171

5.15 Comparaison des réponses pour chaque paire d'énoncés (en pourcentage) pour les scénarios S2 et S1 selon la modalité. 172

5.16 Moyennes des scores selon le niveau d'études et le niveau de français estimé. Les nombres entre parenthèses correspondent au nombre d'annotateurs dans ce sous-ensemble. 173

5.17 Exemples d'énoncés, avec leur catégorie et le taux de bonnes réponses. **AE** correspond à *Avec erreur*, **SE** à *Sans erreur*. 175

Introduction

Introduction

Les ressources langagières constituent une des bases fondamentales du Traitement Automatique des Langues (T.A.L.) et de la Linguistique Computationnelle (L.C.). Ces ressources sont de plusieurs sortes : des productions textuelles, des ressources de synthèse (lexique, dictionnaires, réseaux de connaissances…), des outils de traitement, etc. C'est à partir de ces ressources que les chercheurs et utilisateurs travaillent, soit pour étudier un phénomène linguistique, soit pour créer de nouveaux outils ou applications. Si certaines tâches du T.A.L. et de la L.C. peuvent s'appuyer uniquement sur des productions textuelles brutes, d'autres nécessitent des ressources enrichies, ou annotées.

Par ressources textuelles brutes, nous entendons des textes sur lesquels aucun traitement n'a été effectué, laissés et utilisés dans l'état dans lequel ils ont été récupérés. Les ressources enrichies, quant à elles, contiennent des méta-informations sur les données, obtenues soit manuellement, soit automatiquement. Ces méta-informations peuvent être, entre autres, des renseignements sur la structure des documents ou un ajout interprétatif, appelé aussi *annotation*, sur une donnée. Les annotations, selon le Centre National de Ressources Textuelles et Lexicales (CNRTL), sont des « remarques manuscrites notées en marge d'un texte ». Eshkol-Taravella (2015) distingue trois types d'annotation :

— l'ajout de gloses ponctuelles sur un document ;
— une annotation au niveau du document, par exemple les méta-données du document ;
— une annotation intratextuelle, pour rajouter des informations.

Le dernier type d'annotation constitue l'objet principal de notre mémoire.

Si le corpus annoté est de taille suffisamment importante et représentatif de la langue ou du phénomène, le corpus peut constituer ce que nous nommons corpus de référence, ou *gold standard*. Grâce à ce dernier, nous pouvons alors étudier un phénomène linguistique. Ainsi, il sera possible d'observer sa fréquence d'apparition, les éventuelles conditions dans

lesquelles il a plus de chance de survenir ou encore la structure du phénomène. L'évaluation des sorties d'un système (par exemple, à base de règles symboliques) pour l'analyse de tels phénomènes est permise par ces mêmes corpus de référence. Ces corpus servent d'étalon afin de comparer différents systèmes et leur performance. À l'heure actuelle, sont clairement florissantes les méthodes d'apprentissage. Ces méthodes d'apprentissage ont dans la plupart des cas besoin de données annotées. Il faut suffisamment de données, et l'entraînement d'un système pour détecter ou analyser ces phénomènes devient alors possible.

L'établissement de corpus de référence est donc primordial : la fabrication d'applications et d'outils du T.A.L. en dépend. Certaines applications peuvent s'accommoder de données bruitées, toutefois les erreurs systématiques demeurent problématiques. Si la base se révèle de qualité insuffisante, le reste de la chaîne de traitement est susceptible de s'en trouver altéré et engendre des potentielles dégradations de la qualité. Par exemple, les outils entraînés sur cette base peuvent apprendre des erreurs, et cela peut se répercuter en cascade quand plusieurs traitements sont effectués. Cela devient alors un effet « plafond de verre » (Manning, 2011) : au final, la mauvaise qualité de l'annotation initiale empêche de dépasser un certain seuil, quelle que soit la méthode utilisée.

La construction de ces *gold standard* demeure une tâche non triviale, par leur importance et la délicate question que pose l'établissement d'une référence (Artstein & Poesio, 2008 ; Bayerl & Paul, 2011) : l'annotation reste un processus souvent largement subjectif et qui peut dépendre du contexte. Il est même des cas où une telle référence est impossible ou inatteignable. La création automatique de ces ressources s'avère souvent impossible, et nous avons alors recours à de l'annotation manuelle, réalisée par des humains. Or, le processus d'annotation s'avère délicat : les tâches se révèlent parfois complexes, requérant davantage d'interprétation, et quelques fois un degré de subjectivité notable.

Un usage généralement admis consiste à procéder à une annotation multiple, c'est-à-dire faire annoter les mêmes données par plusieurs annotateurs. De leurs annotations est ensuite établie une annotation de référence, si leur accord est jugé satisfaisant. Mais l'enrichissement collaboratif des données soulève de nombreux questionnements théoriques et pratiques (Fort, 2016 ; Mathet & Widlöcher, 2016) : notamment, la modélisation informatique d'un phénomène n'est pas toujours aisée, ou un bon accord entre les annotateurs n'est pas forcément gage de validité des annotations. L'annotation manuelle

implique aussi des coûts temporels et financiers parfois importants (Böhmová et al., 2003 ; Martínez Alonso et al., 2016).

Se pose la question suivante : comment obtenir une référence fiable ? Nous pouvons étendre cette question en nous interrogeant sur la manière de le faire avec peu de données disponibles, dans le cas de contraintes financières ou temporelles limitantes. En effet, nous avons besoin de garantir la fiabilité de ces annotations, tout en limitant les coûts. Cette interrogation amène des réflexions sur le processus d'annotation, sur la manière dont sont réalisées les campagnes d'annotation, ainsi que des questions plus concrètes, concernant notamment les aspects techniques de l'annotation. Nous présentons dans la suite de cette introduction le contexte de notre recherche, avant d'introduire le principal apport de notre travail. Enfin, nous détaillons la structuration de ce manuscrit.

Méthodologie des campagnes d'annotation

Nous réalisons souvent l'annotation multiple au travers d'une ou plusieurs campagnes d'annotation. Le déroulement d'une campagne s'effectue en plusieurs étapes, notamment la préparation (ce qui est lié à une première approche du phénomène et de la tâche demandée), puis l'annotation d'un corpus de textes, réalisée par plusieurs annotateurs grâce à un outil et à un guide d'annotation, et contrôlée par une évaluation régulière. Enfin, l'exploitation des résultats est rendue possible au moyen de l'établissement d'une référence et de la diffusion de celle-ci.

La construction de données de référence doit donc faire l'objet d'une attention accrue, notamment lors de campagnes d'annotation manuelle qui concentrent de nombreuses et épineuses difficultés. Ces dernières impliquent de multiples aspects, déjà étudiés dans la littérature mais souvent de manière séparée. Dans cette thèse, nous prenons le parti de les appréhender dans leur ensemble.

Dans le cadre du processus d'annotation, nous souhaitons mener un examen critique des conditions dans lesquelles est produite l'annotation. Nous nous interrogeons plus particulièrement sur ce qui peut perturber ce processus et, de ce fait, avoir des conséquences négatives sur la fiabilité des annotations.

Pour ce faire, nous nous appuierons sur des campagnes en environnement contrôlé, que nous avons menées. Nous avons notamment fait le choix d'avoir des tâches d'annotations

qui soient à la fois simples, interprétatives et ne requérant aucun entraînement particulier. Durant ces campagnes, nous avons tenu au maximum à être vigilante sur les phénomènes perturbateurs extérieurs. Nous nous sommes aussi interrogée sur la procédure appliquée et les améliorations futures que nous pourrions y apporter.

Biais d'annotation et recommandations

Dans cette optique, l'apport du travail présenté dans ce mémoire réside en l'introduction d'une définition et d'une classification des biais d'annotation. Pour poursuivre cette étude, nous nous interrogeons sur la méthode d'identification et d'observation de biais susceptibles de perturber l'annotation. Nous cherchons à poser les bases d'une méthode d'analyse à travers des exemples de campagnes dédiées à l'examen de certains biais.

Les biais étudiés dans ce mémoire sont de deux ordres. D'une part, nous étudions l'organisation des items au sein d'une campagne, et plus particulièrement l'ordre de présentation à l'annotateur, ainsi qu'au cas où des items proches par leur contenu surviennent dans le corpus. D'autre part, nous nous intéressons aux modalités de la tâche d'annotation, au travers de la possibilité du retour arrière. Nous nous interrogeons aussi au rapport entre l'accord des annotateurs et la validité des annotations.

Étant implicites, les biais sont difficile à repérer et même décelés, leur traitement n'est pas aisé, voire impossible dans certains cas. La référence ainsi produite ne reflétera peut-être pas la « vérité » du phénomène. Les gestionnaires de campagne doivent donc faire preuve d'une attention active pour contrôler les biais qui peuvent survenir. C'est pour cette raison, en complément de la méthode présentée, que nous désirons présenter des recommandations, afin d'aider les gestionnaires à être vigilants à certains aspects du processus. Ces recommandations s'inscrivent dans la perspective de fournir un guide des bonnes pratiques, afin de garantir la fiabilité des annotations.

Présentation du plan

Ce mémoire s'organise de la façon suivante.

La première partie concerne l'état de l'art des campagnes d'annotation, où nous met-

tons en exergue les points problématiques où des biais sont susceptibles d'être introduits. Dans le chapitre 1, nous clarifions les différentes facettes d'une campagne d'annotation en présentant de manière synthétique les différentes étapes et les points de vigilance associés. Le chapitre 2 s'intéresse à la diversité des objets annotables, en s'appuyant sur des exemples concrets et en soulignant les problématiques qu'ils soulèvent.

La deuxième partie se concentre sur les biais d'annotation. Dans le chapitre 3, nous proposons une définition de la notion de biais, avant d'exposer une classification des biais. Dans les chapitres 4 et 5, nous posons les bases d'une méthode d'analyse des biais à travers deux campagnes dédiées à certains d'entre eux. La campagne « Portraits » (chapitre 4) permet notamment de nous intéresser :

1. au rapport entretenu entre l'accord inter-annotateurs et la validité des annotations ;
2. à une méthodologie pour étudier un biais d'annotation, celui de l'ordre des items.

La campagne « Erreurs » (chapitre 5) nous permet de présenter des premiers résultats concernant l'influence de la possibilité du retour arrière sur les annotations, ainsi que la réaction des annotateurs face à des items proches par leur contenu.

La conclusion (conclusion et perspectives) termine cette thèse et propose des pistes de recherche s'appuyant sur le travail réalisé dans cette thèse. La dernière partie a aussi pour objectif de proposer, à partir de nos observations, des recommandations plus générales (Recommandations).

Première partie

État de l'art

Chapitre 1

Mener une campagne d'annotation

Sommaire

1.1 Appréhender le phénomène à annoter 14
1.1.1 Perception de l'objet étudié 14
1.1.2 Modélisation du phénomène : préparer une première version du guide d'annotation 14
1.1.3 Prendre la complexité en compte 17
1.2 Constituer le corpus de textes à annoter 18
1.2.1 Sélection des textes 19
1.2.2 Sources possibles des textes et corpus déjà disponibles 20
1.3 Choisir l'outil d'annotation 22
1.3.1 Outils existants 22
1.3.2 Critères à considérer pour le choix de l'outil d'annotation . . . 29
1.4 Choisir et accompagner les annotateurs 31
1.4.1 Quelle expertise ? 31
1.4.2 Nombre d'annotateurs 32
1.4.3 Accompagner les annotateurs 33
1.4.4 Rester à l'écoute des annotateurs 34
1.5 Évaluer les annotations 35
1.5.1 Mesurer l'accord inter-annotateurs 36
1.5.2 Mieux appréhender l'accord inter-annotateurs 41
1.6 Établir une référence 43
1.6.1 Méthodes pour établir une référence 43
1.6.2 Problèmes liés à ces méthodes 44
1.7 Diffuser le corpus 45
1.7.1 Formats des annotations 45

1.7.2 Mettre à disposition le corpus 49
1.8 Animer une campagne 50
1.8.1 Des projets parfois de (très) longue haleine 51
1.8.2 L'annotation manuelle, à quel prix ? 51
1.8.3 Conjuguer éthique, réglementation et recherche 52
1.9 Conclusion 53

Ce chapitre a pour objectif de présenter une synthèse des problèmes rencontrés lors des différentes étapes d'une campagne d'annotation. Nous souhaitons ainsi attirer l'attention des gestionnaires de campagnes sur des points de vigilance, afin qu'ils fassent preuve de prudence durant leur campagne. Nous soulevons aussi certains manquements et problèmes qui peuvent être rencontrés par les acteurs d'une campagne d'annotation. Autant que faire se peut, nous étayerons nos propos en les illustrant par des tâches d'annotation linguistiques. Ces exemples permettront de contextualiser les difficultés abordées et de mettre en exergue que la solution dépend de l'environnement ou de la tâche.

Bien que nous ayons choisi de présenter les étapes d'une campagne d'annotation d'une manière linéaire, il est nécessaire de rappeler que certaines étapes (notamment l'écriture du guide d'annotation, l'annotation du corpus et les calculs des accords inter-annotateurs) sont à réaliser de manière cyclique, du moins au début de la campagne. La figure 1.1 expose les différentes étapes d'une campagne d'annotation :

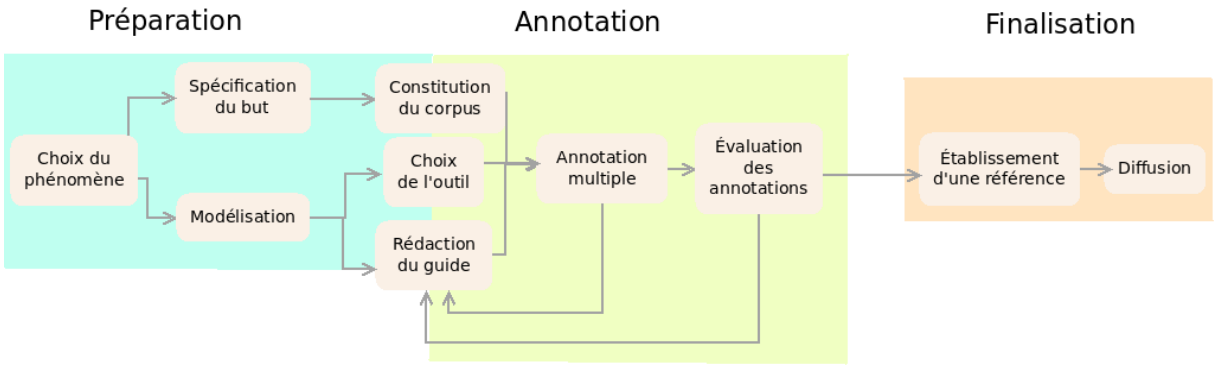
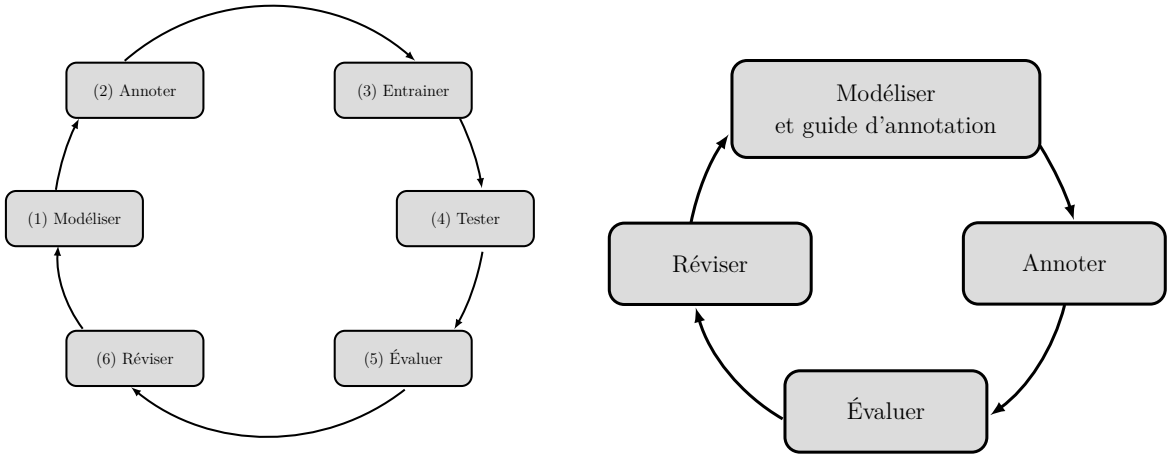


Figure 1.1 – Étapes d'une campagne d'annotation manuelle.

Nous y distinguons trois phases principales, à savoir :

1. La **préparation** : il s'agit de définir et spécifier les objectifs de la campagne et la modélisation choisie du phénomène.
2. L'**annotation** : cette phase est le cœur d'une campagne, mais aussi la plus délicate, à cause des contretemps possibles et des problèmes qui peuvent se produire à ce moment.
3. La **finalisation** : la dernière partie d'une campagne comprend l'établissement d'une référence (si cela est possible), ainsi que la diffusion du corpus et des documents liés à la campagne.

Le fait de procéder par cycle d'étapes rappelle notamment la méthode agile en gestion de projet (Beck, 2011) : ce procédé conseille notamment de travailler par itérations et avec une grande part d'adaptation au fil des étapes. Il semble que Bonneau-Maynard et al. (2005) aient été les premiers à employer cette méthode pour leur campagne d'annotation [1], en recommandant de réviser le guide d'annotation selon les retours des annotateurs, ainsi que de calculer de l'accord inter-annotateur au plus tôt pour contrôler la qualité des annotations. Ce sont Voormann et Gut (2008) qui ont formalisé et répandu le concept d'annotation agile en introduisant trois principes, dont le premier met en exergue le fait de procéder par itération plutôt que par séquence. Plus tard, Pustejovsky et Stubbs (2012) ont aussi défini les cycles MATTER et MAMA, dans une optique d'annotation pour de l'apprentissage machine (voire les figures 1.2).



(a) Cycle MATTER (b) Cycle MAMA

Figure 1.2 – Cycles MATTER et MAMA repris de Pustejovsky et Stubbs (2012).

1. Selon Fort (2012).

1.1 Appréhender le phénomène à annoter

1.1.1 Perception de l'objet étudié

Avant toute chose, les responsables de campagne se doivent de définir clairement et explicitement le but recherché de la campagne : est-ce pour étudier un phénomène linguistique ? pour en proposer un corpus de référence ? pour évaluer un système ? ou encore pour entraîner un système ? Cette spécification de l'objectif permet de fixer le cadre et entraîne certaines décisions. Il est important de bien avoir conscience de l'objectif final pour éviter des changements radicaux qui rendent caducs des choix déjà effectués et qui nécessitent de ré-annoter.

Une des premières étapes d'une campagne d'annotation est de se positionner par rapport à un modèle linguistique — même si Leech (1997) souligne qu'un schéma d'annotation doit idéalement être théoriquement neutre. Selon le modèle, l'objet étudié ne sera pas perçu de la même manière, et cela impacte de manière profonde l'annotation. Cet impact se ressent autant sur le type que sur la tâche d'annotation, et se répercutera ensuite fortement sur les besoins et le choix de l'outil. Ainsi, si nous prenons comme exemple la tâche d'annotation de la coréférence, les unités à repérer et à catégoriser ne seront pas les mêmes selon le modèle choisi. Cette diversité des modèles est bien illustrée dans l'état de l'art réalisé par (Ogrodniczuk et al., 2014, Chap. 3) et présentant les corpus produits pour cette tâche.

Cette décision peut néanmoins vite devenir un problème insoluble, car il est souvent difficile de s'abstraire totalement de toute théorie, voire impossible selon le phénomène. Par exemple, si pour l'annotation des entités nommées, la manière d'appréhender le phénomène est relativement stable selon les campagnes (malgré des divergences au niveau de la caractérisation), il n'en est pas de même pour l'annotation des relations rhétoriques, qui dépend de la perception de l'organisation du discours et des théories qui l'entourent.

1.1.2 Modélisation du phénomène : préparer une première version du guide d'annotation

La modélisation du phénomène dépend, en premier lieu, du modèle adopté et des spécificités du phénomène. Il convient alors d'adapter la tâche, les consignes d'annotations

et, le cas échéant, les catégories pour rendre compte du phénomène le plus précisément possible.

Bien que certains choix découlent de décisions ultérieures ou des fonctionnalités de l'outil d'annotation, nous évoquons dans ce paragraphe certains points que le rédacteur du guide doit prendre en considération, et cela dès la première version du guide. Ici, nous distinguons deux types de recommandations : celles ayant trait au fond de la tâche, et celles sur la manière d'écrire le guide.

Une des premières questions touche à la forme que l'annotation et la tâche devraient prendre ; autrement dit : que faisons-nous annoter aux annotateurs ? Nous distinguons ici ce qui a trait, d'une part, à l'**ancrage** des éléments à annoter, d'une autre à la **caractérisation** de ces objets. Pour l'ancrage des objets, nous proposons la classification suivante :

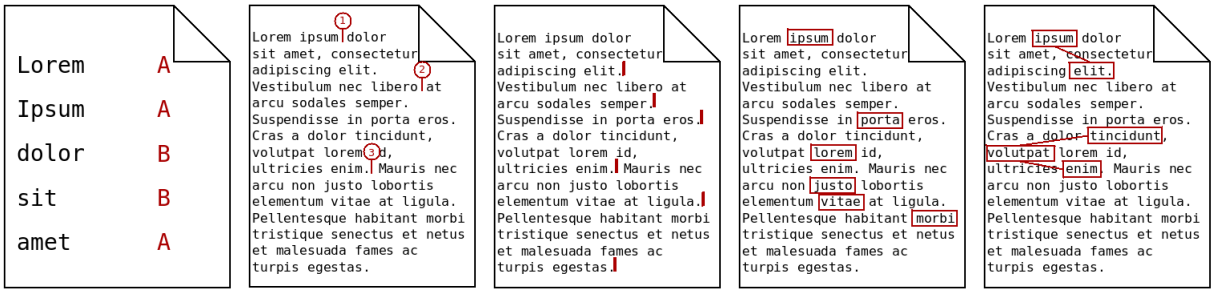
Unités déjà définies : Les unités sont déjà délimitées, l'annotateur doit alors les caractériser.

Ancrage minimal : L'annotateur marque une position dans le flux textuel.

Segmentation : L'annotateur doit paver le texte.

Unitizing : L'annotateur doit repérer dans le texte les unités ; il peut aussi caractériser ces mêmes unités.

Mise en relation : L'annotateur doit relier deux (ou plus) unités entre elles.



(a) Unités déjà définies (b) Position minimale (c) Segmentation (d) Unitizing (e) Mise en relation

FIGURE 1.3 – Illustration des différents types d'ancrage.

En plus de cette classification d'ancrage, le responsable de campagne doit aussi considérer d'autres paramètres d'ancrage : la taille (ou les tailles) des unités à repérer. Sont-elles

de longueur fixe (token, phrase, paragraphe…) ou variable ? Les chevauchements et les structures enchâssées sont-ils autorisés ?

À la question de l'ancrage des unités s'ajoute une réflexion concernant la caractérisation de ces objets : quelles catégories sont les plus pertinentes pour la campagne ? Le jeu d'étiquette peut être fermé, c'est-à-dire avoir un ensemble de catégories déjà défini en amont et qui ne changera pas, ou ouvert, et qui pourra évoluer au fil des annotations. Le type de catégories est aussi un paramètre auquel il faut réfléchir : catégories binaires (présence ou non du phénomène, positif ou négatif, etc.), nominales (spectre de catégories plus large), scalaires (échelle de valeur), etc.

Le schéma peut contenir uniquement des catégories, mais aussi :

— Des sous-catégories : les responsables de campagne peuvent décider de détailler des catégories parfois générales, en rajoutant des sous-catégories. Selon l'application, il est possible de revenir à des catégories plus générales, mais l'inverse est impossible. Le schéma fin Quæro de Grouin et al. (2011) utilise ce type de jeu d'étiquettes.
— Des attributs, ou traits : les attributs et leurs valeurs peuvent être assignés à plusieurs catégories, en plus et parfois indépendamment de cette catégorisation. Cette solution peut être privilégiée pour éviter une explosion du nombre de catégories, si le nombre de catégories et d'attributs est important. Ils sont souvent utilisés en étiquetage morpho-syntaxique, comme proposé dans pour le schéma IL-POSTS (Sankaran et al., 2008).
— Une hiérarchie (Chiril et al., 2020).

Parfois, certains responsables prévoient une catégorie pour les items incertains ou ambigus.

Toutefois, des limites techniques peuvent aussi se poser et entraver la modélisation informatique du phénomène. D'une part, il faut être conscient que tous les outils d'annotation ne possèdent pas toutes les fonctionnalités. D'autre part, certaines réalités linguistiques peuvent être difficilement représentables.

D'un point de vue rédactionnel, Fort et al. (2009) recommandent notamment :

— de définir précisément les termes, les catégories et justifier les choix effectués ;
— d'ajouter des exemples ;
— d'intégrer les potentielles ambiguïtés ;
— de préciser l'objectif de la campagne ;
— de laisser une part d'interprétation pour les annotateurs.

Ces recommandations sont à appliquer avec mesure : donner trop d'exemples peut être

préjudiciable. Certains points seront discutés dans le chapitre 5.7.

Le guide d'annotation ne sera pas parfait dès la première version, et il doit évoluer au fil de la campagne. Il est au départ écrit à partir d'une vision du phénomène ou de la tâche d'annotation qui peut être biaisée ou ne pas rendre compte de l'entièreté du phénomène et des difficultés. L'amélioration du guide ne peut se faire qu'en travaillant conjointement avec les annotateurs (voir en 1.4), qui sont au cœur du processus.

1.1.3 Prendre la complexité en compte

Une question sous-jacente lorsque nous étudions l'annotation de phénomènes linguistiques concerne la définition de la complexité de ces derniers. Il convient alors de distinguer les deux facettes de la complexité :

— la complexité intrinsèque de l'objet étudié ;
— la complexité inhérente de la procédure d'annotation.

Certains chercheurs se sont déjà intéressés à la complexité intrinsèque de la tâche, notamment Gut et Bayerl (2004). Ces derniers estiment que nous pouvons juger la difficulté d'une tâche selon l'accord inter-annotateur obtenu[2] : moins le score d'accord est élevé, plus cette tâche sera considérée comme difficile. Ce procédé est appliqué, par exemple, durant le processus d'annotation du corpus CLISTER (Hiebel et al., 2022) : en effet, les responsables de la campagne ont d'abord fait annoter un échantillon de phrases, sans guide. Ils ont ensuite calculé l'α de Krippendorff et ont obtenu un faible accord (0,239) : la tâche a alors été jugée difficile.

Les autrices de Fort, Nazarenko et al. (2012) ont, quant à elles, proposé d'analyser une tâche d'annotation selon six dimensions, en quantifiant ces dimensions pour les représenter sur un diagramme radar, comme celui de la figure 1.4.

Ainsi, cela permet de mieux saisir les complexités des tâches d'annotation, de même que les différences qui existent. Si une tâche peut se décomposer en plusieurs sous-tâches — ou tâches élémentaires —, alors les autrices conseillent d'afficher sur un même graphique les caractéristiques des sous-tâches, tout en augmentant l'échelle du graphique proportionnellement aux nombres de sous-tâches.

2. La présentation des mesures d'accord inter-annotateurs sera effectuée dans la partie 1.5.

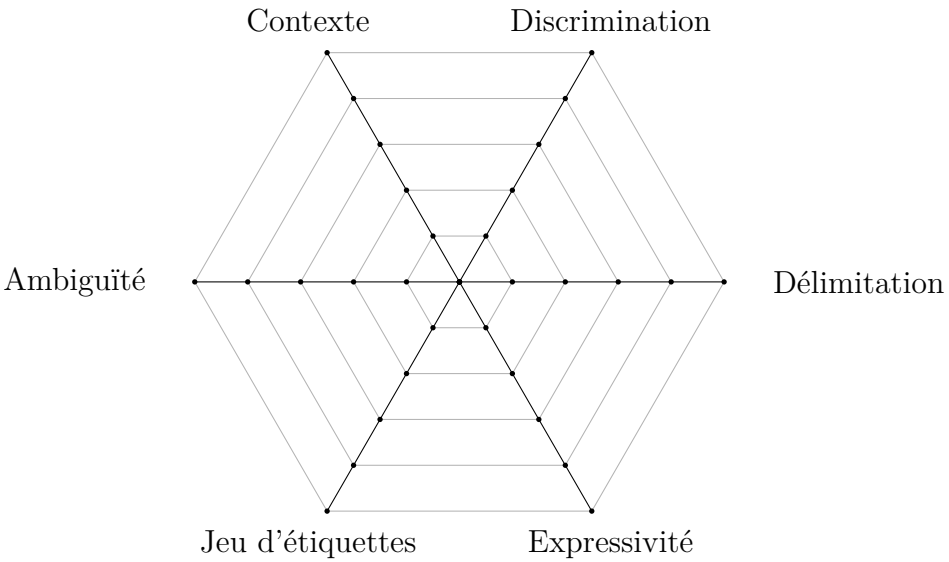


FIGURE 1.4 – Synthèse des six dimensions de complexité définies par FORT, NAZARENKO et al. (2012) (version française du diagramme reprise de la thèse de FORT (2012)).

Outre la complexité intrinsèque de la tâche, il convient aussi de s'interroger sur la manière d'aborder la complexité technique de l'annotation. Il faut pouvoir en effet rendre compte de la réalité du phénomène, tout en gardant une simplicité d'annotation. Ce point sera développé dans la partie 1.3.

1.2 Constituer le corpus de textes à annoter

Le corpus qui sera à annoter est un des points majeurs de la campagne. La construction du corpus se révèle donc une étape primordiale et délicate, car de sa bonne constitution dépendront les annotations et la représentativité de ces dernières. Pour mieux appréhender cette notion de corpus, nous reprenons la définition de ce terme proposée par SINCLAIR (2005) et associée au domaine de la linguistique de corpus : [3]

> A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or

3. Une traduction est proposée par MILLOUR (2020) :

> Un corpus est une collection de textes sous forme électronique sélectionnés selon des critères externes dans l'optique de représenter, autant que possible, une langue ou une variété de langue, et utilisée comme source de donnée pour la recherche en linguistique.

language variety as a source of data for linguistic research.

Les textes du corpus sélectionnés répondent donc à deux besoins principaux. La première exigence est de constituer un corpus représentatif d'une langue tout en gardant une certaine homogénéité dans sa sélection. Le second besoin est de correspondre à la tâche : dans le cadre d'une campagne d'annotation, le phénomène à annoter doit apparaître dans les textes.

Cette notion de représentativité est à nuancer selon la discipline ou le but applicatif dans lesquels la campagne d'annotation s'inscrit. En effet, si l'objectif est d'étudier les caractéristiques du phénomène, il est primordial que le phénomène se rencontre en quantité suffisante. À l'inverse, si le but est par exemple principalement d'étudier la fréquence d'un phénomène, le fait qu'il se retrouve de manière sporadique dans le corpus n'est pas dommageable.

1.2.1 Sélection des textes

La constitution du corpus à annoter renvoie fondamentalement à la linguistique de corpus, discipline qui s'intéresse aux enjeux et aux choix réalisés durant la constitution du corpus, pour l'application finale. Nous renvoyons le lecteur intéressé par la linguistique de corpus aux ouvrages de McEnery et Hardie (2011) ; Nazarenko et al. (1997), détaillant davantage cette discipline.

Comme nous l'avons signalé en introduction de cette partie, le corpus se doit d'être le plus représentatif d'une langue : il convient d'avoir des textes diversifiés, non limités à un critère (genre, auteur, thématique, etc.). Cette notion est d'autant plus importante s'il s'agit d'un corpus ayant pour vocation d'être *de référence* ou servant de base pour produire un outil (comme un étiqueteur morpho-syntaxique) : il doit pouvoir faire état d'une langue et d'un phénomène et donc proposer des textes variés et nombreux (ces corpus sont souvent de grande taille). Des études comme celles de (McClosky et al., 2006 ; Sharoff, 2006) ont démontré qu'utiliser des ressources entraînées sur un corpus dont les textes sont spécialisés induit des baisses de performances sur d'autres types de textes. Les choix effectués lors de la création du corpus peuvent donc limiter sa possible réutilisation dans le futur. Cela est toutefois à nuancer, dans le cadre d'une campagne d'annotation qui s'intéresse spécifiquement à un domaine d'application autre que linguistique : le corpus peut alors ne refléter que le domaine concerné.

Toutefois, le corpus est censé garder une homogénéité suffisante. Il est plus judicieux de ne considérer que les textes provenant d'un mode de production ou de réception spécifique (par exemple, oral ou écrit), sauf si le but de la tâche est précisément d'étudier les différences entre les modes. D'autres caractéristiques sont à prendre en compte, comme le caractère diachronique ou synchronique du corpus, le cadre, le(s) destinataire(s)… Biber (1993) a notamment proposé une grille de catégorisation externe, permettant de donner une description des documents du corpus.

	Critères	**Exemples**
1	Canal	Écrit, oral, discours écrit
2	Format	Publié (et autres formats), non publié
3	Cadre	Institutionnel, cadre publique, privé, personnel, etc.
4	Destinataire	
	(a) Pluralité	Non compté, pluriel, individuel, soi-même
	(b) Présence (lieu et temps)	Présent, absent
	(c) Interactivité	Non, un peu, beaucoup
	(d) Connaissances partagées	Générales, spécialisées, personnelles
5	Émetteur	
	(a) Démographie	Sexe, âge, profession…
	(b) Reconnu	Personne reconnue, institution
6	Factuel	Informations factuelles, indéterminé, fictif
7	Objectifs	Persuader, amuser, informer, expliquer, raconter, décrire, enregistrer, se confier, exprimer des émotions ou des opinions…
8	Domaine	

Table 1.1 – Grille de catégorisation externe : liste hiérarchique des paramètres situationnels. Adaptation de l'anglais depuis Biber (1993).

1.2.2 Sources possibles des textes et corpus déjà disponibles

Les sources de corpus sont variées, ainsi que leur état, et chacune amène son lot de difficultés dont il faut être conscient. Dans cette partie, nous considérons les corpus sous format électronique, qu'ils soient structurés ou en texte brut.

Avec l'émergence et la démocratisation d'Internet, le Web représente une source de données extrêmement importante, par la richesse et la quantité des ressources disponibles

en ligne. Valette (2016) évoque les nouveaux genres d'internet, des textes nativement écrits pour être en ligne. Il s'agit surtout des mails, des messages sur les forums, des blogs, et des microblogs (*tweets*). L'utilisation de telles sources peut poser problème au niveau de la propriété intellectuelle, et le responsable de campagne doit pouvoir justifier de leur réutilisation. D'autres types de données, pour lesquelles la pratique et l'usage se sont transformés, peuvent aussi être cités : c'est notamment le cas des journaux, qui proposent une version en ligne. Nous pouvons citer le site Europresse[4] qui regroupe plusieurs journaux français et internationaux, et qui facilite l'extraction des articles pour un usage numérique. Là encore, il faut être vigilant quant aux licences associées à ces textes.

Il peut être difficile de rassembler des textes, *a fortiori* sous licence réutilisable, pour certaines langues, et surtout pour des langues peu dotées. En ce sens, quelques initiatives, comme l'encyclopédie libre Wikipédia[5], permettent aux usagers de créer du contenu tout en autorisant leur réutilisation. La taille importante et la disponibilité dans plusieurs langues des contenus mis à disposition sont les raisons du succès de Wikipédia comme corpus. Le projet Oscar (Ortiz Suárez et al., 2019) met à disposition, sous licence CC0, des corpus pour 166 langues, initialement issus du projet CommonCrawl[6].

D'autres textes, initialement diffusés sous la forme matérielle d'un livre ou d'un autre support peuvent être aussi disponibles en version numérique par le biais du fichier numérique, d'un éditeur par exemple, et peuvent donc constituer des corpus non bruités. C'est le but du Projet Gutenberg[7], qui vise en effet à proposer des versions électroniques de plusieurs livres et d'éditions. Ce sont essentiellement des œuvres littéraires anciennes : des romans, des pièces de théâtre ou de la poésie. Leur large utilisation est motivée par la facilité à les trouver sur Internet due à leur arrivée dans le domaine public. Toutefois, le passage au format « texte brut » de ce genre de document n'est pas exempt de problèmes.

Il existe également des documents de nature variée (manuscrits, imprimés anciens…) dont l'accès électronique est en cours de création. Le projet français Huma-Num[8] vise notamment à préserver cet héritage culturel et soutient des projets qui se fixent comme objectif de numériser de tels textes pour que la communauté puisse en profiter. Leur numérisation et leur traitement posent toutefois des difficultés, parce que la reconnaissance

4. Voir http://www.europresse.com/fr/.
5. Voir https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Accueil_principal.
6. https://commoncrawl.org/
7. Voir https://www.gutenberg.org/browse/languages/fr.
8. https://www.huma-num.fr

de caractères n'est pas adaptée à de telles ressources (Gabay et al., 2020 ; Jiang et al., 2021). De plus, la transition d'un format d'un objet matériel à un objet dématérialisé ne peut pas retranscrire toutes les spécificités du matériau de base (papier utilisé, ratures…), spécificités se révélant souvent utiles pour l'annotation.

Enfin, certains sites proposent de référencer des corpus déjà existants, parfois même déjà annotés : Ortolang [9], Corli [10], ou encore les corpus mis à disposition par le Défi Fouille de Texte [11]. Au niveau international, nous pouvons citer les initiatives CLARIN [12], l'*European Language Resources Association* (ELRA) [13] et le *Linguistic Data Consortium* (LDC) [14].

1.3 Choisir l'outil d'annotation

L'outil d'annotation sera au cœur du processus d'annotation, car les annotations seront réalisées via cet environnement. Le choix de l'outil est généralement laissé au responsable de campagne, qui peut préférer utiliser tel outil par souci de simplicité ou parce qu'il permet de rendre compte *techniquement* d'un phénomène. Parfois, les annotateurs peuvent donner leur avis sur le choix de l'outil à utiliser, pour des raisons de facilité d'utilisation. Au fil des années et des campagnes, les outils d'annotation se sont multipliés et l'usage a évolué. Il convient alors de faire un état des lieux des outils existants, ainsi que des critères à considérer pour la sélection de l'outil.

1.3.1 Outils existants

Pour une liste complète d'outils d'annotation, nous nous référons au travail de Neves et Ševa (2019) [15] [16].

9. https://www.ortolang.fr/market/corpora.
10. https://corli.huma-num.fr/inventaire-des-corpus-ecrits/ pour les corpus écrits et http://ircom.huma-num.fr/site/corpus.php pour les corpus oraux et multimodaux.
11. https://deft.limsi.fr/
12. https://www.clarin.eu/content/data
13. http://www.elra.info/en/
14. https://www.ldc.upenn.edu/
15. Les auteurs ont aussi développé un moteur de recherche disponible à l'adresse suivante https://annotationsaurus.herokuapp.com/ et permettant de rechercher un logiciel correspondant à des critères.
16. Bien que des critères concernent spécifiquement le domaine du biomédical, les auteurs ont souhaité dresser une liste la plus exhaustive possible des logiciels d'annotation.

1.3.1.1 Méthodes triviales

Si le responsable d'une campagne est novice dans les pratiques d'une campagne d'annotation ou que l'objet est simple, l'adoption d'outils simples peut parfaitement convenir à certaines tâches.

C'est le cas par exemple des traitements de texte *What you see is what you get* (WYSIWYG), comme OpenOffice Writer ou Microsoft Word, qui sont parfois utilisés (Mpouli Njanga Seh, 2016). Il peut aussi s'agir d'un module complémentaire d'un éditeur de texte, comme cela a été le cas pour la correction des pré-annotations du *Penn Treebank* (Marcus et al., 1993), qui ont développé un *plug-in* pour le logiciel Emacs. Comme nous le verrons dans la section 1.7.1, un des formats d'annotation fréquemment utilisé est XML ; certains responsables de campagne préfèrent donc recourir directement à un éditeur XML, tels que Oxygen ou XMLmind [17].

L'avantage de ces méthodes est la connaissance de l'interface pour les annotateurs. Le coût d'entrée est alors moindre et ils peuvent commencer l'annotation plus rapidement. La gestion d'un schéma d'annotation peut aussi être permise grâce au détournement des styles, dans le cas des traitements de textes WYSIWYG : par exemple, un style particulier correspondrait à une catégorie. Dans le cas de la pure catégorisation, un tableur pourrait aussi suffire.

Toutefois, l'utilisation de ces logiciels n'est pas sans risque. Le gros défaut vient d'un problème d'expressivité : il est en effet difficile de rendre compte de certaines spécificités d'un objet avec une interface simple et détournée de sa fonction première. Cela est notamment le cas lorsqu'il y a des risques de chevauchements ou de structures imbriquées des objets. Exprimer les relations ou des attributs n'est toujours pas possible. Par ailleurs, cette solution ne permet pas d'avoir des annotations déportées — une des préconisations de Leech (1993). Un risque de non exploitabilité demeure aussi, sans beaucoup de post-traitements ; ces-derniers peuvent, en outre, entraîner une perte d'information.

17. Voir https://www.oxygenxml.com/ et http://www.xmlmind.com/products_fr.html.

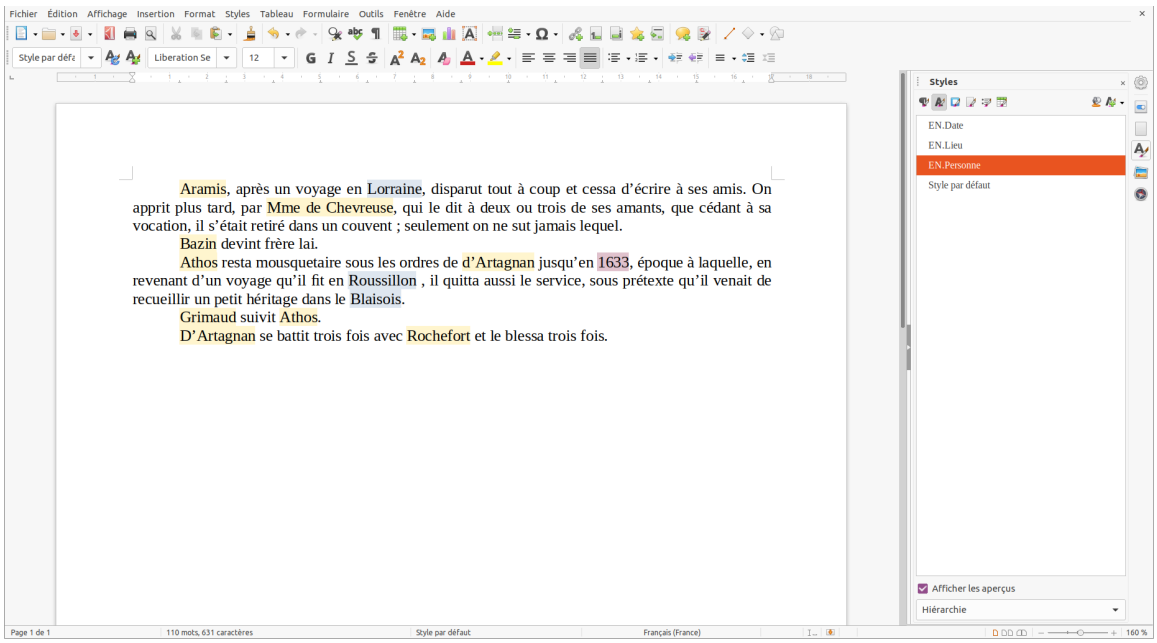


FIGURE 1.5 – Exemple d'annotation via l'interface d'un traitement de textes WYSIWYG. Texte provenant de l'épilogue des « Trois mousquetaires » d'Alexandre Dumas.

Token	Parties du discours
Aramis	NAM
,	PUNCT
après	PREP
un	DET
voyage	NOM
en	PREP
Lorraine	NAM
,	PUNCT

FIGURE 1.6 – Exemple d'étiquetage morpho-syntaxique annoté via un tableur.

1.3.1.2 Logiciels d'annotation (*standalone* ou client lourd)

Les logiciels *standalone* ou client lourds pensés pour l'annotation manuelle représentent une grande majorité des outils utilisés lors des campagnes ; il s'agit d'applications indépendantes, pouvant être installées et utilisées telles quelles, sans logiciels complémentaires. Réfléchis par et pour des annotateurs, leur conception facilite le processus d'annotation. Ils permettent également, lorsque cela est nécessaire, d'imposer aux annotateurs des contraintes. Ils respectent aussi des normes et des standards de formats d'annotation, rendant la diffusion et le traitement des annotations plus aisés. Le développement de tels logiciels implique néanmoins un coût financier et temporel. Si aucun logiciel ne correspond à la tâche, il sera nécessaire de prévoir ces coûts dans la campagne pour la création d'un tel outil.

Ces logiciels sont parfois développés pour répondre à un besoin ou à une campagne spécifique. La question de leur pérennisation se pose alors car, associé à un projet particulier qui s'arrête, leur suivi et leur diffusion ne sont généralement plus assurés. Cela complique la recherche d'un outil déjà existant et qui correspondrait aux critères requis, car il y a alors moins de chance de trouver de tels logiciels, et même dans ce dernier cas, la compétence pour faire fonctionner ou adapter le logiciel n'est plus toujours disponible.

Comme logiciel d'annotation *standalone*, nous pouvons citer Glozz [18] (Widlöcher & Mathet, 2009), Callisto (Day et al., 2004), MMAX2 [19] (Müller, 2006), PALinkA (Orăsan, 2003), Gate [20] (**Cunningham_2022**)…

1.3.1.3 Interfaces Web

Il existe aussi des outils d'annotation que l'on peut installer sur un serveur et rendre accessibles via un site Web. L'interface ressemble alors à un logiciel d'annotation *standalone*, avec les mêmes possibilités de fonctionnalités proposées. Il convient de noter que c'est grâce à l'évolution des technologies du web côté client que nous pouvons désormais mettre en place des interfaces élaborées impossibles auparavant.

Le principal inconvénient repose sur le fait qu'une connexion Internet peut être requise. Néanmoins, ce type d'installation apporte tout de même des avantages, notamment sur

18. http://www.glozz.org/.
19. https://github.com/ottiram/MMAX2.
20. https://gate.ac.uk/

deux aspects :

— la facilité de déploiement et récupération des données : l'annotateur n'a pas à installer directement un logiciel, ce qui évite des soucis d'installation, et, grâce à la gestion de compte, il lui est aisé de reprendre ses annotations s'il travaille sur plusieurs ordinateurs ;
— le travail collaboratif : le gestionnaire d'une campagne a facilement accès à l'ensemble des annotations produites par les différents annotateurs et peut contrôler leur qualité à tout moment.

Les interfaces Web les plus utilisées sont BRAT [21] (STENETORP et al., 2012), WEBANNO [22] (YIMAM et al., 2013) et INCEPTION [23] (KLIE et al., 2018).

FIGURE 1.7 – Exemple de corpus annoté via l'interface BRAT.

21. http://brat.nlplab.org/.
22. https://webanno.github.io/webanno/.
23. https://inception-project.github.io/

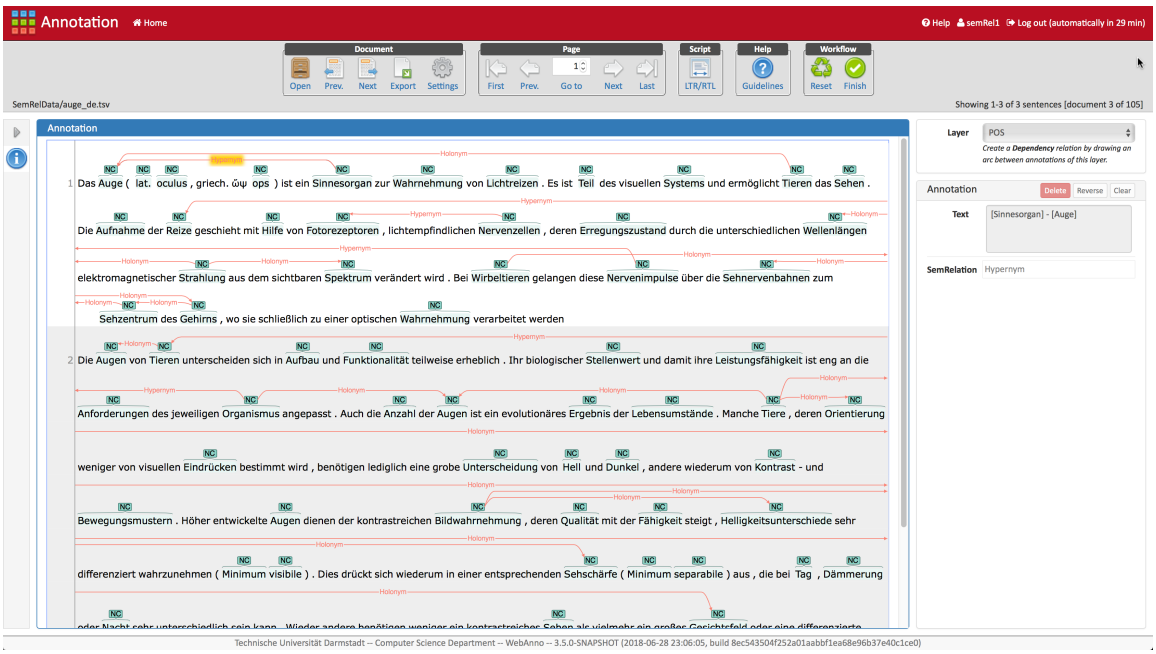


FIGURE 1.8 – Exemple de corpus annoté via l'interface WEBANNO.

1.3.1.4 Plateformes de *crowdsourcing*, ou de myriadisation

Dans certains cas, nous pouvons vouloir que la campagne d'annotation touche un plus large public et l'utilisation d'une plateforme de *crowdsourcing* (**Howe_2006**) — traduit par myriadisation (SAGOT et al., 2011) en français — est alors privilégiée. Plus qu'un site Web dédié exclusivement à l'annotation, ces plateformes proposent à des usagers d'Internet de participer à une tâche d'annotation via une interface appropriée avec l'objectif souhaité (GEIGER et al., 2011). Une telle démarche peut permettre de récolter de nombreuses annotations en peu de temps, bien que leur qualité soit parfois débattue.

Ces plateformes se distinguent des interfaces Web au niveau de l'interactivité avec la communauté d'annotation qu'elles peuvent proposer, avec une campagne qui peut évoluer au fil du temps : rajout de certains items à annoter, de nouvelles fonctionnalités, des récompenses financières, etc. Ceci est particulièrement vrai lorsque ces plateformes prennent la forme d'un jeu ayant un but. La myriadisation permet aussi de toucher un public d'annotateurs potentiellement plus large et plus diversifié.

Nous pouvons citer notamment deux plateformes de myriadisation suivantes pour

l'annotation de phénomènes linguistiques : Amazon Mechanical Turk [24] (Callison-Burch & Dredze, 2010) et LanguageARC [25](Fiumara et al., 2020). Plus spécifiques, nous pouvons aussi citer des plateformes de jeux ayant un but (von Ahn, 2006), comme Phrase Detective [26] (Poesio et al., 2013), JeuxDeMots [27] (Lafourcade & Joubert, 2008) ou encore ZombiLingo [28] (Guillaume et al., 2016).

Figure 1.9 – Interface du jeu JeuxDeMots.

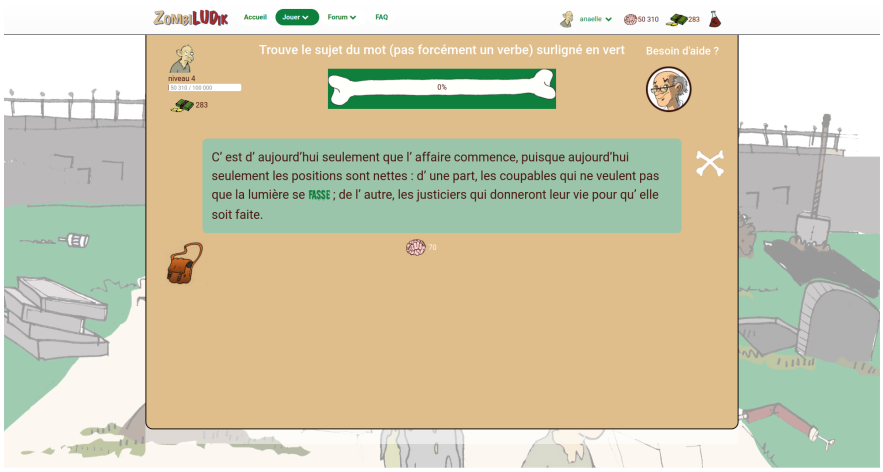


Figure 1.10 – Exemple d'annotation sur le jeu Zombiludik, jeu frère de ZombiLingo.

24. https://www.mturk.com/
25. lhttps://languagearc.com/
26. https://anawiki.essex.ac.uk/phrasedetectives/
27. http://www.jeuxdemots.org/jdm-accueil.php
28. https://zombiludik.org/

1.3.2 Critères à considérer pour le choix de l'outil d'annotation

La décision de privilégier tel ou tel outil constitue un choix important lors d'une campagne d'annotation : plus l'outil sélectionné est adapté à l'objet annoté, plus le processus d'annotation sera efficace, permettra de mieux rendre compte de la réalité du phénomène étudié et limitera les biais. Ce choix repose sur plusieurs critères, touchant à différents aspects, dont nous donnons une liste non exhaustive ci-dessous.

1.3.2.1 Facilité d'installation, d'utilisation et ergonomie

Un premier aspect concerne la facilité d'installation et d'utilisation. Des critères techniques sont notamment à prendre en compte : est-ce que le logiciel est disponible ? Est-il facile d'installation, surtout pour des annotateurs peu familiers avec l'informatique ?

Selon l'outil choisi, les difficultés ne se présenteront pas au même niveau, et aux mêmes personnes. S'il s'agit d'une interface Web, la partie configuration sur un serveur est à gérer, par une personne généralement non impliquée directement dans la campagne d'annotation. Si c'est un logiciel *standalone*, le processus d'installation peut être laissé aux annotateurs. Cela peut décourager les annotateurs non familiers avec l'informatique.

L'interface utilisateur est aussi un élément important, notamment si elle est plus ou moins simple d'utilisation. L'outil doit permettre, facilement, de créer, de modifier et de supprimer des annotations. Des raccourcis ou une intégration de ressources externes (ontologies, dictionnaires, etc.) sont aussi des ajouts appréciables, pour réduire la pénibilité de la tâche d'annotation.

1.3.2.2 Besoins et contraintes de l'objet annoté

Tous les objets annotés n'ont pas les mêmes spécificités, bien que la modélisation de certaines particularités puissent se ressembler. Le logiciel d'annotation doit correspondre à ces spécificités. Par exemple, si l'objet annoté nécessite d'annoter des relations, cette fonctionnalité doit être faisable à partir du logiciel, idéalement sans trop de complications. La majorité des logiciels d'annotation actuels propose toutefois une annotation pour tout type d'ancrage, de la délimitation de segments ou d'unités à la mise en relation, tout en permettant de catégoriser les items repérés.

Une adaptation de la tâche d'annotation est toujours possible, toutefois cela sera au détriment de l'exactitude quant au phénomène étudié ou de la facilité d'utilisation et de traitement. Un ajout au logiciel est, quant à lui, toujours faisable, tant que l'adaptation ne se fait pas au détriment de la facilité d'utilisation.

Néanmoins, il peut être plus pertinent dans certains cas que l'outil ne soit pas trop permissif, et cela pour deux raisons. La première est liée au cas où donner trop de latitude aux annotateurs risque parfois de générer des erreurs d'annotation ou de manipulation de l'outil. La seconde concerne la complexité d'annotation d'un objet. Pour rendre compte au maximum de la réalité d'un phénomène, même des cas les plus rares, nous pouvons vouloir utiliser une procédure d'annotation la plus complète, quitte à en complexifier le processus. Nous pouvons nous demander si une complexité d'annotation est toujours « rentable », ou s'il n'est pas préférable de ne pas pouvoir annoter tous les cas rares afin de faciliter le travail d'annotation.

1.3.2.3 Fonctionnalités propres aux outils

En complément des deux critères listés plus haut, le gestionnaire de la campagne peut aussi vouloir privilégier un outil pour ces fonctionnalités connexes. Ces fonctionnalités touchent à plusieurs aspects. Il n'existe pas d'outil d'annotation universel et possédant toutes les fonctionnalités souhaitées.

Un de ces aspects concerne la prise en charge de ressources. Ces ressources peuvent être internes, comme le schéma et le guide d'annotation, mais aussi des ressources externes, comme l'intégration d'ontologies. Dans les campagnes d'annotation dans le domaine du T.A.L. appliqué au biomédical, par exemple, les annotateurs peuvent avoir besoin d'accéder à une ressource détaillant les maladies.

L'aspect collaboratif peut être un plus pour les gestionnaires de campagne, notamment pour assigner les textes aux annotateurs ou pour contrôler les avancées des annotateurs sur les différents textes.

En plus des fonctionnalités proposées, il convient de vérifier la prise en charge des formats d'annotation. Même si des efforts d'uniformisation ont été entrepris par la communauté au sujet des formats de données, comme l'initiative de la Text Encoding Initiative et d'autres formats de données (qui seront détaillés dans la partie 1.7.1), il

reste encore un travail à apporter concernant une standardisation des outils d'annotation. Nous nous devons toutefois de signaler que des efforts d'interopérabilité entre les différents outils ont déjà été faits.

Néanmoins, les spécificités propres à chaque support d'annotation, à chaque type d'annotation et à chaque phénomène étudié rendent impossible une éventuelle harmonisation de standards à tous les plans. Il convient donc de faire des choix, à différents niveaux (technique, annotation, intégration dans une chaîne de traitement, etc.), et d'avoir conscience des potentiels impacts que ces choix ont sur les annotations recueillies.

1.4 Choisir et accompagner les annotateurs

Les annotateurs constituent le cœur d'une campagne et sont les acteurs les plus essentiels du processus d'annotation. Bien choisir ces annotateurs est donc une étape cruciale. Nous présentons dans cette partie certains points nécessitant l'attention du responsable de campagne pour le « recrutement » des annotateurs.

1.4.1 Quelle expertise ?

Une question fréquente, dont témoignent les réflexions de Péry-Woodley et al. (2011), Stubbs (2012) ou encore Candito et al. (2014), au sujet des annotateurs concerne leur expertise. Même s'il existe plusieurs degrés d'expertise, dans les campagnes d'annotation, nous établissons souvent une opposition binaire entre :

— les annotateurs **experts** : ce sont généralement des personnes ayant suivi une formation dans le domaine (soit linguistique, soit disciplinaire) ;
— les annotateurs **non-experts** (parfois appelés naïfs) : des annotateurs sans formation particulière pour la tâche.

En plus de cette distinction, se pose la question du *type* d'expertise : s'agit-il d'annotateurs déjà au fait du processus d'annotation, ou des annotateurs spécialiste d'un domaine ? À cet effet, Fort (2017) caractérise trois types d'expert : du domaine du corpus (médical, juridique, politique…), du domaine de l'annotation (sémantique, transcription de l'oral, syntaxique…), de la tâche.

Toutefois, une interrogation subsiste quant à la frontière entre les annotateurs experts et « naïfs » : à partir de quand un annotateur est-il considéré comme expert ? Cette appellation est-elle le reflet d'une expérience acquise au fil des années, ou au fil des campagnes d'annotation ? Ou est-elle liée à la confiance placée dans les annotations, souvent davantage accordée pour les annotateurs experts ? Peut-être serait-il plus judicieux de rajouter une catégorie intermédiaire. Cette distinction peut aussi être nuancée selon la tâche. En effet, certaines tâches d'annotation requièrent plutôt un avis, un point de vue, plutôt qu'une expertise, comme cela est le cas pour l'annotation d'opinion ou de sentiment. Pour multiplier les points de vue, faire appel à différents types d'annotateurs (expert ou non, du domaine du corpus ou de l'annotation), semble être une bonne solution.

Dans le cas de campagnes de myriadisation, l'expertise des annotateurs se pose d'autant plus : pour palier le manque d'expert, une foule de non-experts peut-elle suffire pour atteindre la qualité attendue avec des annotations d'experts ? Certains auteurs, comme Chamberlain et al. (2009) ; Snow et al. (2008), ont démontré que les annotations recueillies dans le cadre d'une campagne de myriadisation sont d'une qualité égale à celles obtenues avec des annotateurs experts. Le propos est néanmoins nuancé dans Chamberlain et al. (2013), car les responsables de la plateforme Phrase Detectives indiquent obtenir une qualité d'annotation variable selon les tâches (une différence de 20 points de pourcentage).Cette différence pourrait par exemple s'expliquer par le fait que certaines données, certains objets ne peuvent qu'être perçus par des spécialistes.

En plus de l'expertise, nous pouvons aussi vouloir connaître l'indice de confiance de l'annotateur. En effet, une présupposition courante est de juger plus fiables les annotations des annotateurs s'attribuant un bon indice de confiance. Être au fait de la confiance d'un annotateur sur des items particuliers se révèle aussi un atout non négligeable. Cela est d'autant plus utile lorsque le schéma d'annotation ne prévoit pas une catégorie pour les annotations incertaines. Les annotations en question pourront être alors étudiées avec plus de précision ou traitées avec plus de précaution.

1.4.2 Nombre d'annotateurs

En plus de l'expertise, il se pose aux gestionnaires de campagne la question du nombre d'annotateurs sur un segment donné, et par extension, du nombre total d'annotateurs à engager.

Souvent au cours des campagnes, les gestionnaires attribuent à chaque annotateur un certain nombre de segments du corpus à annoter, qui varient selon les annotateurs, créant ainsi une division du travail. Cette technique a notamment été utilisée lors de la campagne d'Annodis (Péry-Woodley et al., 2011). Le nombre d'annotateurs par segment peut donc être moindre que le nombre total d'annotateurs.

Il convient aussi de définir le nombre d'annotateurs travaillant sur les mêmes parties du corpus. Bayerl et Paul (2011) remarquent une corrélation entre un nombre élevé d'annotateurs et une hausse du désaccord. Toutefois, avoir une cohorte large d'annotateurs permettrait :

— d'accorder une confiance plus élevée aux annotations pour lesquelles un grand nombre d'annotateurs sont d'accord (en se basant sur l'hypothèse discutable que plus il y a d'annotateurs en accord sur une annotation, plus cette annotation est valide) ;
— d'avoir un échantillon élargi des désaccords possibles.

Le nombre d'annotateurs serait aussi à faire varier selon la complexité intrinsèque de la tâche : plus la tâche est complexe, plus il faudrait d'annotateurs. Enfin, dans le cas de la myriadisation, où les annotateurs sont généralement non-experts, Snow et al. (2008) estiment qu'il faudrait davantage d'annotateurs pour compenser le manque d'expertise : ils concluent que le travail de quatre non-experts équivaudrait au travail d'un expert. Toutefois, Passonneau et al. (2012) n'arrivent pas à une conclusion aussi catégorique.

1.4.3 Accompagner les annotateurs

Une fois les annotateurs choisis, le responsable de campagne est tenu de réfléchir à la formation que les annotateurs ont besoin de suivre. Cette phase d'entraînement consiste souvent, pour les annotateurs, à annoter une petite partie du corpus, *via* l'outil d'annotation, et à se familiariser avec l'environnement d'annotation et la documentation (guide d'annotation, manuel du logiciel, etc.). Cette phase sert également aux responsables de campagne de recueillir le retour des annotateurs, notamment pour améliorer le guide, en évaluant l'accord inter-annotateur.

La formation des annotateurs donne lieu à des annotations plus fiables, c'est-à-dire augmente l'accord entre les annotateurs et diminue les chances qu'un annotateur, ne sachant pas quoi répondre, réponde au hasard. Maîtriser l'outil permet aussi une annotation

plus rapide et évite des erreurs de manipulation. De nombreux auteurs (Bayerl & Paul, 2011 ; Bhardwaj et al., 2010 ; Dandapat et al., 2009 ; Fort et al., 2010) ont souligné l'importance de la formation des annotateurs, autant pour les annotateurs experts que naïfs [29].

Les expériences menées par Dandapat et al. (2009) ont montré que les annotateurs formés, que cela soit à la tâche ou à l'outil, annotent plus rapidement. Plus précisément, la formation sous la supervision d'un responsable serait plus efficiente qu'une « auto-formation ». Dans Bhardwaj et al. (2010), les auteurs comparent l'accord inter-annotateurs entre des annotateurs entraînés et des annotateurs de la plateforme de *crowdsourcing* Amazon Mechanical Turk. Bien que l'accord inter-annotateur varie selon les items, il reste plus élevé au sein du groupe des annotateurs entraînés que les annotateurs d'AMT.

1.4.4 Rester à l'écoute des annotateurs

Une fois qu'une première version du guide est réalisée (ainsi qu'éventuellement un manuel de l'outil) et que les annotateurs ont été choisis et éventuellement entraînés, une première phase d'annotation commence. Elle permet notamment aux annotateurs de se familiariser avec la tâche et l'environnement d'annotation, de se former. Dans le même temps, le guide d'annotation va connaître des modifications, selon le retour des annotateurs et la réalité de la tâche, mais aussi de la qualité des annotations produites (qualité contrôlée en mesurant l'accord inter-annotateurs, voir partie suivante 1.5).

Après cette phase, lorsque le guide sera stabilisé, que les annotations seront de qualité suffisante, voire qu'une référence sera disponible sur une petite partie du corpus, l'annotation pourra se faire sur l'ensemble du corpus. Même pendant cette étape, les gestionnaires de campagne doivent rester à l'écoute des annotateurs. En effet, ils pourront être confrontés à des items plus difficiles ou ambigus non traités dans le guide. D'autres remarques peuvent aussi émerger pendant une annotation plus longue que celle de la première phase. Une suggestion plus fine et complète de division du processus d'annotation existe dans la thèse de Fort (2012).

29. À juste titre, Péry-Woodley et al. (2011) s'interrogent si un annotateur naïf l'est encore s'il est formé à la tâche d'annotation.

1.5 Évaluer les annotations

Au cours d'une campagne ou à la fin du processus d'annotation, l'évaluation des annotations produites est un point crucial : cette étape permet en effet de contrôler la qualité des annotations produites et, si cette qualité est suffisante, prépare à l'établissement d'une référence. Durant la campagne, calculer l'accord inter-annotateurs permet notamment d'identifier des problèmes liés au guide d'annotation ou des annotateurs dont les productions se distinguent trop fortement des autres. Un accord inter-annotateur faible peut aussi trouver sa source dans la difficulté de la tâche.

En plus de l'accord inter-annotateurs, nous pouvons calculer l'accord intra-annotateur : au lieu de comparer les annotations de plusieurs annotateurs, nous comparons les annotations d'un seul annotateur sur un même jeu de données à des périodes différentes (par exemple, au début et au milieu du processus). Gut et Bayerl (2004) considèrent que cet accord intra-annotateur se révèle tout aussi nécessaire : cette mesure permettrait de vérifier le reproductibilité des annotations, mais aussi de mettre en lumière des annotateurs dont les annotations manqueraient de consistance, c'est-à-dire de cohérence dans les annotations d'items voisins, par manque d'expérience ou d'implication.

L'accord est surtout utile pour vérifier la **fiabilité** de la tâche d'annotation, pour laquelle Krippendorff (2013) distingue, repris plus tard par Artstein et Poesio (2008), trois types de fiabilité :

- la **stabilité** (*stability*) d'un annotateur, grâce à l'accord intra-annotateur, pour vérifier que sa manière d'annoter est constante ;
- la **reproductibilité** (*replicability*), grâce à l'accord inter-annotateur, si les annotateurs annotent de la même façon en travaillant indépendamment des uns des autres ;
- l'**exactitude** (*accuracy*) : en plus d'observer des sources de désaccord intra- et inter-annotateur, cette méthode permet de mesurer l'écart par rapport à une référence s'il en existe déjà une.

La fiabilité constitue un pré-requis pour la **validation** du schéma d'annotation, c'est-à-dire si les catégories choisies et définies permettent de retranscrire une « vérité » du phénomène. Toutefois, un bon accord inter-annotateur n'est pas forcément gage de validité (Mathet & Widlöcher, 2016).

1.5.1 Mesurer l'accord inter-annotateurs

Cette partie s'appuie sur l'article très complet portant sur les mesures d'accord inter-annotateurs de Artstein et Poesio (2008).

Les mesures d'accord s'appuient sur une première valeur brute, appelée *accord observé* (A_o) : il s'agit du pourcentage d'annotations sur lesquelles les annotateurs sont d'accord. Cet accord observé reste néanmoins simpliste et insuffisant pour évaluer proprement les annotations. En effet, il n'intègre pas la notion de chance, très importante dans les mesures d'accord inter-annotateurs : il peut arriver que des annotateurs soient d'accord sur une annotation de manière fortuite. En plus de l'accord observé, les coefficients prennent donc en compte ce que la communauté appelle l'accord attendu, ou l'*excepted agreement* (A_e), c'est-à-dire l'accord que les annotateurs obtiendraient si tous annotaient au hasard. La formule, présentée en 1.1, est calculé à partir du nombre d'éléments i pour lesquels il y a un accord (agr), divisé par le nombre total d'éléments I.

$$A_o = \frac{1}{|I|} \sum_{i \in I} arg_i \tag{1.1}$$

1.5.1.1 Mesures dédiées à la catégorisation seule

Nous traitons dans un premier temps les mesures d'accord conçues pour comparer les annotations entre deux annotateurs et dédiées à la comparaison de la caractérisation seule. Nous étudions ici trois mesures : le S de Bennett et al. (1954), le R de Finn (1970) et le κ de Cohen (1960). Ces trois coefficients se fondent sur une formule commune présentée dans l'équation 1.1. Leur différence se situe dans la manière de calculer l'accord attendu : ce calcul de l'accord attendu dépend de la manière de prendre en considération la distribution des catégories.

$$S, \pi, \kappa = \frac{A_o - A_e}{1 - A_e} \tag{1.2}$$

S de Bennett Le calcul de l'accord attendu pour ce coefficient se base sur une distribution uniforme des catégories choisies au hasard pour la valeur A_e Ainsi, si le nombre de catégories est élevé, l'accord attendu sera faible et le score général sera important. Scott (1955) critique cela et met en exergue l'effet pervers de rajouter des catégories artificielles

(non-utilisées durant l'annotation) pour gonfler le score.

π de Scott En réponse au calcul de S, Scott (1955) s'appuie sur la distribution de catégories observées parmi les annotations produites.

κ de Cohen Le κ de Cohen (1960) est une des mesures les plus connues et les plus utilisées. Ici, les tirages au sort s'appuient sur les distributions individuelles des deux annotateurs pour calculer A_e.

Néanmoins, comme nous l'avons vu en partie 1.4, il est rare qu'une campagne d'annotation ne fasse intervenir que deux annotateurs. Il convient alors d'adapter les mesures à ce cas de figure. Nous présentons deux coefficients, toujours appropriés pour les annotations catégorielles, et pouvant comparer plusieurs annotateurs.

Multi-π de Fleiss Fleiss (1971) propose une généralisation du π de Scott pour plus de deux annotateurs. Pour ce faire, il se base sur l'accord par paires : il s'agit du pourcentage du nombre de paires d'annotateurs en accord sur le nombre total de paires possibles pour un item. Il reprend la formule 1.2. Ce coefficient est parfois appelé aussi κ, sans qu'il ne possède de lien avec le κ de Cohen.

Multi-κ Davies et Fleiss (1982) suggèrent, quant à eux, une généralisation pour le κ de Cohen. Comme la mesure originale, les distributions des probabilités pour chaque annotateur doivent être calculées.

Il reste une dernière sous-catégorie de mesures d'accord inter-annotateurs dédiées à la catégorisation : celles avec une pondération des catégories. En effet, tous les désaccords entre des catégories n'ont pas forcément la même importance : par exemple, en annotation morpho-syntaxique, assigner la catégorie « Nom propre » à un « Nom commun » peut être jugée comme une erreur moins grave qu'assigner la catégorie « Verbe » à cette même unité.

Les mesures pondérées intègrent une notion de distance, c'est-à-dire donner un coût plus ou moins important aux désaccords entre les différentes catégories, en pénalisant moins sévèrement deux catégories proches, et inversement [30]. Nous pouvons citer en particulier deux mesures pondérées : α et κ_ω, et elles s'appuient toutes deux sur le désaccord (et non l'accord, à l'inverse des précédentes mesures présentées) et leur implémentation de la pondération est la même. Elles se différencient principalement par le calcul du désaccord attendu, ainsi que sur le nombre d'annotateurs à comparer simultanément.

α de Krippendorff Krippendorff (2013) reprend l'hypothèse de π : le calcul s'appuie sur des distributions globales et non individuelles considérant que les annotateurs (qui peuvent être plus de deux) sont interchangeables [31]. Pour calculer α, Krippendorff s'inspire largement du calcul de la variance. De plus, comme nous le verrons dans la partie 1.5.1.3, des adaptations de la mesure sont possibles pour traiter des annotations non exclusivement catégorielles. Cette mesure gère aussi les annotations manquantes (lorsque tous les annotateurs n'ont pas annoté toutes les unités).

κ_ω de Cohen Ce coefficient a été introduit par Cohen (1968). L'auteur s'appuie sur la même hypothèse que le κ initial, c'est-à-dire une distribution des catégories variant selon les annotateurs. Attention, il convient de noter que ce coefficient ne permet de comparer que les productions de deux annotateurs. À notre connaissance, il n'existe pas encore de version d'un κ pondéré pour comparer les productions de plusieurs annotateurs.

1.5.1.2 Mesures intégrant l'*unitizing*

L'*unitizing*, terme proposé par Krippendorff (1995), est un type d'annotation qui regroupe deux tâches imbriquées : la délimitation des items à annoter et la catégorisation de ces derniers. La figure 1.11 illustre les principales problématiques de l'*unitizing*, à savoir : le caractère sporadique des unités, leur positionnement libre, la possibilité d'avoir des unités enchâssées ou encore des unités qui se chevauchent.

L'évaluation de telles annotations dépend de la manière d'appréhender l'*unitizing*. Pour Krippendorff, il s'agit surtout de mesurer la quantité d'intersections d'unités avec

30. Fort (2012) souligne néanmoins que cette approche nécessite d'avoir une présupposition sur la tâche d'annotation, ce qui peut déjà introduire un biais.

31. Si tous les désaccords entre les catégories sont égaux, cette mesure est presque identique au multi-π.

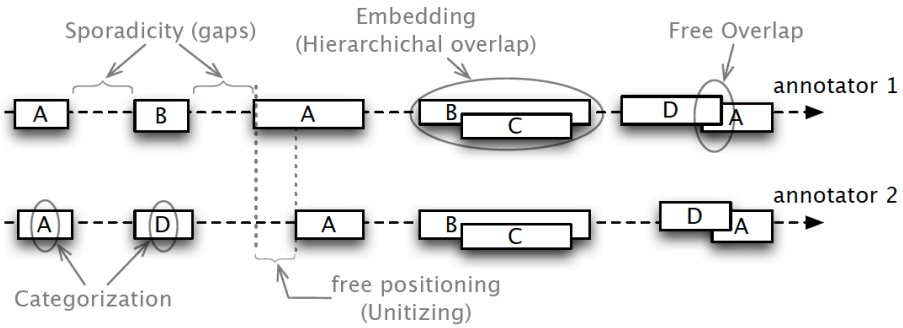


FIGURE 1.11 – Problématiques de l'*unitizing*, figure reprise de MATHET et al. (2015).

la même catégorie. D'autres, comme MATHET et al. (2015), considèrent qu'une phase d'alignement est nécessaire : cela consiste à associer à chaque élément d'un ensemble un élément d'un autre ensemble.

À notre connaissance, il n'y a que deux mesures permettant de comparer les annotations relevant de l'*unitizing* :

${}_u\alpha$ de Krippendorff Dès KRIPPENDORFF (1995), l'auteur a présenté une mesure dédiée à l'*unitizing*, ${}_u\alpha$: l'accord est alors défini par la quantité de chevauchement entre des unités ayant une catégorie identique. Depuis, l'auteur a présenté trois autres mesures d'accord, complétant la première mesure :

- ${}_{|u}\alpha$: ce coefficient prend en compte seulement le chevauchement des unités, sans tenir compte des catégories qui leur sont assignées. Cette mesure est notamment utile pour évaluer uniquement l'alignement des unités entre les différents annotateurs, pour savoir s'ils sont d'accord sur le positionnement des unités ;
- ${}_{cu}\alpha$: cette mesure permet de mesurer le degré d'accord entre les catégories assignées aux unités repérées, tout en intégrant une pondération entre les catégories (la même que α). Toutefois, le calcul ne peut se faire que sur des unités qui se recoupent entre les ensembles des annotateurs ;
- ${}_{(k)u}\alpha$: présenté dans KRIPPENDORFF et al. (2016), ce coefficient mesure la fiabilité de chaque catégorie.

γ de Mathet, Widlöcher et Métivier Cette mesure, dédiée principalement à l'*unitizing*, se veut unifiée et holistique :

- **unifiée :** prend en compte la mesure et l'alignement en même temps ;

— **holistique :** considère les annotations dans leur globalité.

Les auteurs considèrent en effet que l'alignement et la mesure doivent être envisagées en même temps et ne devraient pas être traitées séparément lorsqu'on nous évaluons les annotations manuelles. L'algorithme utilisé, dont les détails se retrouvent dans Mathet et al., 2015, se révèle toutefois d'une complexité algorithmique importante — par comparaison aux autres mesures d'accord —, surtout lorsque les items à annoter sont nombreux. Le calcul de cette mesure est donc assez conséquente en temps et en ressource. Il est à noter qu'il existe des extensions de γ, γ_{cat} et γ_k (Mathet, 2017), se concentrant sur les catégories — il s'agit des pendants respectifs de ${}_{cu}\alpha$ et de ${}_{(k)u}\alpha$ et s'en inspirent sur le principe.

Paun et al. (2022) soulignent toutefois la faible popularité et utilisation de ces mesures dédiées à l'*unitizing* en T.A.L. Ils évoquent trois raisons à cela :

— ces coefficients sont somme toute assez récents ;
— leur calcul n'est pas aisé, bien que des logiciels soient disponibles pour les calculer [32] ;
— la valeur de sortie demeure difficile à interpréter [33].

1.5.1.3 Mesures adaptables pour d'autres types d'annotation

Bien qu'une grande majorité des campagnes d'annotation concerne une catégorisation nominale des unités (prédéfinies ou non), certaines annotations peuvent faire intervenir des catégories plus fines ou dont les frontières sont perméables (comme des échelles de valeurs). Toutefois, souvent par méconnaissance, les responsables de campagne utilisent des métriques propres aux annotations nominales (notamment le κ, très populaire). Cette manière de faire évoque à Mathet et Widlöcher (2016) le fameux « marteau de Maslow » (expression venant de la théorie de l'instrument de Maslow (1966)), qui consiste à considérer tous les problèmes avec une solution unique. Dans le même article, les auteurs soulignent aussi la nécessité d'utiliser les mesures adaptées aux spécificités de l'annotation, pour mieux rendre compte des accords et désaccords des annotations.

Dans une certaine mesure, les coefficients intégrant une pondération peuvent être

32. Voir par exemple https://mathet.users.greyc.fr/agreement/.
33. Même si le problème d'interprétabilité se pose déjà avec les autres mesures d'accord inter-annotateurs.

adaptables à une annotation autre que nominale. Cela est notamment le cas de la famille α de Krippendorff (2013), dont les mesures peuvent être utilisées dans l'évaluation d'annotation en « échelle » : $\alpha_{Ordinal}$, α_{Ratio}, $\alpha_{Interval}$.

Les mesures disponibles se concentrent essentiellement sur la catégorisation des unités, voire des bornes de ces dernières dans le cas de l'*unitizing*. Par exemple, les attributs rajoutés à certaines catégories ne sont initialement pas pris en compte, mais peuvent l'être en transformant les attributs en deux catégories distinctes, en pondérant la distance entre ces catégories.

L'annotation d'une relation entre deux unités, en revanche, n'est pas gérée par les mesures présentées ci-dessus. Notons toutefois qu'il existe des métriques qui comparent une annotation de référence avec une annotation candidate pour évaluer la mise en relation des unités : il s'agit des métriques *Labeled Accuracy Score* (LAS) et *Unlabeled Accuracy Score* (UAS), utilisées en analyse syntaxique.

L'existence de telles métriques donne une alternative aux responsables de campagne cherchant à contourner ce manque de mesures. Néanmoins, ces dernières n'intègrent pas la notion de chance dans leurs calculs, et leur utilisation est donc sujette à précaution.

D'autres décomposent leur processus d'annotation en plusieurs étapes, afin de revenir à des tâches pour lesquelles il existe des mesures adaptées. Nous pouvons citer Lefeuvre, Antoine et Schang (2014) qui décomposent le processus d'annotation en deux étapes. En effet, les annotateurs devaient d'abord repérer et délimiter les entités référentielles, puis relier les différentes mentions entre elles (qui avaient été au préalable vérifiées et avaient donné lieu à un consensus). Les auteurs évaluent ensuite les annotations avec la mesure κ.

1.5.2 Mieux appréhender l'accord inter-annotateurs

Calculer l'accord inter-annotateurs n'est que la première étape pour évaluer les annotations manuelles. Il faut aussi savoir interpréter les valeurs : à partir de quel seuil d'accord les annotations peuvent-elles être considérées comme fiables ? la mesure d'accord dépend-elle de la tâche ? Il convient de noter que les interprétations fournies sont une aide, qui reste subjective selon les auteurs et la tâche.

Nous avons repris en figure 1.12 les différentes échelles de jugements de fiabilité de

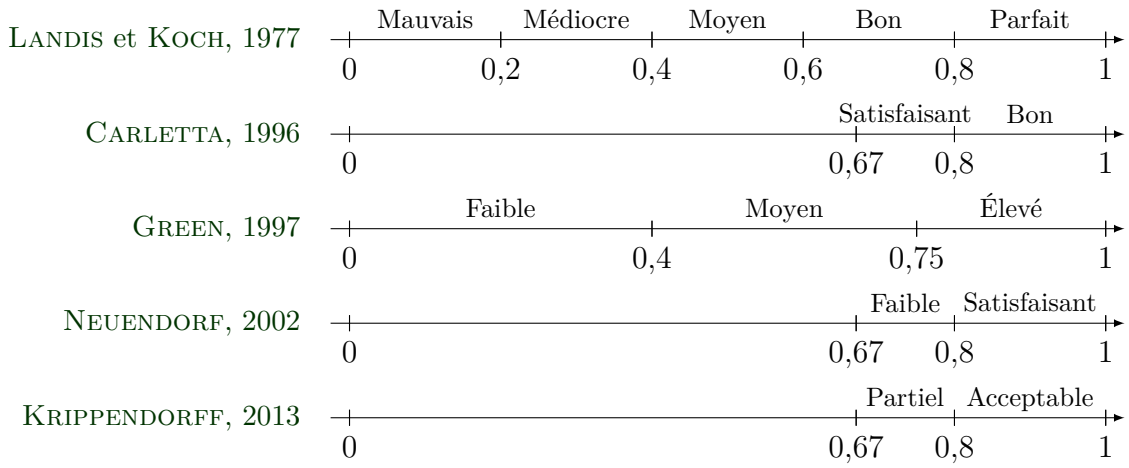


Figure 1.12 – Seuils de fiabilité de l'accord inter-annotateur (pour la mesure κ) (Bregeon et al., 2019 ; Fort, 2022).

l'accord inter-annotateurs. Généralement, au-dessus de 0,8 l'accord est jugé bon, voire parfait pour Landis et Koch (1977). Bregeon et al. (2019) estiment que ce manque de consistance des échelles souligne une « fragilité méthodologique » et, dès lors, il devient difficile d'interpréter l'accord inter-annotateurs. Les différentes échelles proposées peuvent parfois déprendre d'une tâche linguistique particulière ou d'un type de tâche.

Il convient aussi de mettre en perspective l'accord obtenu avec la complexité de la tâche. Pour illustrer cela, nous prenons comme exemple la campagne d'annotation menée et décrite par Gut et Bayerl (2004) : six annotateurs ont participé à une tâche de transcription prosodique, en annotant selon plusieurs niveaux (segmentation en phrases, puis en mots, puis en syllabes, puis en intervalles vocaliques, consonantiques et pauses, avant l'indication des tons et des hauteurs). En utilisant le κ de Cohen, les gestionnaires de la campagne ont calculé les accords inter-annotateurs des paires d'annotateurs pour chaque type d'annotation. Certaines paires d'annotateurs, ayant obtenu un accord proche de 1 pour l'annotation des mots et des intervalles, ne parviennent qu'à des accords inférieurs à 0,4 pour l'annotation des syllabes et des tons. Cette expérience montre bien toute la difficulté — voire l'impossibilité — de produire une échelle d'interprétation de l'accord inter-annotateurs qui se voudrait universelle, tant l'accord dépend de la tâche.

Bayerl et Paul (2011) rapportent aussi des accords inter-annotateurs plus élevés sur des tâches liées à la prosodie que sur les tâches d'analyse sémantique (notamment la désambiguïsation du sens des mots). Amidei et al. (2018), quant à eux, enquêtent sur les

raisons des accords faibles sur la qualité de la génération automatique de texte.

Enfin, il reste encore un travail à fournir quant à l'intelligibilité et l'interprétation des mesures d'accord inter-annotateurs : la correspondance est souvent floue entre le résultat obtenu et ce que ce résultat signifie véritablement, en terme de qualités de l'annotation. Mathet et al. (2012) ont initié un tel travail, pour mieux comprendre les principes et les failles des différentes mesures d'accord. Dans cette publication, les auteurs ont mené plusieurs expérimentations qui consistaient à « dégrader » des annotations de référence, selon des degrés variables et des manières différentes, et observer le comportement des mesures d'accord (notamment à quel point leur score était affecté). Plus récemment, nous pouvons citer les travaux de Bregeon et al. (2019), qui étudient la corrélation d'un bon niveau d'accord avec un bon niveau de stabilité (reproductibilité).

In fine, les mesures des accords inter- et intra-annotateurs demeurent des outils permettant de détecter des problèmes, qu'ils soient liés à la tâche (guide d'annotation, complexité) ou aux annotateurs (annotations cohérentes et fiables, expérience reproductibles). La variété et la complexité des tâches sont aussi des facteurs impactant l'accord inter-annotateurs et la manière dont nous devons utiliser la mesure. En effet, certaines tâches demandent des mesures plus appropriées selon leurs spécificités, comme l'*unitizing* et la délimitation des unités, ou lorsque les frontières entre les catégories sont perméables. De plus, les échelles présentées en 1.12 sont à adapter suivant la complexité ou le degré d'interprétation ou d'ambiguïté de la tâche. Le risque pris à ne pas utiliser des mesures adaptées demeure en l'occultation des désaccords d'annotation.

1.6 Établir une référence

1.6.1 Méthodes pour établir une référence

Une fois l'accord inter-annotateurs jugé satisfaisant, nous pouvons procéder à l'établissement d'une référence. Cette étape nécessite, pour chaque item annoté, d'examiner et de comparer les différentes annotations réalisées, afin de proposer idéalement une seule annotation finale. Conserver la multiplicité des points comme référence peut néanmoins être envisagée. Il existe différentes méthodes pour établir une référence, que nous listons ci-dessous ; ces méthodes concernent d'abord la catégorisation.

Vote à l'unanimité (consensus) : Pour cette méthode, seuls les items pour lesquels les annotateurs sont en parfait accord sont gardés.

Vote à la majorité relative : Pour chaque item, l'annotation de référence est définie par la catégorie ayant eu le plus de vote. La majorité est variable selon le nombre de catégories et le nombre d'annotateurs, et il n'y a parfois pas de majorité.

Révision collégiale par adjudication : Les annotateurs et un référent (généralement un expert) se réunissent afin de discuter des annotations produites, et particulièrement des items dont les annotations sont fortement en désaccord. Cette méthode a été utilisée dans des campagnes telles que Annodis (Péry-Woodley et al., 2011) ou TCOF-POS (Benzitoun et al., 2012).

Ces méthodes conviennent surtout pour une catégorisation seule, mais dès que nous quittons le terrain de la catégorisation, comme dans le cadre de l'*unitizing* ou des relations, cela devient difficile de les utiliser telle quelle. Il est en effet difficile d'utiliser un vote à l'unanimité ou à la majorité relative lorsqu'il s'agit de statuer, en plus de la catégorie, sur les bornes des éléments — à moins de faire une « moyenne » des bornes des unités. La seule solution qui demeure envisageable est la révision collégiale.

1.6.2 Problèmes liés à ces méthodes

Le principe du vote à l'unanimité est problématique : seuls des items faciles (c'est-à-dire les items pour lesquels il ne semble pas y avoir de difficultés pour l'annotation) sont gardés, tandis que ceux dont l'annotation est plus délicate sont supprimés du corpus de référence final. Certes, la référence ainsi acquise peut garantir une meilleure qualité, néanmoins exclure les cas problématiques est préjudiciable pour la suite de la chaîne de traitement. D'un point de vue linguistique, il s'agit souvent des cas les plus intéressants.

Cela pose aussi problème pour un but applicatif. Par exemple, entraîner un système sur un corpus constitué avec une telle méthode ne permettrait pas au système de catégoriser, voire simplement de repérer, des occurrences du phénomène. La mesure de validité est aussi artificiellement forte lorsque nous excluons les cas difficiles, les systèmes étant alors surévalués par la mesure de validité.

Pour le vote à majorité, la majorité n'est pas toujours atteinte ou une égalité peut survenir. Une manière de procéder est alors d'accorder des « poids » différents à certaines

annotations, selon le degré d'expertise ou de confiance des annotateurs. Plus récemment, avec la forte croissance de la myriadisation, les responsables utilisent des méthodes d'agrégation (Dawid & Skene, 1979), qui reposent très largement sur le vote à majorité, même pour les modèles probabilistes les plus élaborés (Paun et al., 2022 ; Zhou et al., 2012).

Nous rejoignons (Mathet & Widlöcher, 2016) en estimant que la révision collégiale semble être la meilleure solution pour établir une référence. Elle reste néanmoins coûteuse, et peut amener un effet de conformisme : les annotateurs peuvent vouloir, inconsciemment, se ranger à la majorité ou se fier à l'avis du référent. Dans tous les cas, il convient de garder une trace des items problématiques.

1.7 Diffuser le corpus

La diffusion est la dernière étape, et concerne aussi bien le corpus que ce qui devrait l'accompagner. Même si la communauté T.A.L. s'est largement inquiétée de la diffusion, certains responsables la méconnaissent, voire la négligent. Elle doit toutefois être pensée en avance, dès la conception. Elle constitue aussi une phase essentielle pour la reproductibilité et l'amélioration *a posteriori*, ainsi que pour des campagnes futures.

1.7.1 Formats des annotations

Bien qu'abordé dans cette partie, le format des annotation est aussi étroitement lié à l'outil utilisé qu'à la diffusion. Nous avons cependant préféré détailler la question des formats ici, car la large diffusion et la pérennité des sources en dépendent : si le format est facilement utilisable ou répandu, le corpus pourra être repris aisément. Il existe différents principes d'annotation, dont quelques-uns sont listés ci-dessous. Plusieurs formats peuvent implémenter l'un ou l'autre de ces principes :

Annotation insérée (*inline*) ou déportée (*stand-off*) (Fort, 2012) :

Insérée : Les annotations produites sont directement incluses dans les textes sources, permettant par la même occasion de modifier les textes bruts si les annotateurs y détectent une erreur.

Déportée : Les annotations sont présentes dans un fichier séparé.

Dans une solution hybride, le positionnement des unités pourrait être réalisé au contact du texte (*inline*) et la caractérisation serait pour sa part totalement ou partiellement séparée (*stand-off*). La caractérisation pourrait être éventuellement plus riche qu'une simple catégorisation, avec des structures de traits par exemple qui sont assez communes en T.A.L. Si nous ajoutons des relations au modèle, se posera aussi la question de savoir si elles doivent être avec le texte ou avec les caractérisations, les deux solutions ayant leurs avantages.

EXEMPLE 1.1 : Annotation directement insérée dans le texte, repris de ANNODIS (PÉRY-WOODLEY et al., 2011).

```
<structure>
<context type="before">...  en plus dépendantes de l'aide et de l'investissement étrangers. </context>
<CT NbCar="913" startofs="11600" para="1" list="0" heading="-1" id="geop_28CT_coder2_1280479728969" file="geop_28">
<firstCOREF>L'IDE (investissement direct étranger)</firstCOREF>
<tags><tag nature="Im_COREFdef" start="11600">L'IDE (investissement direct étranger)</tag><tag nature="Ia_COREFpro" start="11715">il
</tag><tag nature="Im_COREFdef" start="11830">l'IDE</tag><tag nature="Im_COREFdef" start="11886">l'IDE par habitant</tag><tag nature=
"Ia_COREFdem" start="12001">ce ratio</tag><tag nature="Im_COREFdef" start="12127">l'IDE</tag><tag nature="Ia_COREFdef_R" start="12151
">l'IDE</tag><tag nature="Im_COREFind" start="12273">des IDE investis</tag><tag nature="Ia_COREFdem" start="12329">Ce chiffre</tag></
tags>
<segment schema="CT_coder2_1280479728969" start="11600" end="12513"><fullVersion><cue type="Im_COREFdef">L'IDE (investissement
direct étranger)</cue> est nécessaire pour mettre en valeur les ressources de la région. Pourtant, <cue type="Ia_COREFpro">il</cue>
reste encore très faible. Parmi les pays en transition, l'Asie centrale est le parent pauvre du point de vue de <cue type="
Im_COREFdef">l'IDE</cue>. La BERD a calculé, sur la période 1989-1999, que <cue type="Im_COREFdef">l'IDE par habitant</cue> avait
été de 668 dollars pour les pays d'Europe centrale et orientale. Pour les pays de la CEI, <cue type="Ia_COREFdem">ce ratio</cue>
était près de cinq fois inférieur, s'élevant à 140 dollars. Si on excepte le Kazakhstan qui a attiré près de 80 % de <cue type="
Im_COREFdef">l'IDE</cue> en Asie centrale, <cue type="Ia_COREFdef_R">l'IDE</cue> est inférieur à 50 dollars par habitant. Malgré les
hydrocarbures et les métaux, l'Asie centrale n'a reçu que 0,3 % <cue type="Im_COREFind">des IDE investis</cue> dans le monde sur la
période 1998-2000 <cue type="Ia_COREFdem">Ce chiffre</cue> était nul dix ans plus tôt mais seuls les pays en développement du
Pacifique sud ont attiré moins de capitaux que les pays d'Asie centrale sur cette période de trois années.<break type="paragraph"/></
fullVersion>
<shortVersion>L'IDE (investissement direct étranger) est nécessaire ...</shortVersion>
</segment>

</CT>
<context type="after"> L'investissement est faible. Les pays de la région ont ainsi ...</context>
</structure>
```

EXEMPLE 1.2 : Annotation déportée. Exemple repris de ANNODIS PÉRY-WOODLEY et al. (2011).

Texte brut :

Amélioration de la sécurité Le maire a invité les membres du conseil à élaborer le programme d'amélioration de la voirie communale et de la sécurité routière pour l'année 1999. Il a rappelé que plusieurs automobilistes ont quitté la chaussée à l'intersection de la RD192 et du chemin rural de la Vaux des Fossés et qu'il convient de modifier le régime de priorité à cet endroit. La pose d'un panneau stop paraît être la formule la mieux adaptée pour assurer la sécurité des usagers. En délibérant, l'assemblée a accepté la proposition du maire et l'a chargé de faire établir par les services de la DDE un dossier de demande de subvention dans le cadre de la répartition des amendes de police 1999.

Annotation déportée :

```
<unit id="gold_1">
<metadata>
<author>gold</author>
<creation-date>1</creation-date>
</metadata>
<characterisation>
<type>UDE</type>
<featureSet>
<feature name="type">UDE</feature>
</featureSet>
</characterisation>
<positioning>
<start>
<singlePosition index="0"/>
</start>
<end>
<singlePosition index="27"/>
</end>
</positioning></unit>

<unit id="gold_30">
<metadata>
<author>gold</author>
<creation-date>30</creation-date>
</metadata>
<characterisation>
<type>UDE</type>
<featureSet>
<feature name="type">UDE</feature>
</featureSet>
</characterisation>
<positioning>
<start>
<singlePosition index="30"/>
</start>
<end>
<singlePosition index="70"/>
</end>
</positioning></unit>
```

Exemple 1.3 : Format hybride inline–stand-off. Phrase et annotations reprises de Sequoia Candito et al., 2017.

Positionnement *inline*

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xmlns:txm="
http://textometrie.org/1.0">
    <s>À peu près au même moment que <entite id="1">
    Gutenberg</entite> inventait l'imprimerie, <entite
    id="2">Gillet Bonnemire</entite> créait en 1450 la
    première forge à <entite id="3">Saint-Dizier</
    entite>, à l'actuel emplacement du CHS. </s>
</TEI>
```

Caractérisation *stand-off*

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xmlns:txm="
http://textometrie.org/1.0">
    <annotations>
        <annotation id="1">EN.Personne</annotation>
        <annotation id="2">EN.Personne</annotation>
        <annotation id="3">EN.Lieu</annotation>
    </annotations>
</TEI>
```

Linéaire : Les annotations sont incorporées au texte, séparées des unités par un symbole délimiteur. C'est le format utilisé par exemple pour le Brown Corpus (Kucera & Francis, 1967) et le Penn Treebank (Marcus et al., 1993).

Balisé : Les annotations sont présentes dans des balises, encadrant les unités, et autorisant des structures hiérarchiques. Lors des campagnes d'annotation en Humanités Numérique, le format utilisé est en général une extension de XML.

Exemple 1.4 : Format linéaire

Scarlett_NAM joue_VER:pres avec_PRP le_DET:ART chat_NOM ._PUNCT

Exemple 1.5 : Format balisé

<w cat="NAM">Scarlett</w> <w cat="VER">joue</w> <w cat="PRP">avec</w> <w cat="DET">le</w> <w cat="NOM">chat</w> <w cat="PUNCT">.</w>

Format horizontal ou vertical (Leech, 1997) :

Format horizontal : Les annotations reprennent le format en ligne du texte.

Format vertical : Ce format se présente sous la forme d'une unité (souvent un token) par ligne, suivie par sa catégorie, voire aussi par des attributs et d'autres informations (s'il fait partie d'un *chunck* ou non). Les formats CoNLL-X (Buchholz & Marsi, 2006) et sa version révisée CoNLL-U, ainsi que *Inside-Outside-Beginning* (ou IOB) sont des exemples de format vertical. Pour Leech (1997), ce format

permet des étiquettes plus verbeuses ; toutefois il ne permet pas d'intégrer des structures enchâssées.

Exemple 1.6 : Format horizontal

Scarlett/NAM joue/VER:pres avec/PRP le/DET:ART chat/NOM ./PUNCT

Exemple 1.7 : Format vertical

#	**Forme**	**Lemme**	**Catégorie**
1	Scarlett	Scarlett	NAM
2	joue	jouer	VER:pres
3	avec	avec	PRP
4	le	le	DET:ART
5	chat	chat	NOM
6	.	.	PUNCT

Certains chercheurs ont proposé des formats pour « lisser » les usages et rendre plus facilement interopérables les annotations selon les corpus et les traitements ultérieurs. C'est notamment le but de l'initiative d'encodage de textes (Text Encoding Initiative en anglais, ou TEI). Le format TEI est un format balisé (à la manière de SGML, puis de XML), et propose une grande variété dans ses applications. Ce format est surtout utilisé dans des projets en Humanités Numériques. Pour une application plus linguistique, Ide et Suderman (2007) ont présenté l'extension XML GrAF, qui se veut indépendante de la tâche linguistique ; toutefois, à notre connaissance, cette extension a été peu reprise au cours de ces dernières années.

1.7.2 Mettre à disposition le corpus

La mise à disposition peut s'effectuer via des plateformes dédiées, telles celles déjà citées en 1.2.2 : Corli, Ortolang, Cocoon, LRE Map, Clarin, ELRA ou encore LDC. Cette multiplicité de plateformes indique surtout une grande diversité des corpus disponibles (oral ou écrit, type d'annotation, sujets…).

Certains corpus sont aussi directement accessibles depuis le site du projet ou d'un chercheur qui y est associé. Dans des cas, plus rares, le corpus n'est disponible que sur

demande. Cette pratique pose des problèmes quant à la pérennité de la diffusion, si le site devient indisponible ou si le responsable n'est plus joignable.

Il convient aussi de considérer quels fichiers doivent être mis à disposition, lors de la diffusion d'un corpus. Nous distinguons notamment :

— le **corpus** et les **annotations** : ce sont bien sûr les fichiers les plus importants lors du processus de diffusion du corpus ;
— la **documentation** : il s'agit des documents tels que le guide et le schéma d'annotation, mais aussi tout document concernant les choix effectués durant le processus d'annotation et de l'établissement de la référence ;
— les **références relatives au projet** : indiquer *a minima* la référence à citer lorsque nous utilisons le corpus ; les références ayant aidé au projet peuvent aussi être utiles (outils, modèle linguistique adopté…).

En ce sens, nous pouvons citer la distribution du corpus Ancor[34] de (Muzerelle et al., 2014), qui fait preuve d'une grande exemplarité concernant une bonne mise à disposition d'un corpus, notamment grâce à un rapport récapitulant les informations importantes relatives au corpus et reprenant toute la documentation (Antoine et al., 2014).

1.8 Animer une campagne

Une campagne d'annotation, ce n'est pas seulement les étapes présentées ci-dessus. C'est aussi un projet, avec des contraintes à prendre en compte, connaissant inévitablement des retards et problèmes, comme tout autre type de projet. Nous abordons dans cette section certains de ces problèmes potentiels liés à l'annotation. Notre objectif, ici, n'est pas d'apporter des solutions à ces difficultés, mais plutôt d'attirer l'attention sur certains aspects plus généraux de l'annotation.

34. Voir https://tln.lifat.univ-tours.fr/version-francaise/ressources/ancor-centre/corpus-ancor-centre-corpus-de-francais-parle-annote-en-coreference.

1.8.1 Des projets parfois de (très) longue haleine

L'aspect peut-être le plus important d'une campagne d'annotation est la durée d'un tel projet. Il s'agit en effet d'une entreprise chronophage et prenante, subissant souvent des retards. Si certaines campagnes s'étalent sur quelques mois — ce qui constitue déjà un travail important —, d'autres peuvent se prolonger sur plusieurs années, comme cela a été le cas pour le Penn Treebank (Marcus et al., 1993) ou le Prague Dependency Treebank (Böhmová et al., 2003) (ces deux projets ont duré cinq ans).

Il n'est pas aisé d'estimer précisément la durée de chaque étape d'une campagne : cela dépend de différents paramètres intrinsèques à la campagne (par exemple la taille du corpus à annoter, la complexité de la tâche ou de l'outil), mais aussi des possibles contretemps inévitables (personnes non disponibles durant une période, découverte d'un problème dans les données ou d'un bug sur le logiciel, etc.). S'il s'agit d'un corpus oral nécessitant une transcription, cette première étape nécessite un temps non négligeable : Bazillon (2011) a estimé que transcrire manuellement 10 heures de parole spontanée équivaudrait à 80 heures de travail.

Enfin, les contextes (financier, de recherche, sociétal, etc.) évoluent aussi au cours du temps, et entre les différentes étapes. Ces changements ont parfois des incidences sur le processus d'annotation, sur les bases ou les objectifs du projet, et donc parfois sur les annotations.

1.8.2 L'annotation manuelle, à quel prix ?

Le coût financier d'une campagne d'annotation comprend des dépenses variées : la rémunération des annotateurs (s'il ne s'agit pas de *crowdsourcing*, sauf dans le cas d'Amazon Mechanical Turk), les ingénieurs (d'autant plus si le développement d'un outil est requis), voire aussi l'acquisition du corpus dans certains cas.

La précision de l'estimation financière d'une campagne d'annotation est encore peu répandue. Citons toutefois l'exemple, souvent repris, du Prague Dependency Treebank (Böhmová et al., 2003), dont le coût financier total est estimé à 600 000 $ sur l'ensemble des cinq ans du projet. Plus récemment, Martínez Alonso et al. (2016) ont tenu à estimer le coût de plusieurs projets pour la construction de corpus en syntaxe de dépendance ; pour le corpus Sequoia (Candito & Seddah, 2012), le coût revient à

59 000 €.

Une campagne d'annotation peut revenir très chère si le projet prend du retard ou a besoin de nouvelles annotations. Les dépenses sont aussi plus élevées lorsque la tâche d'annotation est complexe ou inédite, ainsi que s'il s'agit d'une langue peu dotée (Seddah et al., 2020).

1.8.3 Conjuguer éthique, réglementation et recherche

Une autre difficulté, non triviale, d'une campagne d'annotation concerne l'éthique. Cet aspect est parfois délaissé, amenant certains responsables de campagne à prendre des décisions malencontreuses, parfois graves. En ce sens, Couillault et Fort (2013) ont proposé une « Charte Éthique et Big Data », visant notamment à documenter au mieux le processus de création d'un corpus. Plus récemment, **Mitchell_2019** ; E. M. Bender et Friedman (2018) ; Gebru et al. (2021) ont proposé des canevas pour la documentation des corpus développés en T.A.L., pour accroître la transparence des données et leur ré-utilisation.

Depuis le début des années 2010, la communauté du T.A.L. prend de plus en plus conscience de ces problématiques. En témoignent diverses actions, comme :

— des ateliers concernant l'éthique de plus en plus nombreux lors des conférences du domaine (par exemple ACL, TALN, LREC) ;
— des articles avec des prises de position sur certaines problématiques en T.A.L. : E. Bender (2019) ; E. M. Bender et al. (2021) ; Fort et Névéol (2018) ; Hovy et Spruit (2016) ; Lefeuvre et al. (2015).

Nous attirons aussi la vigilance des responsables de campagne et des annotateurs sur la nécessité d'anonymiser les données annotées. La réglementation et l'éthique imposent en effet de retirer les données permettant d'identifier les individus. Dans le cadre des campagnes, en particulier orales ou traitant d'un domaine comme le médical ou le juridique, ces informations peuvent être directes (un nom, une caractéristique rare), ou une combinaison de plusieurs caractéristiques amenant à une identification. La procédure d'anonymisation doit prendre en compte ces entités dénommantes (Eshkol et al., 2014). Des études (de Mazancourt et al., 2015 ; Eshkol et al., 2014 ; Grouin, 2013) se sont déjà intéressées à ces problèmes d'anonymisation, en relevant qu'il était complexe, voire

impossible, d'arriver à une anonymisation complète des données. Ceci rejoint les principes FAIR énoncés par Wilkinson et al. (2016), et qui sont : **F**acile à trouver, **A**ccessible, **I**nteropérable, **R**éutilisable.

Les responsables de campagne sont aussi confrontés à la question de la licence du corpus. D'une part, il y a la licence des textes du corpus à prendre en compte, et elle doit permettre leur réutilisation, et leur diffusion [35]. D'une autre, pour une science ouverte efficiente, il est primordial de distribuer le corpus et les annotations sous une licence permettant une réutilisation aisée.

1.9 Conclusion

Au cours de ce chapitre, nous avons présenté les grandes étapes d'une campagne d'annotation. Notre objectif principal était d'aider les responsables de campagnes à saisir l'ampleur d'une telle démarche, de les accompagner au fil de leur questionnement et dans leur progression. Nous avons mis en lumière certains points noirs, auxquels les responsables doivent particulièrement faire attention. Notamment, la modélisation d'un phénomène linguistique n'est pas exempte de manquements dus à des limites techniques et de neutralité théorique. Les mesures d'accord inter-annotateurs peuvent aussi se relever un sujet épineux, surtout lorsqu'aucune mesure n'est prévue pour le type d'ancrage ou de caractérisation.

Nous avons aussi souhaité souligner qu'une campagne d'annotation est constellée de choix et de décisions à prendre, et qu'aucune campagne ne peut être parfaite. Les responsables doivent prêter une attention toute particulière avant le début de la campagne à la manière d'effectuer la tâche d'annotation, et des conséquences qui impactent le choix d'un outil d'annotation. Durant la campagne, il s'agit aussi de noter toutes les décisions prises et leurs justifications.

35. Les annotations sous format déporté peuvent répondre en partie au problème de données non librement accessibles.

Chapitre 2

Typologie des différents types d'annotation

Sommaire

2.1 Comment aborder une tâche complexe ? 56
2.1.1 Tout annoter simultanément 56
2.1.2 Décomposition en plusieurs tâches d'annotation 57
2.2 Illustration pour chaque type d'ancrage 57
2.2.1 Unités déjà définies 58
2.2.2 Ancrage avec position minimale 62
2.2.3 Segmentation 63
2.2.4 Unitizing 64
2.2.5 Mise en relation 66
2.3 Deux exemples ciblés 68
2.3.1 Entités nommées 69
2.3.2 Coréférence 73
2.3.3 Un manque d'harmonisation 77
2.4 Conclusion 78

Les tâches d'annotation sont diverses et peuvent prendre différents formes selon la manière dont les responsables de campagnes les abordent. Dans ce chapitre, nous souhaitons présenter certaines tâches d'annotation selon le type d'ancrage et de caractérisation qu'elle requièrent.

Nous détaillons dans la deuxième partie des exemples de tâches d'annotation, selon le type principal. Nous nous intéressons plus particulièrement à deux tâches d'annotation retenues pour leur richesse illustrative, du fait de leurs caractéristiques linguistiques et des problématiques soulevées durant leur annotation et leur évaluation. Ces deux tâches sont :

1. le repérage et la catégorisation d'entités nommées ;
2. le repérage et la mise en relation des entités référentielles dans le cadre de la coréférence.

2.1 Comment aborder une tâche complexe ?

Les phénomènes linguistiques peuvent faire intervenir plusieurs couches d'interprétation et être vus comme une composition de plusieurs tâches d'annotation. De là découlent deux manières d'appréhender l'annotation de phénomènes complexes :

— réaliser les phases de l'annotation simultanément ;
— décomposer en plusieurs tâches, ou phases.

2.1.1 Tout annoter simultanément

Dans le premier cas, il s'agit d'annoter le phénomène dans son ensemble et en une seule fois. Pour appuyer cette façon de procéder, nous pouvons avancer que, bien que complexe, l'analyse d'un phénomène est difficilement décomposable : les couches d'annotation sont souvent intriquées et dépendantes l'une de l'autre. Par exemple, dans le cadre de l'*unitizing*, la catégorisation peut dépendre de la segmentation, et inversement (voir exemple ci-dessous). Par ailleurs, lorsque nous repérons des unités, nous avons déjà une présupposition quant à la catégorie que nous lui attribuerons.

Exemple 2.1 : Exemple de catégorisation dépendant de l'*unitizing*

	L'	hôpital	Ambroise	Paré
Identification$_1$			Ambroise	Paré
Catégorisation$_1$			Personne	
Identification$_2$		hôpital	Ambroise	Paré
Catégorisation$_2$			Lieu	

Des campagnes telles que Fort et Claveau (2012) ; Péry-Woodley et al. (2011) ont suivi cette procédure d'annotation. Dans l'idéal, si nous décidons d'annoter le tout

simultanément, la mesure d'accord inter-annotateurs utilisée se doit de prendre en compte toute la tâche (par exemple, la mesure γ de Mathet, Widlöcher et Métivier). Nous pouvons néanmoins annoter plusieurs objets en une même passe, mais les analyser ensuite de façon séparée.

2.1.2 Décomposition en plusieurs tâches d'annotation

Les responsables de campagne peuvent préférer décomposer une tâche complexe en plusieurs tâches. Cette décomposition permet, entre autres, de réduire la charge cognitive de l'annotation. Dans ce cas, il convient de séparer véritablement les phases entre elles et de construire la campagne en conséquence. Cette séparation, pour qu'elle soit efficace, doit prendre la forme de sous-campagnes d'annotation, c'est-à-dire :

1. réaliser la première tâche d'annotation ;
2. évaluer les annotations alors produites ;
3. (étape optionnelle) construire une référence pour cette première tache ;
4. à partir de ces annotations de référence, réaliser la deuxième tâche ;
5. évaluer les annotations de cette deuxième tâche ;
6. (étape optionnelle) établir une référence pour cette deuxième phase ;
7. et ainsi de suite.

En ce sens, nous pouvons notamment citer la campagne du corpus ANCOR (Muzerelle et al., 2014), pour laquelle les responsables ont d'abord tenu à faire délimiter les mentions possibles (tâche d'unitizing), puis, une fois une référence obtenue sur les mentions, à les relier entre elles (tâche de mise en relation).

2.2 Illustration pour chaque type d'ancrage

Pour présenter certaines tâches d'annotation de phénomènes linguistiques, nous reprenons la classification des tâches d'annotation selon l'ancrage évoqué en partie 1.1.2, à savoir : les unités déjà définies, l'ancrage minimal, la segmentation, l'*unitizing* et la mise en relation d'items. Pour chacun de ces ancrages, plusieurs caractérisations des unités sont possibles.

2.2.1 Unités déjà définies

L'ancrage déjà défini correspond au cas où les unités à annoter sont déjà délimitées et auxquelles l'annotateur doit assigner une catégorie. De nombreuses campagnes d'annotation sont basées sur ce modèle, et il y a une pléthore d'exemples de tâches avec différentes caractérisations.

Dans le cas d'annotation binaire, nous pouvons citer des campagnes d'annotation où il s'agit de définir si un item est pertinent par rapport à un sujet :

— courriels pouvant être des spams (Metsis et al., 2006) ;
— textes relatifs aux transports ou non (Paroubek et al., 2018) ;
— phrases qui contient un segment obsolescent (Laignelet, 2009) ;
— etc.

Exemple 2.2 : Catégorisation de segments obsolètes, repris de Laignelet (2009).

Phrase	Catégorie
Aujourd'hui, le PIB par habitant de la France est de 27 600 dollars.	Obsolète
En 2004, le PIB par habitant de la France est de 27 600 euros.	Non obsolète

Selon le modèle utilisé, l'annotation de (micro-)textes en analyse de sentiment peut être binaire, si les seules catégories retenues sont `Positif` et `Négatif` (Hamon et al., 2015). Toutefois, avoir seulement deux catégories pour cette tâche est rare, le jeu d'étiquettes prévoyant généralement deux étiquettes supplémentaires : `Neutre` et `Mixte`.

Une majorité des campagnes d'annotation se réalise avec un schéma d'annotation avec plus de deux catégories. Historiquement, une des premières tâches d'annotation utilisant ce type d'annotation est celle en étiquetage en parties du discours — ou P.O.S. pour l'anglais *parts of speech* : Brown Corpus (Francis et al., 1982), Penn TreeBank (Marcus et al., 1993), TCOF-POS (Benzitoun et al., 2012). Les schémas d'annotation diffèrent selon les langues et le projet, et sont plus ou moins larges et fins, selon les sous-catégories et les attributs retenus : ainsi, le schéma du PTB possède 36 étiquettes, tandis que celui du Brown Corpus en possède 77 ; le nombre d'étiquettes peut même aller jusqu'à 300 pour le corpus Susanne (Sampson, 1995).

Exemple 2.3 : Phrase extraite du Brown Corpus.

She	was	now	enjoying	the	voyage	very	much	.
PPS	BEDZ	RB	VBG	AT	NN	QL	RB	.

Exemple 2.4 : Phrase en français oral annotée dans le cadre du corpus TCOF-POS (Benzitoun et al., 2012).

ouais	FNO	ouais
ben	INT	ben
je	PRO:cls	je
l’	PRO:clo	le
ai	AUX:pres	avoir
eu	VER:pper	avoir
au	PRP:det	au
tel	NOM:trc	téléphone
tout à l’heure	ADV	tout à l’heure
au	PRP:det	au
téléphone	NOM	téléphone
il	PRO:cls	il
m’	PRO:clo	me
a	AUX:pres	avoir
dit	VER:pper	dire
qu’	KON	que
il	PRO:cls	il
devrait	VER:cond	devoir
passer	VER:infi	passer

Comme autre tâche d’annotation en catégorisation, citons aussi l’annotation en actes de dialogues, qui consiste à annoter des tours de parole selon leur fonction illocutoire (Austin, 1975 ; Searle, 1972, 1982). Traditionnellement, les tours sont déjà segmentés et reprennent la segmentation originelle des locuteurs. En langue anglaise, il existe notamment des corpus tels que MapTask (Carletta et al., 1997), Switchboard (Stolcke et al., 2000), ou encore Communicator (Doran et al., 2001). Plus récemment, et en français, nous pouvons citer le corpus DATCHA, produit par Perrotin et al. (2018).

Exemple 2.5 : Dialogue annoté selon le manuel d'annotation de Asher et al. (2017), repris de Perrotin et al. (2018).

OPE	1	TC	Bonjour, je suis _TC1_, que puis-je pour vous ?
PRO	2	C	impossible pendant la lecture d'avancer la lecture
STA	3	C	_NUMTEL_
CLQ	4	TC	Si je comprends bien, le problème concerne la vidéo à la demande ?
STA	5	C	mais aussi l'enregistreur et la tv à la demande
INQ	6	C	pouvez vous m'appeler sur le portable ?
INQ	7	TC	Est ce que vous avez un message d'erreur ?
STA	8	C	non
STA	9	TC	Si vous avez débuté le visionnage, mais que le téléchargement n'est pas terminé, l'avance et le retour rapides sont indisponibles. Vous pouvez uniquement stopper ou reprendre le visionnage au début de votre vidéo.
PRO	10	C	seulement l'enregistreur avait terminé l'enregistrement et au cours de la lecture je n'arrive pas à avancer
CLQ	11	TC	Donc le téléchargement a terminé mais vous n'y arrivez pas à l'avancer ?
STA	12	C	après l'avoir débranché puis rebrancher ça refonctionne merci
INQ	13	TC	Ca fonctionne maintenant ?
STA	14	C	oui
ACK	15	TC	Parfait.
INQ	16	TC	Puis-je faire autre chose pour vous ?
STA	17	C	non merci
CLO	18	TC	Je vous en prie Mr _CLIENT_.
CLO	19	TC	Orange vous remercie de votre confiance. Je vous souhaite une bonne journée.

La désambiguïsation lexicale (tâche qui consiste à assigner aux mots ambigus le sens selon le contexte donné) est aussi une tâche d'annotation avec seulement une catégorisation. Les catégories sont néanmoins relatives à chaque mot. La campagne souvent citée est celle de Passonneau et al. (2012).

Le *zoning* argumentatif, tel que défini par Teufel et al. (1999), en est un autre

exemple : l'annotateur assigne à chaque phrase d'un texte (généralement un discours scientifique, un article par exemple) une catégorie selon sa fonction argumentative (contexte, objectif, structure textuelle…).

Il existe aussi des tâches de catégorisation qui ne font pas intervenir des ensembles d'étiquettes fixées ou nominales. Nous listons ci-dessous certaines caractérisations moins répandues :

Multi-étiquettes Parfois, nous pouvons vouloir assigner plusieurs catégories à une même unité. Cette assignation multiple est parfois utilisée en analyse de sentiments, lorsqu'un item présente les caractéristiques de plusieurs émotions.

Ensemble d'étiquettes dynamique Pour certaines tâches, les catégories peuvent évoluer selon le contexte, selon les items du corpus : cela est notamment le cas de l'annotation d'anaphores et de la deixis du discours. Il s'agit par exemple de la méthode utilisée par Landragin et al. (2017) pour le corpus Democrat [1].

Catégories non nominales Il y a aussi des catégories avec une forte interdépendance. Il s'agit, par exemple, d'annotations sous forme d'échelle, ou annotations scalaires, dont les catégories sont contiguës (les frontières sont perméables). Les exemples de ce type d'annotation sont variés :

— Une extension d'une catégorisation binaire (ou ternaire) d'analyse de sentiment ou d'opinion, avec l'ajout d'une intensité et d'une valence (Abbott et al., 2011 ; Bradley & Lang, 1999 ; Bregeon et al., 2019)

— Des échelles de notation, par exemple la pertinence d'une phrase pour être incluse dans un résumé (Fisas et al., 2016). Bien que non réellement issu d'une campagne d'annotation, nous pouvons aussi citer le corpus PeerRead (Kang et al., 2018), qui reprend les relectures et les notes d'articles soumis dans certaines conférences du domaine du T.A.L..

Catégories hiérarchisées Pour certaines campagnes, les responsables utilisent des schémas d'annotation hiérarchisés : il y a des catégories principales, raffinées avec des sous-catégories. Les campagnes d'annotation de Grouin et al. (2011) et Chiril et al. (2020) ont adopté des schémas hiérarchisés.

1. Notons que cela pose un problème méthodologique, si nous considérons que nous avons autant de catégories possibles que de personnes.

2.2.2 Ancrage avec position minimale

Ce type d'ancrage est peu commun dans les tâches d'annotation linguistique : il consiste à ajouter une remarque à un endroit précis du texte. Cela est comparable, en Sciences Humaines et Sociales, aux éditions critiques et savantes, où les éditeurs cherchent à opposer diverses sources pour un texte, pour une traduction par exemple. Cela peut se manifester par des notes de bas de page, qui peuvent comporter sur l'établissement du texte ou des commentaires savants.

EXEMPLE 2.6 : Édition critique de *Histoire du Grand Comte Roger et de son frère Robert Guiscard* de MALATERRA (2016, cop. 2016).

Certains aspects de l'annotation prosodique requièrent parfois uniquement un ancrage avec une position minimale. La fonction principale de la prosodie est de segmenter le discours, et cela est notamment visible lorsque nous faisons des pauses (plus ou moins longues) à l'oral. Il est donc logique que des campagnes d'annotation s'intéressant au langage parlé s'appuient sur la segmentation d'une transcription pour analyser la prosodie [2]. Les auteurs de BUHMANN et al. (2002) ont réalisé une telle annotation. En plus de la segmentation, les annotateurs devaient aussi distinguer les types de pauses (faibles et fortes), ajoutant donc une phase de catégorisation à la segmentation. Citons aussi les campagnes

2. Dans le cas d'une annotation d'un flux audio, marquer les pauses revient à une segmentation.

menées par Pitrelli et al. (1994) et Syrdal et McGory (2000).

Exemple 2.7 : Annotation prosodique, reprise de Buhmann et al. (2002). || représente les pauses fortes, tandis que | représente les pauses faibles.

he was there || and so was his girl-friend
I can tell you | this was un|be|lievable

2.2.3 Segmentation

La segmentation consiste à « paver » le continuum textuel. La principale source de désaccord apparaît lorsque les annotateurs ne bornent pas les segments aux mêmes endroits.

Il existe peu de tâches impliquant uniquement de la segmentation. Nous pouvons toutefois citer la segmentation thématique, qui consiste à délimiter les passages d'un texte selon les thèmes émergents du texte. Les initiatives TREC (Voorhees & Harman, 1998) et TDT (Wayne, 2000) ont permis d'avoir des corpus segmentés thématiquement. En 2006, les organisateurs du DÉfi Fouille de Texte ont aussi proposé un corpus de segmentation thématique de textes politiques (Azé et al., 2006).

Exemple 2.8 : Segmentation thématique, extrait du corpus DEFT2006. Les couleurs permettent de différencier les différents segments.

Le ministre de l'Agriculture, Christian Bonnet. C'est un élu breton, très actif et qui a été un excellent secrétaire d'état au logement, et qui va prendre en charge, à un moment difficile pour le fonctionnement du marche commun agricole, l'avenir de l'agriculture française. Le ministre du Travail, c'est M. Durafour, maire de Saint-Étienne, c'est-à-dire maire de la plus grande ville ouvrière de France, et qui a pu donc, dans la pratique de la vie municipale, connaître le monde du travail, sa représentation, ses problèmes, et qui établira, j'en suis sûr, les meilleurs échanges de vues possibles avec les travailleurs, leurs représentants et leurs organisations syndicales. Le ministre de l'Industrie est Michel d'Ornano, président du Conseil Régional de Basse-Normandie et qui a, je crois, les qualités d'organisation et d'efficacité nécessaires pour que nous poursuivions le développement de notre industrie, notamment le développement des créations d'emploi nécessaires pour assurer l'activité de la jeunesse française. J'ai tenu, avec le Premier ministre, à ce qu'il y ait, dans cette liste, pourtant restreinte, un ministre du Commerce et de l'Artisanat. Nous aurions pu l'appeler le ministre de l'Entreprise individuelle, nous avons gardé son titre traditionnel, et c'est M. Ansquer, vice-président du groupe UDR, à l'Assemblée nationale, et qui est aussi président du Conseil Régional des Pays de la Loire. À côté de cette liste, il y a deux nouveautés que je voulais vous signaler, deux autres ministres : d'abord une femme, Mme Simone Veil, qui est ministre de la Santé. Madame Simone Veil a été déportée avec sa famille à l'âge de 17 ans, à Ravensbruk ; à son retour en France, elle a fait des études et elle a accédé au poste important de secrétaire général du Conseil supérieur de la magistrature. La voici maintenant chargée de l'ensemble du ministère de la Santé, c'est-à-dire des problèmes considérables de l'administration de la santé publique en France, mais aussi, vous le savez, à propos de la santé, d'un certain nombre de problèmes qui intéressent directement les femmes.

2.2.4 Unitizing

La tâche d'*unitizing* est le fait de repérer des unités dans un flux textuel. Cette tâche s'accompagne souvent d'une catégorisation de la séquence délimitée. Selon le but ou la méthodologie de la campagne, il y a trois manières d'aborder ce genre de tâches, qui dépendent aussi du procédé pour traiter les tâches complexes :

— réaliser une seule de ces deux tâches ;
— repérer et catégoriser en même temps ;
— repérer dans un premier temps, puis catégoriser à partir de la référence obtenue d'après ces premières annotations.

Nous pouvons citer comme illustration de ce type d'ancrage l'annotation des expressions polylexicales. Ces dernières sont « des groupements de mots dont le sens individuel "ne permet pas d'interpréter l[a] combinaison" » (Gross, 1982 ; Pasquer, 2017). Leur annotation nécessite, dans un premier temps, de repérer les expressions polylexicales. Une catégorisation peut alors être réalisée sur les expressions repérées, par exemple pour distinguer les expressions nominales des expressions verbales, voire aussi différencier les types d'expressions polylexicales verbales (Candito et al., 2017).

Exemple 2.9 : Repérage et catégorisation d'expressions polylexicales, extraite du corpus Parseme (Candito et al., 2017).

Les associations du village ont vu leur contribution **revue**$_{\text{VID}}$ **à la baisse**.
Sur la première, des fumées ont été rajoutées par ordinateur sur une photo représentant un quartier de Beyrouth ayant **subi**$_{\text{LVC.full}}$ une **attaque** aérienne.
Par ailleurs, la dissolution des roches superficielles a **entraîné**$_{\text{LVC.cause}}$ la **formation** de lapiazs.

L'annotation des structures énumératives (Afantenos et al., 2012) requiert de repérer les énumérations et d'indiquer des unités amorçant et clôturant les énumérations (bien que ces derniers traits puissent être optionnels). Il y a, d'une certaine façon, deux tâches (voire plus) d'*unitizng* en une. Certaines de ces structures sont aussi imbriquées, comme le montre l'exemple ci-dessous.

EXEMPLE 2.10 : Structures énumératives enchâssées, comprenant chacune une amorce et une clôture. Extrait de PÉRY-WOODLEY et al. (2011).

3. Fondements sociaux du concept en Occident	SE1	Amorce	
3.1 Les principes moraux		Item 1	
3.2 Le point de vue du droit		Item 2	
3.3 Le point de vue médical		Item 3	
3.4 Le point de vue psychologique		Item 4	
[…] C'est une notion assez vague, où l'on peut distinguer deux aspects :		SE2	Amorce
– la maturité sociale, c'est-à-dire la capacité de […]			Item 1
– la maturité sexuelle, ou en d'autres termes la capacité […]			Item 2
Ce qu'on peut en tout cas affirmer sur les deux alinéas précédents, c'est […]			Clôture
3.5 Rapprochements		Clôture	
Les approches explicitées ci-dessus forment l'essentiel des principes qui justifient la manière dont nos sociétés perçoivent la pédophilie			

Un dernier exemple de tâche requérant de l'*unitizing* que nous pouvons citer est la transition thématique. Présentée par LABADIÉ et al. (2012, 2010), cette tâche vise à identifier des unités (dont la taille minimale est la phrase) qui peuvent être de différents types (segment, introduction, conclusion ou transition).

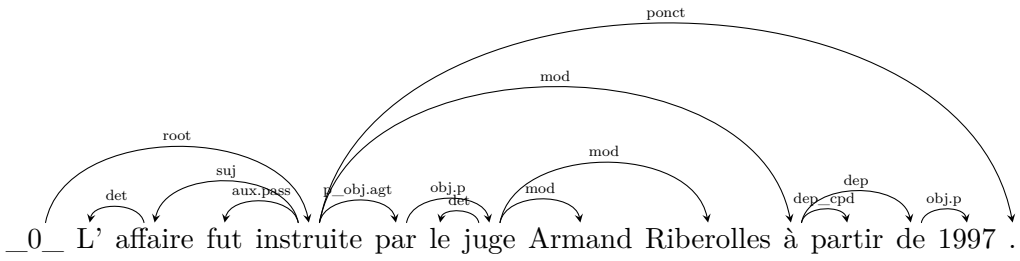
2.2.5 Mise en relation

L'annotation en relation suppose, dans la plupart des cas, de relier des unités entre elles, généralement deux unités, mais parfois plus. Les relations peuvent faire l'objet d'une caractérisation ou non, si la campagne annote plusieurs catégories de relations.

Un des exemples les plus connus de ce type d'annotation est l'annotation en dépendances syntaxiques. Il s'agit de relier une unité avec celles qu'elle contrôle, régit (ou gouverne). Ainsi, une unité peut régir plusieurs autres unités, tout en étant régie par une

autre unité. Les relations sont aussi étiquetées (sujet, objet, déterminant…). Généralement, cette annotation suppose l'annotation en parties du discours. Pour les campagnes d'annotation manuelle, nous pouvons citer les plus connues comme Susanne (Sampson, 1995) et le Penn TreeBank (Taylor & Santorini, 2003) pour l'anglais, le French TreeBank (Abeillé et al., 2019, 2003) et Sequoia (Candito & Seddah, 2012) pour le français. En 2003, Anne Abeillé (Abeillé, 2003) a dirigé un ouvrage regroupant certaines campagnes pour construire ces corpus arborés (ou *treebanks*). Depuis 2016, Nivre et al. (2016) proposent des *treebanks* pour de très nombreuses langues.

Exemple 2.11 : Annotation en syntaxe de dépendance (exemple repris de Sequoia).



Un autre exemple de tâche d'annotation où interviennent des relations est celle des relations rhétoriques. Ici, l'annotation concerne les unités discursives (qui peuvent être des propositions, des phrases, voire des paragraphes, mais aussi des sections et des chapitres entiers) et les relations qu'elles entretiennent entre elles au sein de l'argumentation d'un discours. Par exemple, une unité B explique l'unité A ; une unité B peut aussi être un but de l'unité A [3]. En français, il n'existe, à notre connaissance, que le corpus Annodis (Péry-Woodley et al., 2011) qui a été annoté en relations rhétoriques. Pour la langue anglaise, il existe notamment trois corpus : RST (Carlson et al., 2002), Discor (Baldridge et al., 2007) et Penn Discourse TreeBank (Miltsakaki et al., 2004).

3. Exemples repris des catégories utilisées dans Annodis, disponibles à l'adresse http://redac.univ-tlse2.fr/corpus/annodis/annodis_rr.html.

EXEMPLE 2.12 : Phrase annotée en relations rhétoriques (corpus ANNODIS). EDU correspond à Unités Élémentaires de Discours, et CDU à Unités Complexes de Discours.

[Principes de la sélection naturelle.]_1 [La théorie de la sélection naturelle [telle qu'elle a été initialement décrite par Charles Darwin,]_2 repose sur trois principes :]_3 [1. le principe de variation]_4 [2. le principe d'adaptation]_5 [3. le principe d'hérédité]_6

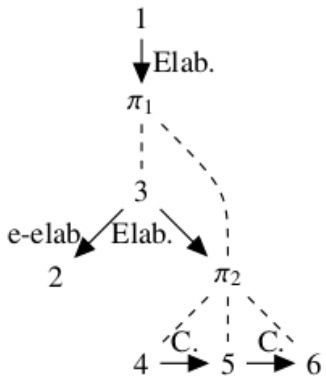


Figure 1. *Exemple de graphe discursif. Les nœuds correspondent aux unités discursives : les EDU représentées par leur numérotation et les CDU par un nœud étiqueté π_n. Les arêtes avec flèches représentent les relations rhétoriques, les arêtes en pointillé sans flèches représentent l'inclusion d'EDU dans un CDU. Elab. = Élaboration, e-elab = Élaboration d'entité, C. = Continuation.*

Un autre exemple de tâche nécessitant une mise en relation d'unités est la coréférence. Cette tâche sera développée dans la section suivante, en raison des autres problématiques qu'elle illustre.

2.3 Deux exemples ciblés

Dans cette partie, nous nous penchons sur deux tâches : la reconnaissance des entités nommées et la résolution de la coréférence. Le but de cette partie est de mettre en exergue deux tâches linguistiques, fort populaires parmi les communautés du T.A.L. et de la L.C. — comme en témoignent les nombreuses conférences et publications dédiées à ces sujets — , aux caractéristiques linguistiques intéressantes par leur richesse et par les problématiques d'annotation qu'elles soulèvent. Cette sélection demeure aussi motivée par la possibilité d'y puiser ensuite prioritairement des exemples. Nous n'escomptons pas donner une vision exhaustive de ces phénomènes linguistiques, mais plutôt mettre l'accent sur ces enjeux d'annotation et des problématiques qu'ils soulèvent.

2.3.1 Entités nommées

2.3.1.1 Définition et applications

Les entités nommées (EN) sont des expressions (mot ou groupe de mots), se référant à une entité du monde. Grishman et Sundheim (1996) ont catégorisé cinq types d'entités, à savoir : personne (`M. Chirac`), organisation (`l'Organisation des Nations Unies`), géographie (`Andorre`), temporel (`le 7 août`) et numérique (20 `€`). Depuis, le concept a été élargi à d'autres types, comme aux évènements (`les attentats du 15 novembre 2015`), aux maladies (`l'Alzheimer`), etc. Ehrmann (2008) a proposé la définition suivante pour les entités nommées :

> Étant donnés un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus.

Dresser une liste exhaustive de toutes les entités semble donc impossible car, selon le corpus, le domaine et l'objectif, les entités nommées possibles sont amenées à changer. Par exemple, à partir d'un même corpus, les entités nommées à repérer peuvent varier selon le but applicatif et le jeu d'étiquettes choisis.

Le repérage et la catégorisation des entités nommées interviennent dans plusieurs cadres d'application, comme la recherche ou l'extraction d'information, le résumé automatique ou encore la construction d'une base de connaissance. Les entités nommées sont aussi utiles lors d'autres traitement linguistiques, notamment la résolution de coréférences, la désambiguïsation lexicale, la traduction automatique…

Le lecteur intéressé pourra se référer à Nouvel et al. (2015) et Ehrmann et Rosset (2018) pour une étude plus en profondeur.

2.3.1.2 Annotation d'entités nommées

L'annotation d'entités nommées comprend deux aspects fondamentaux : la reconnaissance et la catégorisation. À ces deux tâches s'ajoute parfois celle de la désambiguïsation, où il s'agit de préciser et de lier les entités avec une référence unique. L'exemple ci-dessous explicite le processus :

Exemple 2.13 : Exemple d'annotation d'entité nommée

	Montparnasse	se	situe	à	Paris	.
Identification	Montparnasse				Paris	
Catégorisation	Localisation				Localisation	
Désambiguïsation	quartier gare cimetière boulevard				ville (France) département ville (Texas)	

Plusieurs difficultés peuvent survenir lors de l'annotation d'entités nommées. Une majorité de ces difficultés se recoupent avec celles de l'*unitizing* : la catégorisation des unités, le positionnement libre des bornes ou encore l'enchâssement des unités. D'autres sont plus spécifiques aux entités nommées (les exemples proviennent de Ehrmann et Rosset (2018)), qui imposent alors des contraintes notamment sur les environnements et les mesures :

— les discontinuités : une entité nommée peut être divisée/ou séparée en plusieurs segments, comment l'annoter ?
  — `Les Banques centrales américaine et européenne` → `Banque centrale américaine` et `Banque centrale européenne`
  — `Bill et Hillary Clinton` → `Bill Clinton` et `Hillary Clinton`
— l'imbrication : certaines entités nommées peuvent être imbriquées, et selon l'entité nommée en question,
  — `Eiffel` (personne) ne sera pas annoté dans l'entité nommée `Tour Eiffel` (Bâtiment). En revanche, on annotera aussi `Chicago` (Localisation) dans `1:30 p.m. Chicago time` (Heure) [4].
— les frontières : les titres et les fonctions, ainsi que les modifieurs tels que les déterminants, font-ils partie de l'entité nommée, doivent-ils être annotés ?
  — **`la candidate`** `Ségolène Royal`, **`Professeur`** `Paolucci`, **`La`** `Mecque`, **`l'Abbé`** `Pierre`
— la désambiguïsation pas toujours très claire, comment être sûr du liage à la bonne

4. Cet exemple provient de https://cs.nyu.edu/~grishman/NEtask20.book_16.html#HEADING43.

entité ?

— `Jacques Chirac` renvoie-t-il à la personne de Jacques Chirac, ou au président de la République ?

2.3.1.3 Campagnes d'annotation

Plusieurs campagnes d'annotation manuelle ont été réalisées, permettant ainsi la constitution et la publication de corpus variés.

Une liste [5], qui était maintenue jusqu'en 2020, proposait un catalogue (non exhaustif) des corpus annotés en entités nommées, et ce, pour plusieurs langues ; l'arrêt de sa maintenance est dû au nombre croissant de corpus apparaissant chaque année, rendant l'exhaustivité de la liste difficile, voire impossible. Une autre liste est disponible à l'adresse suivante https://damien.nouvels.net/resourcesen/corpora.html, recensant plus d'informations, mais qui ne semble plus à jour depuis 2017.

Les corpus les plus connus sont, pour l'anglais, CoNLL 2003 (Tjong Kim Sang & De Meulder, 2003) et Muc-6 (Grishman & Sundheim, 1996). En langue française, nous pouvons citer les corpus Ester 2 (Galliano et al., 2009), Quaero (Rosset et al., 2012) et Etape (Gravier et al., 2012) [6]. Plus récemment, une nouvelle couche d'annotation, comprenant les entités nommées, a été ajoutée au corpus Sequoia (Candito et al., 2020). Il existe aussi le corpus Fenec (Millour et al., 2022).

5. Disponible à l'adresse https://github.com/juand-r/entity-recognition-datasets

6. Respectivement disponibles aux adresses suivantes : http://catalogue.elra.info/en-us/repository/browse/ELRA-S0338/, http://catalogue.elra.info/en-us/repository/browse/ELRA-W0073/, http://catalogue.elra.info/en-us/repository/browse/ELRA-S0349/ et http://catalogue.elra.info/en-us/repository/browse/ELRA-E0046/.

EXEMPLE 2.14 : Exemples d'entités nommées extraits du corpus Sequoia. Les couleurs correspondent à : NE-PERS, NE-LOC, NE-ORG.

À peu près au même moment que Gutenberg inventait l'imprimerie, Gillet Bonnemire créait en 1450 la première forge à Saint-Dizier, à l'actuel emplacement du CHS.
Ensuite, fut installée une autre forge à la Vacquerie, à l'emplacement aujourd'hui de Cora.
En 1953, les hauts fourneaux et fonderies de Cousances virent le jour, puis Jean Baudesson, maire échevin de Saint-Dizier, autorisé par lettres patentes d'Henri IV, installa à Marnaval- qui signifiait val ou vallée de la Marne ou bien en aval de la Marne-, une forge qui connut son apogée au XIXe siècle.

2.3.1.4 Évaluation

L'évaluation des annotations manuelles des entités nommées pose encore des problèmes. Une solution naïve est d'utiliser le κ de Cohen, en se positionnant au niveau du token et d'avoir une catégorisation binaire `Appartient/N'appartient pas à une entité nommée` (BRANDSEN et al., 2020 ; RUOKOLAINEN et al., 2020). Toutefois, cette solution est biaisée car le nombre de tokens n'appartenant pas à une entité nommée est trop élevé (GROUIN et al., 2011 ; HRIPCSAK & ROTHSCHILD, 2005) [7]. BEJČEK et STRAŇÁK (2010) ont proposé une adaptation du κ, selon des nœuds d'arbres syntaxiques et des pondérations selon les sources de (dés)accords. CANDITO et al. (2020) évoquent la mesure γ, néanmoins cette mesure ne gère pas les annotations discontinues (fréquentes en raison de la coordination). Les auteurs ont finalement préféré une F-mesure, en arguant qu'intuitivement un accord par la chance était relativement faible, en raison du peu de contraintes formelles pour cette tâche. Pour MILLOUR et al. (2022), seul α a été utilisé.

Pour contrer le manque de mesures d'accord inter-annotateurs, certains responsables optent alors pour des métriques prévues initialement pour des évaluations automatiques entre une référence et une annotation candidate. Dans le cadre de l'annotation des entités nommées, il s'agit principalement de F-mesures plus ou moins strictes au niveau des délimitations et des types des entités nommées : certaines demandent de respecter les

7. Cela fait écho au problème de prévalence des catégories, qui sera abordé dans la section 3.1.

bornes, d'autres ne requièrent qu'une correspondance partielle. Chinchor et Sundheim (1993) proposent notamment cinq métriques, reprises plus tard par Segura-Bedmar et al. (2013). Galibert et al. (2011), puis Caubrière et al. (2020), s'appuient sur le *Slot Error Rate* (Makhoul et al., 1999).

2.3.2 Coréférence

Cette partie s'appuie principalement sur les deux premiers chapitres de la thèse de Marine Delaborde (Delaborde, 2020).

2.3.2.1 Définition

La coréférence est un phénomène linguistique où « deux syntagmes nominaux (SN) peuvent être interprétés comme se référant à une même entité dans le monde du discours » (Guéron, 1979) ; Delaborde (2020) ajoute, à cette définition, la mention de « relation symétrique », et propose la schématisation présentée en 2.1.

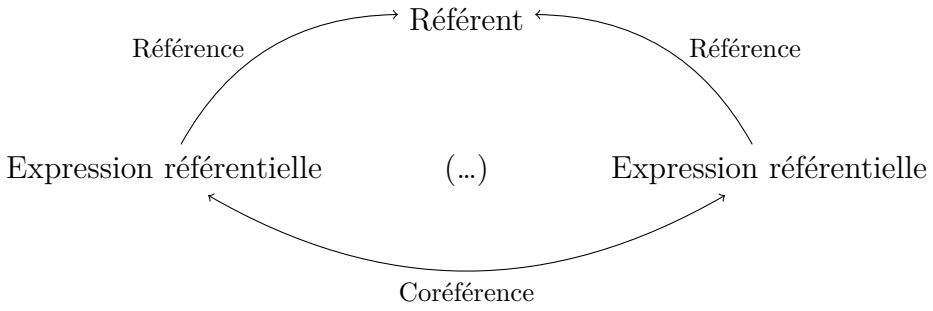


Figure 2.1 – Schéma de la coréférence, repris de Delaborde (2020).

Différentes notions sont associées à la coréférence, définies dans le tableau 2.1.

Exemple 2.15 : Exemple d'une **chaîne de coréférence** et d'un singleton. Extrait de la thèse de Delaborde (2020), et phrase provenant de *Pauline* (Sand, 1881).

La cuisine de l'auberge n'était éclairée que par **une lanterne de fer suspendue au plafond**. Le squelette de **ce luminaire** dessinait une large étoile d'ombre tremblotante sur tout l'intérieur de la pièce, et rejetait sa pâle clarté vers les solives enfumées du plafond.

Notions	Définitions
Référence	Phénomène linguistique permettant de désigner un référent dans le discours.
Expression référentielle	Expression linguistique servant de support à la référence.
Anaphore	Relation asymétrique entre deux expressions référentielles, qui ne sont pas nécessairement coréférentes, dans laquelle l'interprétation de l'une dépend de l'autre.
Chaîne de coréférence	Ensemble des expressions référentielles coréférentes désignant une même entité.

TABLE 2.1 – Notions liées à la coréférence, repris de DELABORDE (2020).

2.3.2.2 Annotation de la coréférence

L'annotation de la coréférence nécessite dans un premier temps d'identifier les expressions référentielles. Il s'agit donc plutôt d'une tâche d'*unitizing*. Les expressions référentielles sont de plusieurs types, notamment des noms (propres et communs), des pronoms et des syntagmes nominaux. La seconde étape de cette tâche d'annotation, la résolution de coréférence, peut prendre différentes formes (LANDRAGIN et al., 2017) : soit l'annotateur construit une chaîne avec les expressions référentielles, en mettant en relation toutes les expressions entre elles, soit il identifie un référent choisi dans un ensemble ouvert, qui devient alors une sorte de catégorie — cette étape devenant alors une tâche de catégorisation.

EXEMPLE 2.16 : Exemple d'annotation en chaîne de coréférence.

Chaîne

C'était une jeune femme d'une beauté vive et saisissante, mais pâlie par la fatigue. Elle refusa l'offre d'une chambre, et, tandis que ses valets préféraient s'enfermer et dormir dans la berline, elle s'assit devant le foyer, sur la chaise classique, ingrat et revêche asile du voyageur résigné.

Autre

C'était une jeune femme d'une beauté vive et saisissante, mais pâlie par la fatigue. Elle refusa l'offre d'une chambre, et, tandis que ses valets préféraient s'enfermer et dormir dans la berline, elle s'assit devant le foyer, sur la chaise classique, ingrat et revêche asile du voyageur résigné.

Laurence

Une question régulièrement soulevée concerne l'annotation des expressions référentielles qui ne possèdent pas de lien de coréférence — nommées alors des singletons. Dans certains cas, les singletons ne sont pas annotés, comme c'est le cas dans le corpus OntoNotes (Pradhan et al., 2011). Ils représentent toutefois la majorité des expressions référentielles (entre 60 et 80 % des mentions dans un texte, selon Recasens et Hovy (2011)).

2.3.2.3 Campagnes d'annotations

Ogrodniczuk et al. (2014, Chap. 3) ont présenté un état de l'art des corpus annotés en coréférences, plus particulièrement pour l'anglais et pour les langues où il y a des sujets implicites (polonais, japonais, arabe et chinois). Le plus connu est OntoNotes (Pradhan et al., 2011). Plus récemment, l'initiative Corbon (Ogrodniczuk & Ng, 2016, 2017) permet de constituer des corpus de coréférence dans différentes langues. Dernièrement, Nedoluzhko et al. (2022) ont présenté le corpus CorefUD, qui propose une annotation standardisée pour la résolution des coréférences et compatible avec les annotations du projet Universal Dependencies.

Pour le français, il existe peu de corpus annotés en coréférence. Nous pouvons toutefois citer les deux corpus Ancor [8] (Muzerelle et al., 2014) — pour l'oral — et Democrat [9] (Landragin, 2021) — pour l'écrit.

2.3.2.4 Évaluation

À notre connaissance, il n'y existe pas de mesure dédiée à l'évaluation des annotations manuelles pour la résolution de coréférence. Durant la campagne Ancor, les responsables ont préféré découper la tâche en deux temps, notamment pour retrouver des tâches pour lesquelles des mesures existaient.

Concernant les métriques comparant une annotation de référence et une annotation candidate, Lion-Bouton et al. (2020) exposent la multiplicité des techniques d'évaluation de la coréférence. Si la majorité s'appuie sur les liens entre les mentions, elles ne traitent pas les relations de la même manière : parfois il s'agit d'une séquence (une men-

8. Voir https://tln.lifat.univ-tours.fr/version-francaise/ressources/ancor-centre/corpus-ancor-centre-corpus-de-francais-parle-annote-en-coreference.

9. Voir https://www.lattice.cnrs.fr/democrat/.

tion est liée à la mention précédente et/ou suivante), d'autres fois toutes les mentions sont liées entre elles.

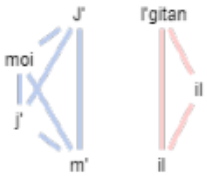
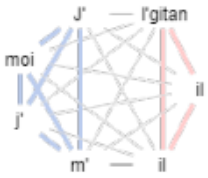
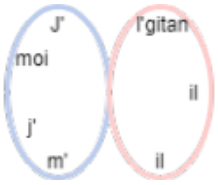
Nom	Séquence de liens (chaîne)	Ensembles des liens (relations)	Liens de coréférences et non coréférences	Ensembles de mentions
Schéma				
Métrique	*MUC* (Vilain et al., 1995)	*LEA* (Moosavi & Strube, 2016)	*BLANC* (Recasens & Hovy, 2011)	B^3 (Bagga & Baldwin, 1998), CEAF (Luo, 2005)

Table 2.2 – Évaluation de la coréférence, repris de Lion-Bouton et al. (2020). Les images sont reprises de cet article.

Cette multiplicité de scores reflète les perceptions divergentes sur la tâche de résolution de la coréférence. Il demeure donc primordial d'être informé quant à ces disparités de point de vue, d'autant plus si les responsables veulent comparer les corpus en utilisant le score adapté à la perception de la coréférence telle qu'annotée. En outre, ces métriques présentent certains problèmes, notamment le fait qu'elles ne prennent en compte que l'identité stricte et que les singletons sont mal gérés — alors qu'ils représentent la majorité des expressions référentielles dans les textes.

Le travail mené par Lion-Bouton et al. (2020) s'inscrit dans une optique d'une recherche réfléchie, s'interrogeant sur la méthodologie et les pratiques lors d'une campagne d'évaluation. Les auteurs s'intéressent notamment à l'interprétation des scores donnés par les métriques pour la résolution des coréférences : ils ont cherché à savoir si ces métriques vérifiaient les propriétés d'une métrique de similarité normalisée, en partant du principe qu'une telle métrique rend « intelligibles les écarts de performance qu'elle peut mesurer ». Leurs expériences ont montré que seule $CEAF_m$ (Luo, 2005) respectait les propriétés.

2.3.3 Un manque d'harmonisation

Les deux tâches présentées ci-dessus sont exposées à des problèmes dus au manque d'harmonisation.

Le premier problème repose sur la diversité des schémas proposés. Il est certes impossible de proposer un schéma harmonisé pour toutes les langues et tous les objectifs d'annotation. Néanmoins il en résulte une difficulté pour comparer les corpus et les outils entraînés sur ces corpus ; Millour et al. (2022) soulèvent ce problème. De plus, si l'on souhaite adapter un jeu d'étiquettes riche pour un jeu plus strict, nous perdons aussi fatalement de la richesse d'annotation.

Le manque de mesures d'accord inter-annotateurs adaptées à ces tâches est un autre obstacle à prendre en compte. Cela rend délicat l'établissement d'une référence, peut-être à cause de l'absence de consensus sur le phénomène. La pratique de séparer les différentes étapes de la tâche d'annotation (segmentation puis catégorisation et/ou relation) contourne ce problème, sans forcément le résoudre. Par exemple, pour la mise en relation d'unités, il reste toujours la question de l'évaluation de l'annotation *manuelle* des relations. Nous pouvons aussi nous demander si le fait de transformer la tâche d'annotation ne transforme pas aussi la perception du phénomène étudié : passer d'une vue macro à une vue micro change-t-il les manières d'annoter chaque étape ? Une question sous-jacente émerge alors : n'est-il pas préférable d'avoir une mesure qui prend en compte l'entièreté du phénomène, plutôt que de le morceler ?

Dans le cas où les responsables des campagnes utilisent des métriques, la multiplicité existante complexifie le processus pour comparer les annotations et les résultats de leur évaluation. En raison des différences entre les modèles dont sont issus les métriques, une métrique peut pénaliser un corpus annoté avec un modèle différent, ou à l'inverse, avantager un corpus annoté spécialement pour présenter cette métrique. La comparaison devient alors artificielle. Parfois, il peut être aussi difficile d'adapter des annotations à des métriques, si l'annotation ne prend pas en considération un type d'ancrage ou de caractérisation.

2.4 Conclusion

Ce chapitre a permis d'exhiber des campagnes d'annotation selon le type d'ancrage et de caractérisation. Même si toutes les combinaisons ancrage–caractérisation n'ont pas pu être représentées, ce panorama permet de mettre en évidence certains éléments importants pour la suite. Notamment, il met en exergue la multiplicité des tâches linguistiques et des campagnes d'annotation réalisées.

Cette vue d'ensemble montre aussi que pour une même tâche, les approches peuvent différer. Les responsables des campagnes peuvent décider de décomposer une tâche pour qu'elle paraisse moins complexe à annoter. Parfois, les différences proviennent des manières d'aborder le phénomène, liées aux théories linguistiques qui régissent le domaine. Toutefois, ce manque d'harmonisation soulève des problèmes de comparaison.

Deuxième partie

Étude des biais et expérimentations

Chapitre 3

Biais d'annotation

Sommaire

3.1 Vers une première classification des biais 82
3.1.1 Classification thématique 83
3.1.2 Classification temporelle 90
3.2 Méthodologie des campagnes d'annotation « Portraits » et « Erreurs » 92
3.2.1 Comment étudier un biais d'annotation ? 92
3.2.2 Nos hypothèses et nos attentes initiales vis-à-vis des expériences 93
3.2.3 Un outil transversal : la consensualité 94
3.3 Conclusion 100

Une campagne d'annotation fait intervenir une chaîne importante et variée d'acteurs qui ont chacun des choix et des contraintes (temporels, financiers, etc.) (Fort, 2012). L'annotation étant un travail généralement collectif et collaboratif, faisant interagir plusieurs aspects — qu'ils soient linguistiques, informatiques, ou autres –, la fiabilité des annotations produites n'est pas garantie par les annotateurs, fussent-ils des experts. Il est donc important de considérer la campagne d'annotation dans son ensemble afin d'identifier et prévenir les éléments pouvant perturber le résultat (Braffort et al., 2011).

Dans ce cadre, nous proposons alors d'introduire la notion de *biais d'annotation* : il s'agit de phénomènes perturbateurs, de nature et d'origine variées, pouvant survenir à toutes les étapes du processus d'annotation et qui sont susceptibles d'en affecter le résultat.

Ces biais peuvent intervenir à toutes les étapes du processus d'annotation, de la phase

de conception de la campagne d'annotation à la finalisation de la campagne et à l'élaboration du corpus de référence — si cela est possible. De nature et d'origine variées, ces biais se trouvent par exemple dans les choix et les contraintes techniques, conscients ou non, implicites ou pas. Leur apparition peut fausser la suite du processus d'annotation et avoir des incidences sur la chaîne de traitement et, par extension, sur les résultats obtenus. Selon la situation, l'impact du biais ne sera pas toujours forcément négatif et peut avoir des effets positifs sur l'annotation ; toutefois, cela peut être contre-productif, se révélant alors d'une fausse qualité. D'un point de vue épistémologique, les biais d'annotation engendrent des problèmes de fidélité et de neutralité quant aux phénomènes étudiés. Si certains biais sont difficilement évitables (par exemple, les problématiques liées à la représentation numérique des données ou les choix imposés par les contraintes logicielles), d'autres résultent d'un jugement ou de choix arbitraires, par facilité ou par commodité.

Dans la suite de ce chapitre, nous classifions thématiquement, en reprenant les étapes d'une campagne, les biais d'annotation que nous avons décelés au cours de notre état de l'art des campagnes d'annotation. Loin d'avoir un but résolument prescriptif, la présentation de cette contribution se veut informative : en avoir conscience permet avant tout « d'intégrer au plus tôt des processus de compensation ou de contrôle » (Brafford et al., 2011). Dans une seconde partie, nous présentons les bases d'une méthodologie pour étudier un biais d'annotation. Cette démarche a été utilisée lors de nos deux campagnes, pour lesquelles nous énonçons, dans cette même partie, nos hypothèses et nos attentes, ainsi que nos moyens d'observation.

3.1 Vers une première classification des biais

Dresser une typologie des biais d'annotation n'est pas chose aisée, car ces phénomènes sont diversifiés et hétérogènes. Certains biais ont tendance à impacter des annotateurs de manière isolée (par exemple, les connaissances connexes d'un annotateur (Amidei et al., 2018)), alors que d'autres biais se répercutent sur un ensemble d'annotateurs (un guide d'annotation mal écrit (Nédellec et al., 2006), par exemple). Les biais d'annotation peuvent déprendre de l'annotateur (par exemple, de son niveau d'entraînement (Dandapat et al., 2009)), mais aussi de l'objet à annoter : chaque tâche a ses spécificités et la combinaison de ces dernières implique une complexité unique (Fort, Nazarenko et al., 2012). Enfin, un biais peut aussi être présent dans l'ensemble d'une campagne,

selon les choix réalisés concernant la représentation numérique des données ou la vision du phénomène adoptée.

3.1.1 Classification thématique

Dans cette partie, nous présentons d'abord une classification des biais de manière thématique, en reprenant les étapes d'une campagne tels que nous avons présentées sur la figure 1.1. Nous sommes néanmoins consciente que cette première typologie peut paraître contestable ou arbitraire : en effet, certains biais sont dépendants de plusieurs étapes, ou ont des influences diverses selon le moment où ils surviennent. Notre but, ici, est avant tout d'édifier une base pour des réflexions futures.

Phénomène

Modélisation du phénomène La perception du phénomène implique déjà un choix. Bien qu'idéalement l'annotation d'un phénomène devrait être théoriquement neutre (comme préconisé par Leech (1993)), il est parfois ardu, voire impossible, de s'affranchir d'un modèle et d'atteindre une neutralité. Il faut alors être conscient que la vision adoptée n'est pas générale ou ne fait pas consensus. Par ailleurs, la modélisation du phénomène ne demeure pas toujours fixe, elle peut évoluer pendant la campagne, selon les contraintes techniques ou le retour des annotateurs. Il en résulte alors que l'annotation peut ne pas rendre compte de l'entièreté du phénomène à annoter. Une modélisation particulière peut aussi indûment conforter ce que pensait le responsable de campagne.

Tâche d'annotation La manière de modéliser le phénomène, ou une partie, pour l'annotation va naturellement influencer la tâche. D'une part, le type de tâche à réaliser ne sera pas forcément le même entre deux campagnes annotant le même phénomène. D'autre part, les schémas d'annotation ne sont pas forcément harmonisés. Citons par exemple l'annotation en analyse de sentiment, dont le schéma peut être `{Positif;Négatif}`, `{Positif;Neutre;Négatif}`, mais encore une échelle de valeur intégrant la valence ($\{-2;-1;0;1;2\}$). Cette grande variabilité de schéma peut impacter l'annotateur qui, habitué à un schéma, garderait ses automatismes pour différencier les catégories, mais aussi le dérouter et l'amener à faire des erreurs d'annotations.

Corpus

Passage d'un format à un autre L'annotation s'effectue parfois sur du texte brut, format qui ne permet pas de rendre compte des mises en pages — telles que les typographies, les alignements de texte ou d'autres éléments (tableaux, figures, etc.). Cette perte d'information peut être préjudiciable à l'annotation, car nous perdons fatalement du contexte qui peut se révéler fort utile. Les documents initialement non numériques, comme les textes anciens, sont aussi touchés par le passage au numérique, la numérisation et l'océrisation ne restituant pas l'intégrité et les particularités (physiques, notamment) de l'œuvre. Cela peut engendrer des erreurs d'annotation.

Absence d'accès au document source En lien avec le biais précédent, pour un aspect plus pratique, l'absence d'un accès au document original peut être avantageux quant à l'annotation de certaines unités, qui peuvent être difficiles à annoter selon l'état du document. C'est surtout le cas pour des campagnes d'annotation où les documents sources ne sont pas nativement sous format électronique (souvent le cas dans le domaine des Humanités Numériques). L'accès au document source peut apporter des informations supplémentaires sur le document à annoter, détérioré à cause d'une mauvaise océrisation ou une erreur de transcription, ou ne reproduisant pas une mise en page particulière, etc.

Corpus non représentatif La sélection des textes du corpus se doit d'être réfléchie si nous voulons généraliser les annotations produites. Le phénomène linguistique étudié ne se rencontrerait pas en même quantité, ni forcément dans les mêmes « conditions » d'apparition selon le type, le domaine ou le genre du texte. Il a aussi été démontré qu'un outil entraîné sur un corpus d'un domaine sera moins performant sur des textes provenant d'autres domaines ou genres littéraires (Amalvy et al., 2022 ; McCLosky et al., 2006 ; Sharoff, 2006).

Outil

Outil non adapté Si l'outil utilisé ne permet pas d'annoter les caractéristiques du phénomène (par exemple l'annotation de relations ou de discontinuités, de même que l'annotation avec des traits), nous pouvons risquer soit de se priver (sciemment) de représenter une réalité du phénomène, soit de perdre en qualité suite au détournement de l'outil. Il convient de noter que l'outil peut ne pas être adapté à

des degrés différents. Se priver de représenter une partie du phénomène est contre-indiqué, et ce quel que soit l'objectif de la campagne d'annotation (produire un corpus de référence ou entraîner un système), car les applications futures seront aussi biaisées pour l'étude de ce phénomène. Quant à détourner l'outil, cela signifie généralement complexifier l'annotation et amplifier les risques d'erreur. Lefeuvre, Antoine, Savary et al. (2014) ont relevé ce problème, lors de leurs travaux portant sur l'annotation de la temporalité, notamment la délimitation des unités discontinues, type d'ancrage non pris en charge initialement par le logiciel.

Difficulté de prise en main La prise en main d'un outil d'annotation, du moins au début de l'annotation, est une étape importante et ne doit pas être négligée. Plus un outil est simple d'utilisation, moins le processus d'annotation est pénible pour les annotateurs (Landragin et al., 2017), ce qui augmente la stabilité des annotations. Par ailleurs, pour les annotateurs moins familiers avec l'environnement informatique, ils peuvent se décourager ou sont plus à même de se tromper. Dans les deux cas, il faut prévoir un manuel d'utilisation et un temps de formation.

L'outil parfait n'existe pas L'outil parfait, qui remplit toutes les conditions (techniques et fonctionnelles) ou qui permet d'annoter tous les phénomènes n'existe pas. Souvent, il y a des compromis à réaliser, parfois au détriment d'une fonctionnalité ou d'un confort d'utilisation. Les conséquences de ces choix peuvent donc impacter de plusieurs manières l'annotation, et les responsables doivent décider en accord avec les annotateurs.

Guide et schéma d'annotation

Caractéristiques du schéma d'annotation Le schéma d'annotation choisi possède certaines caractéristiques qui peuvent influencer les annotateurs. Nous pouvons citer notamment trois exemples :

- Dimension du jeu d'étiquettes : si le jeu d'étiquette est grand, l'annotateur peut éprouver des difficultés soit pour choisir quelle catégorie assigner, soit pour se rappeler toutes les spécificités. Cela peut provoquer des erreurs d'annotation.
- Schéma non hiérarchisé : s'il y a des catégories proches sémantiquement, l'annotateur peut éventuellement avoir des difficultés à avoir une vue d'ensemble claire des catégories ainsi que choisir la catégorie adaptée. Un schéma structuré peut alors aider les annotateurs dans leurs prises de décisions, surtout si les étiquettes sont nombreuses ; l'annotation pourrait aussi s'effectuer en deux

temps : catégorisation des items selon les grandes catégories, puis catégorisation plus fine grâce aux sous-catégories, réduisant potentiellement la charge cognitive et le risque d'erreur.

— Cas spéciaux : le fait de forcer l'annotateur à toujours attribuer une catégorie peut poser problème si l'annotateur hésite ou n'est pas sûr de son annotation. Il peut alors être préférable de prévoir dans le schéma d'annotation une catégorie regroupant des unités « problématiques » (incertaines, ambiguës, ou tout simplement indéfinies) ; une telle catégorie permet à l'annotateur d'indiquer son manque de confiance sans trop alourdir la tâche et la charge cognitive. Nous pouvons aussi imaginer qu'un annotateur assigne, ponctuellement, plusieurs catégories pour un seul objet, s'il hésite.

Qualité rédactionnelle du guide Nédellec et al. (2006) ont déjà montré que la qualité des annotations dépendait au moins en partie du guide d'annotation : un manuel d'annotation doit parvenir à être exhaustif sur le phénomène à annoter (description des catégories, exemples, objectif…), tout en laissant une latitude d'interprétation aux annotateurs. En ce sens, nous rappelons que Fort et al. (2009) ont formulé des recommandations pour la rédaction du guide d'annotation, que nous reprenons dans notre partie 1.1.2. Ces instructions sont en partie contradictoires (être précis dans les définitions tout en laissant une marge d'interprétation), ce qui peut résulter à un guide d'annotation inégal et peu clair pour l'annotateur.

Annotateurs

Niveau d'expertise L'expertise des annotateurs joue un rôle dans la qualité des annotations, comme l'a démontré la méta-analyse de Bayerl et Paul (2011) : les annotations d'experts sont jugées de meilleure qualité et plus fiables.

Niveau et de formation Toutefois, plus que l'expertise, la formation des annotateurs impacte de manière décisive les annotations (Dandapat et al., 2009). Plus les annotateurs sont formés à la tâche, plus leurs annotations sont stables et fiables : leur jugement est moins soumis aux incertitudes et leur temps moyen d'annotation diminue.

Caractéristiques des annotateurs Des caractéristiques des annotateurs influencent directement l'annotation. Amidei et al. (2018) classifient, dans le cadre de l'annotation de la qualité de textes générés automatiquement, les sources de désac-

cord : ils évoquent notamment le style d'écriture de l'annotateur, ses connaissances connexes, ou encore son attention aux détails. Dans un autre ordre d'idée, pour détecter du contenu haineux (sexiste, raciste, homophobe, etc.), est-il préférable de choisir des annotateurs concernés par ces contenus ? En effet, d'un côté ils sont les plus à même de savoir si un contenu est haineux ou non, d'un autre côté, ils sont juges et parties et leur perception peut être biaisée.

Annotation

Pré-annotations du phénomène L'étude de Dandapat et al. (2009) a montré l'effet positif des pré-annotations pour l'annotation manuelle : elles permettent aux annotateurs de gagner du temps et minimisent les gestes à effectuer sur l'outil. Toutefois, mentionnons deux points négatifs à l'utilisation de pré-annotations automatiques, repris des conclusions de l'article de Fort et Sagot (2010) :

— l'utilisation de pré-annotations issues d'un système qu'il faudra ensuite évaluer engendre un problème de neutralité ;

— les annotateurs, lorsqu'ils ne sont pas sûrs ou sont en désaccord avec l'annotation proposée, ont parfois tendance à davantage se fier aux pré-annotations qu'à leur jugement ;

— les pré-annotations doivent être de qualité suffisante, sous peine d'être contre-productives pour les annotateurs qui doivent les corriger.

Subdivision de la tâche Décomposer les tâches complexes permet, notamment, de réduire la charge cognitive liée à l'annotation. Néanmoins, la tâche initiale est déformée et ce glissement de tâche peut modifier les annotations. Si, dans un premier temps, nous demandons à l'annotateur de segmenter les unités, sans chercher à les catégoriser, sa segmentation ne sera pas forcément la même si l'annotateur devait les catégoriser dans la foulée. Finalement, annotons-nous une partie du phénomène ou son ensemble ?

Ordre des items L'annotateur peut être influencé dans ses annotations par l'ordre dans lequel lui sont présentés les items à annoter. En effet, si l'ordre respecte une quelconque logique (par exemple, dans une catégorisation binaire, il y a un passage ne contenant quasiment que des A), l'annotateur peut s'y habituer et avoir tendance à annoter systématiquement les A et ne repérer les B. Ce biais peut apparaître de manière beaucoup plus subtile, par exemple si une catégorie est quasiment systématique en début de phrase, et une autre en fin de phrase.

Distribution des catégories Selon la tâche d'annotation ou le corpus, la distribution des catégories pourra être déséquilibrée, soit globalement, soit ponctuellement. Si ce déséquilibre est trop important, il y a un risque de biais impactant les annotateurs, qui peuvent avoir deux réactions possibles à cela : soit leur attention est « endormie » et ils repèrent moins bien la catégorie rare, soit ils sont davantage attentifs et prompts à annoter une unité de la catégorie rare s'ils ont un doute.

Items proches par leur contenu Si les annotateurs rencontrent deux unités presque identiques au sein du corpus, ils peuvent avoir tendance à tenir compte de l'une pour l'annoter, soit par contraste, soit par rapprochement.

Impossibilité du retour arrière En lien avec le biais précédent, les annotateurs peuvent aussi être influencés par le fait d'avoir accès ou non au travers de l'outil aux précédentes annotations : si l'annotateur rencontre une unité qui ressemble à une autre déjà annotée, il peut vouloir revenir sur cette précédente unité pour vérifier son annotation, soit pour la modifier soit pour annoter cette nouvelle unité ? Si l'accès est plus ou moins aisé, cela aussi peut avoir une incidence sur son utilisation, et donc sur les annotations.

Poids du contexte (Fort, Nazarenko et al., 2012) Annoter des unités « indépendantes », qui ne requièrent pas d'être interprétées au sein d'un contexte plus large, se révèle une tâche avec une charge cognitive moindre par rapport à une annotation où la taille du contexte est importante (phrase, paragraphe, texte entier…). En effet, plus la taille du contexte à prendre en compte est large, plus le travail d'annotation est long et plus l'annotateur se fatigue rapidement. Cela peut l'amener à commettre plus d'erreurs d'inattention. Dans un domaine moins répandu, et même si cela est un cas extrême, une campagne d'annotation d'entités nommées, d'actions et de relations dans les matchs de foot a nécessité d'avoir accès aux vidéos des matchs pour annoter (Fort & Claveau, 2012).

Connaissances du domaine La taille du contexte se trouve souvent liée aux connaissances connexes intervenant durant l'annotation. Selon la tâche ou l'expérience de l'annotateur, celui-ci peut avoir besoin d'accéder à des informations complémentaires pour pouvoir annoter ou interpréter le contexte. Cela est notamment le cas lors de campagnes d'annotation dans le domaine biomédical, entre autres, où les annotateurs ont souvent accès à des bases de connaissances sur les maladies ou à PubMed.

Temps d'annotation Il existe deux façons d'appréhender le temps d'annotation :

— au niveau de la session : annoter est un travail qui prend du temps et, par conséquent, amène une fatigue. Si l'annotateur effectue une longue session d'annotation, il y a le risque qu'il fasse davantage d'erreurs à cause de la fatigue.
— au niveau de l'intégralité de la campagne : rejoignant l'idée de la formation, l'annotateur s'accoutume à la tâche au fil des annotations et celles-ci deviennent (normalement) plus fiables (voir la notion de courbe d'apprentissage (Fort & Sagot, 2010)).

Évaluation des annotations

Mesures d'accord inter-annotateurs non adaptées Un biais répandu est d'utiliser une mesure d'accord inter-annotateur qui ne reflète pas forcément les caractéristiques du phénomène étudié. Par exemple, comme l'ont montré Mathet et Widlöcher (2016) dans le cas où la tâche d'annotation nécessite de l'*unitizing*, utiliser un κ de Cohen en combinaison avec de l'atomisation peut amener à considérer comme excellentes des annotations avec des désaccords flagrants. Malheureusement, il n'existe pas de mesures adaptées à toutes les tâches.

Catégories aux frontières perméables L'utilisation de telle ou telle mesure d'accord peut aussi dépendre du schéma d'annotation, notamment si les distances entre les catégories diffèrent selon les catégories (comme le cas d'annotation en échelle de valeur). Utiliser des mesures pondérées, comme le κ_ω, α ou encore le γ_{cat}, permet de mieux rendre compte de l'accord inter-annotateurs dans ce genre de cas.

Biais en évaluation Di Eugenio et Glass (2004), repris par Paun et al. (2022), soulignent deux biais principaux dans les mesures d'accord inter-annotateurs :
— **Biais de l'annotateur**, parfois appelé **paradoxe du** κ : le κ favorise les distributions déséquilibrées entre les annotateurs, *a contrario* de π et α : quand les annotateurs vont être en désaccord avec la distribution des catégories, alors cela va faire croître l'accord. ;
— **Prévalence des catégories** : si la distribution des catégories est déséquilibrée (par exemple, une catégorie commune et une rare), l'accord obtenu sera focalisé presque exclusivement sur la catégorie rare. En effet, les mesures corrigées par la chance sont sensibles à l'accord sur les catégories rares (Fort et al., 2010).

Référence et Diffusion

Être juge et partie Ce biais concerne davantage l'éthique. Il s'agit, lorsque différentes théories existent pour un phénomène linguistique, d'être responsable ou annotateur lors d'une campagne et de proposer une mesure ou une métrique pour évaluer les annotations. Il est alors difficile d'être neutre et :

— soit d'annoter sans favoriser la manière dont la mesure ou la métrique calcule le score ;

— soit d'élaborer une mesure ou métrique sans avantager la théorie utilisée pour l'annotation.

Cela peut amener des risques en matière d'évaluation des systèmes.

Manque d'harmonisation Comparer les corpus annotés pour une même tâche ou les outils entraînés à partir des corpus de référence n'est pas toujours aisé et réalisable, faute d'harmonisation (parfois à raison) entre les différents schémas. Le manque d'harmonisation provient des guides et schéma d'annotation choisis, dépendant parfois d'un domaine (par exemple en contexte de recherche orienté Humanités ou biomédical), d'un modèle (comme pour la coréférence) ou d'une vision de la tâche (Amidei et al., 2018). Se pose alors la question de leur comparaison. Des études (Grouin, 2018 ; Millour et al., 2022) ont tenté de répondre à cette question et, s'il est possible de réduire un jeu d'étiquettes fin à un jeu à gros grain, l'inverse est impossible. Dès lors, il est difficile de comparer les différents jeux de données et les outils qui en dépendent.

Problème pour l'apprentissage automatique Celui-ci, introduit pendant l'annotation, affecte des tâches situées en eval. L'efficacité des outils entraînés dépend aussi du schéma utilisé pour annoter le corpus d'entraînement. Comme le démontrent Millour et al. (2022), si l'outil a été entraîné sur un schéma d'annotation fin, il sera pénalisé s'il doit annoter automatiquement selon un schéma d'annotation moins précis.

3.1.2 Classification temporelle

D'autres classifications des biais sont possibles, notamment en suivant l'ordre chronologique dans lequel les biais peuvent apparaître durant une campagne. Nous présentons une telle classification dans la table 3.1, en reprenant les biais dégagés précédemment. Nous distinguons, en plus, les biais collectifs, touchant à l'ensemble des annotateurs, et

les biais individuels. Il s'agit d'une première proposition, qu'il conviendrait d'approfondir et d'évaluer.

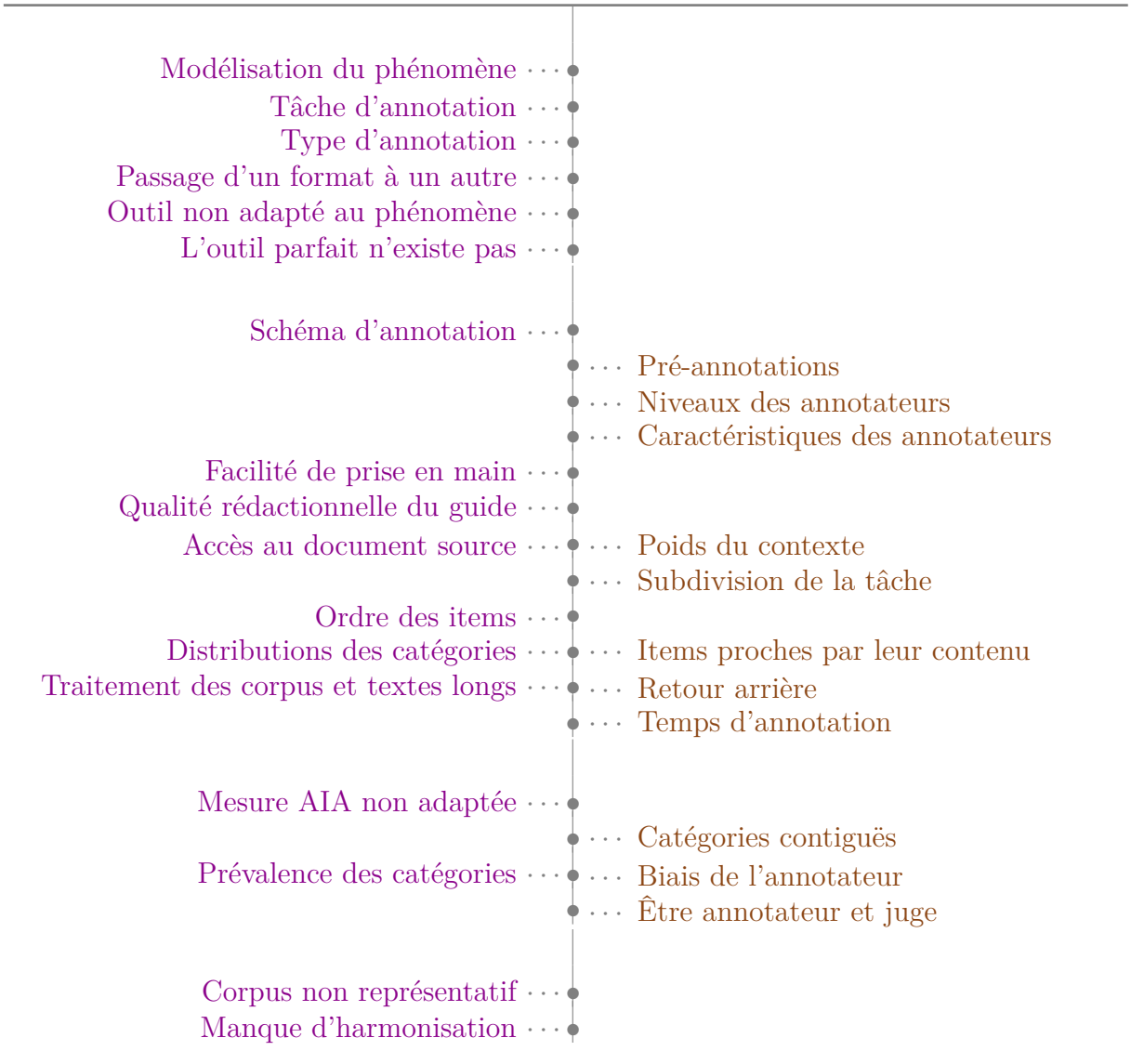


TABLE 3.1 – Une autre classification des biais, selon les dimensions Collectif/Individuel.

3.2 Méthodologie des campagnes d'annotation « Portraits » et « Erreurs »

Les deux prochains chapitres sont consacrés aux campagnes d'annotation « Portraits » et « Erreurs ». Chacune a été pensée et menée avec des hypothèses différentes à tester. Nous décrivons dans cette partie les hypothèses que nous voulions démontrer grâce à ces expériences, en détaillant la méthodologie suivie.

3.2.1 Comment étudier un biais d'annotation ?

Notre travail s'inscrit dans le projet visant à identifier et quantifier les biais d'annotation. Toutefois, l'étude d'un biais lors d'une campagne n'est pas chose aisée : il y a souvent plusieurs paramètres ou biais qui interviennent en même temps et influencent les annotateurs. Nous nous sommes interrogée sur la manière de réaliser et faciliter cette étude, sur les pistes disponibles pour isoler un biais ou pour limiter les influences d'un autre biais. De cette réflexion sont apparus certaines conditions et problèmes potentiels, listés et décrits ci-dessous, pour prévenir d'éventuelles difficultés durant l'analyse du biais :

Biais à étudier L'étude d'un biais ne peut se faire que si l'on a déjà une idée du biais ou du paramètre. Nous pouvons étudier un biais déjà traité pour une tâche particulière, et en observer l'impact sur une autre tâche ou un autre jeu de données. Nous pouvons également vouloir en identifier et en analyser de nouveaux.

Possibilité de variation dans la campagne Pour mettre en exergue l'impact du biais, adopter une approche d'étude contrastive est nécessaire. Dans la mesure du possible, il convient de proposer à plusieurs cohortes différents scénarios, où seul le paramètre à observer change. Il s'agira ensuite de comparer les annotations entre ces scénarios, tout en les mettant en perspective avec les variations introduites. Les variations peuvent être bivalentes (la présence ou non d'un paramètre) ou dans un ensemble de classes plus importantes (un paramètre avec plusieurs degrés de « dégradation »).

Généricité du biais Idéalement, le biais et les observations tirées de l'expérience doivent pouvoir s'appliquer à des tâches comparables ou des campagnes similaires. Si les effets d'un biais pourront évidemment varier d'une campagne à l'autre, nous

cherchons au moins à définir des moyens d'observations reproductibles.

Accès à une référence qui fait autorité Pour appréhender et mesurer l'impact du biais sur les annotations, il est préférable d'avoir un corpus avec une référence qui fasse autorité. Ainsi, nous pourrons mesurer l'écart entre la « vérité » et les annotations. Il est possible de comparer avec d'autres annotations qui ne sont pas forcément jugées de référence, toutefois l'impact du biais peut être minimisé, voire non détecté. Il faudrait néanmoins s'interroger sur les conséquences à en tirer en l'absence de référence, qui constitue le cas le plus fréquent.

Ayant fait appel, dans le cadre de nos deux campagnes d'annotation menées durant la thèse, à des participants ne connaissant pas la notion d'annotation et volontaires pour les tâches, deux autres pré-requis ont été de rigueur, listés ci-dessous :

Tâche requérant de l'interprétation : Bien que la tâche d'annotation doive disposer d'une référence faisant autorité, elle doit toutefois rester sujet à interprétation suffisante, pour que les annotations diffèrent suffisamment selon les annotateurs.

Tâche qui ne requiert pas d'entraînement : Puisque l'étude porte à chaque fois sur un biais spécifique, nous souhaitons minimiser les autres et notamment les éléments de différentiation entre annotateurs. Nous souhaitons donc que la tâche d'annotation soit néanmoins suffisamment simple, pour que la compétence des annotateurs ne soit pas un paramètre invisible potentiellement perturbateur.

3.2.2 Nos hypothèses et nos attentes initiales vis-à-vis des expériences

Au travers de nos campagnes, nous souhaitons nous intéresser plus particulièrement à certains biais présentés dans la section précédente.

Un des phénomènes que nous souhaitons étudier est l'*ordre de présentation des items* et l'influence d'ordres particuliers sur les annotations. Notamment, les annotateurs peuvent-ils avoir tendance à suivre une certaine logique si un motif se répète dans les annotations ? Pour ce faire, nous concevrons une campagne avec des scénarios, où l'ordre de présentation des items sera contrôlé. Nous étudierons ensuite les annotations selon les différents scénarios afin de détecter des variations qui seraient liées à l'ordre du scénario.

Nous désirons aussi observer le *rapport entre l'accord* inter-annotateurs et la *validité* des annotations produites. Nous nous interrogeons sur le bien-fondé de l'hypothèse discutable qu'un accord inter-annotateurs jugé satisfaisant est suffisant pour accéder à des annotations de référence. Au moyen de la consensualité (mesure présentée dans la section suivante), nous étudierons les liens entre la performance des annotateurs et leur consensualité pour les deux campagnes menées, pour appuyer ou non l'hypothèse ci-dessus. Nous espérons notamment montrer une corrélation entre la performance et la consensualité des annotateurs, qui puisse compléter l'accord inter-annotateur.

La *possibilité du retour arrière* et son impact sur les annotations nous intéressent aussi. Nous proposerons ainsi aux annotateurs deux variantes d'une campagne : une où les annotateurs auront accès à leurs anciennes annotations et y pourront revenir, une autre où il sera impossible de revoir ses anciennes annotations. Nous cherchons à savoir si une différence dans les annotations, que cela soit quantifiable par rapport à une référence, ou dans la manière d'annoter.

Un phénomène supplémentaire que nous souhaitons étudier est l'annotation d'*items proches par leur contenu*. Dans un corpus, certains objets à annoter possèdent parfois des structures ou des particularités voisines, et nous souhaitons analyser la manière dont les annotateurs prennent en compte ces ressemblances pour annoter. À la manière des scénarios, nous inclurons dans notre campagnes « Erreurs » des items comparables entre eux et nous regarderons en détail l'annotation de ces items.

3.2.3 Un outil transversal : la consensualité

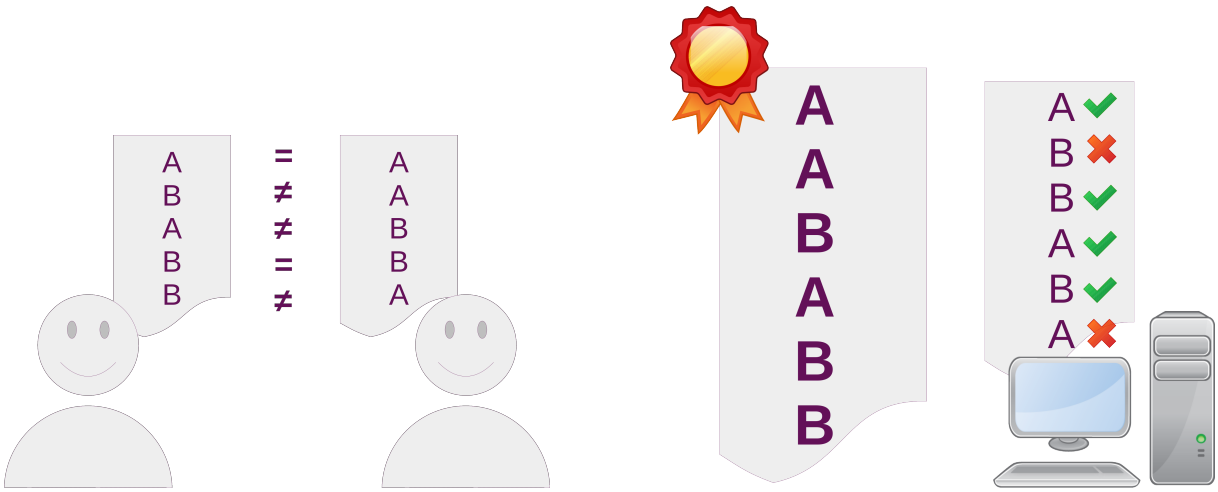
3.2.3.1 Accord et validité

Avant toute chose, il convient de bien faire la distinction entre les notions d'accord et de validité, qui sont parfois mal comprises, voire confondues :

1. **Accord** : dans le cadre de l'annotation multiple, une mesure d'accord inter-annotateurs tente de proposer un indice sur le degré de similarité entre les annotations provenant d'annotateurs distincts ;
2. **Validité** : généralement utilisée dans le cadre d'un processus automatique d'annotation, une mesure de validité tente, quant à elle, de se prononcer sur le degré

de similarité entre une annotation candidate, et une référence (généralement une annotation jugée valide par un expert).

Les deux figures ci-dessous illustrent ces deux notions[1] :



(a) Accord (b) Validité.

FIGURE 3.1 – Schématisation des deux notions.

Si ces deux mesures peuvent se ressembler puisqu'elles se prononcent sur des similarités entre des jeux d'annotations, mentionnons une première différence fondamentale : dans le cas (1), il n'y a pas de référence, les entrées ont toutes le même statut, alors qu'il y a une profonde asymétrie dans le cas (2), où l'on compare un candidat à la « vérité ». D'autres différences peuvent exister, comme le fait que dans le cas (1), il peut y avoir autant de jeux d'annotation qu'on le souhaite (par exemple 10 annotateurs), tandis qu'il y en a systématiquement 2 dans le cas (2). Par ailleurs, dans le cas (1), les mesures tentent généralement de retirer la part de « chance » qui intervient dans l'accord entre annotateurs.

Il est important d'insister sur le fait que si plusieurs annotateurs annotent tous parfaitement, leurs annotations seront similaires, et leur accord sera total. Mais la réciproque n'est malheureusement pas établie : des annotateurs peuvent être en accord (sur tout ou une partie des éléments annotés) sans pour autant que leurs annotations soient valides, car ils peuvent commettre les mêmes erreurs — le cas du Polish Treebank (WOLIŃSKI

1. Pour la figure représentant la notion de validité, nous avons fait le choix de représenter l'annotation candidate avec une icône d'ordinateur. Bien évidemment, l'annotation candidate n'est toujours pas issue d'un système automatique.

et al., 2011) illustre bien cette réalité. Cela souligne la pertinence des expérimentations proposées.

3.2.3.2 Vers la définition de la consensualité

Dans le cadre de la campagne « Portraits » (chapitre 4), qui a pour but d'annoter des images, nous avons utilisé la notion de la consensualité. Nous présentons ici les formules et notations que nous avons définie ; la partie 5.4.1 permettra d'étendre ces définitions pour la campagne « Erreurs ».

Pour une image i, μ_i désigne la référence et $x_{i,a}$ réfère à l'annotation de l'annotateur a pour cette image. N est le nombre total d'images. Soit un sous-groupe d'annotateurs, $G = \{a_1, ..., a_n\}$, $|G|$ représente la cardinalité de ce groupe et $\sigma_i(G)$ la variance de ce groupe pour l'image i. Nous définissons ensuite les concepts et les formules suivants :

— l'**imperfection d'un annotateur** (le contraire de la **performance de l'annotateur**) est inspirée de la formule de l'écart-type, et elle est donnée par la formule 3.1 :

$$\text{imperfection}(a) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{i,a} - \mu_i)^2} \tag{3.1}$$

Quand elle est nulle, les annotations de l'annotateur sont toutes valides.

— l'**imperfection d'un groupe d'annotateurs** est la moyenne de toutes les imperfections des annotateurs du groupe. Voir la formule 3.2 :

$$\text{imperfection}(G) = \frac{1}{|G|}\sum_{j=1}^{|G|}\text{imperfection}(a_j) \tag{3.2}$$

Quand elle est nulle, toutes les annotations de tous les annotateurs du groupe sont valides.

— le **désaccord d'un groupe d'annotateurs** (donné par la formule 3.3) : pour chaque image i, nous prenons la variance des annotations (σ_i) ; le désaccord est la moyenne de cette valeur pour toutes les images.

$$\text{desaccord}(G) = \frac{1}{N}\sum_{i=1}^{N}\sigma_i(G) \tag{3.3}$$

Quand ce désaccord est nul, les annotateurs ont tous, pour chacune des photos,

donné le même âge (mais pas forcément le bon).

— le **degré de consensualité d'un annotateur vis-à-vis d'un groupe** (auquel l'annotateur appartient) est donné par la différence algébrique entre le désaccord de ce groupe privé de cet annotateur a, et le désaccord de ce groupe. Voir la formule 3.4, pour l'annotateur $a \in G$:

$$\text{consensualite}(a, G) = \text{desaccord}(G \setminus a) - \text{desaccord}(G) \tag{3.4}$$

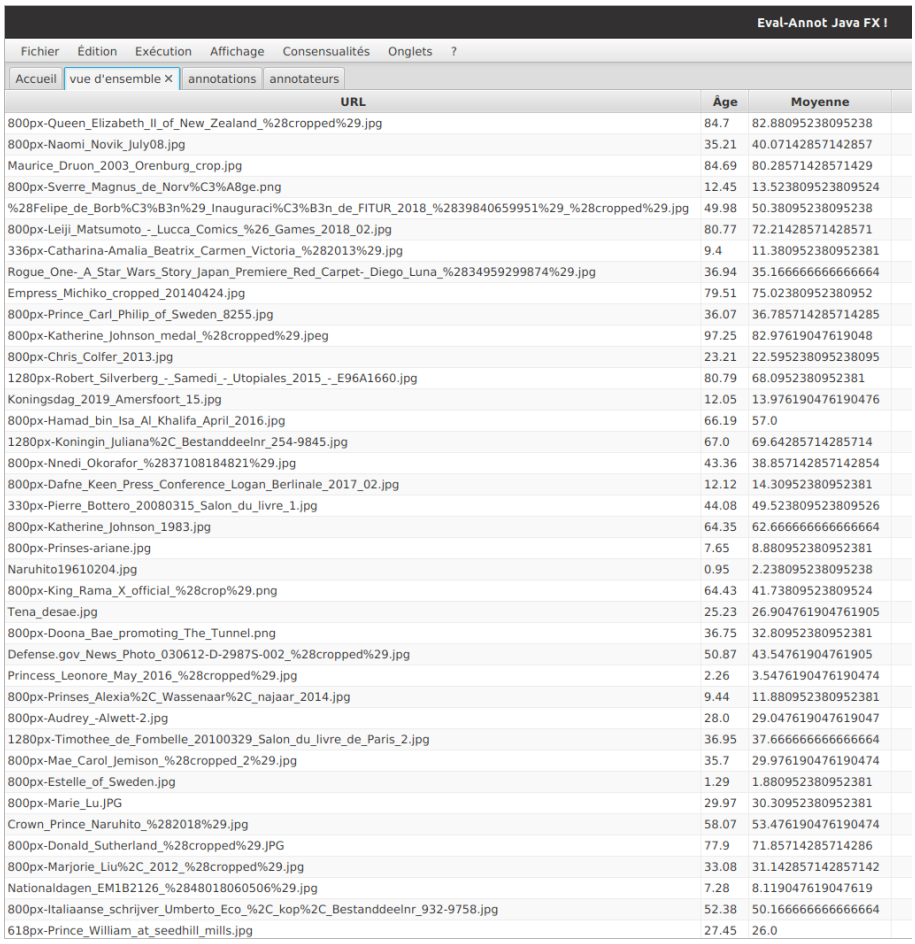
Si cette valeur algébrique est positive, cela signifie que l'annotateur génère de l'accord, et donc qu'il est consensuel. Attention, le degré de consensualité de l'annotateur est relatif à un groupe : on peut être consensuel vis-à-vis de certains annotateurs, et non consensuel vis-à-vis d'autres annotateurs. Nous distinguons deux façons d'établir la consensualité individuelle :

1. **Consensualité initiale** : pour chaque annotateur du groupe considéré, nous calculons sa consensualité vis à vis de l'ensemble du groupe, et nous trions les annotateurs selon leur valeur de consensualité.
2. **Consensualité progressive** : dans cette variante, nous procédons de façon itérative. À partir de la consensualité initiale, nous retirons l'annotateur le moins consensuel. Nous réitérons le processus à partir du sous-groupe restant, en recalculant à chaque itération toutes les consensualités à partir du groupe restant. Ainsi, nous essayons de garder, petit à petit, les annotateurs les plus consensuels parmi ceux déjà les plus consensuels.

3.2.3.3 Plateforme de mise en œuvre : l'application Éval-Annot

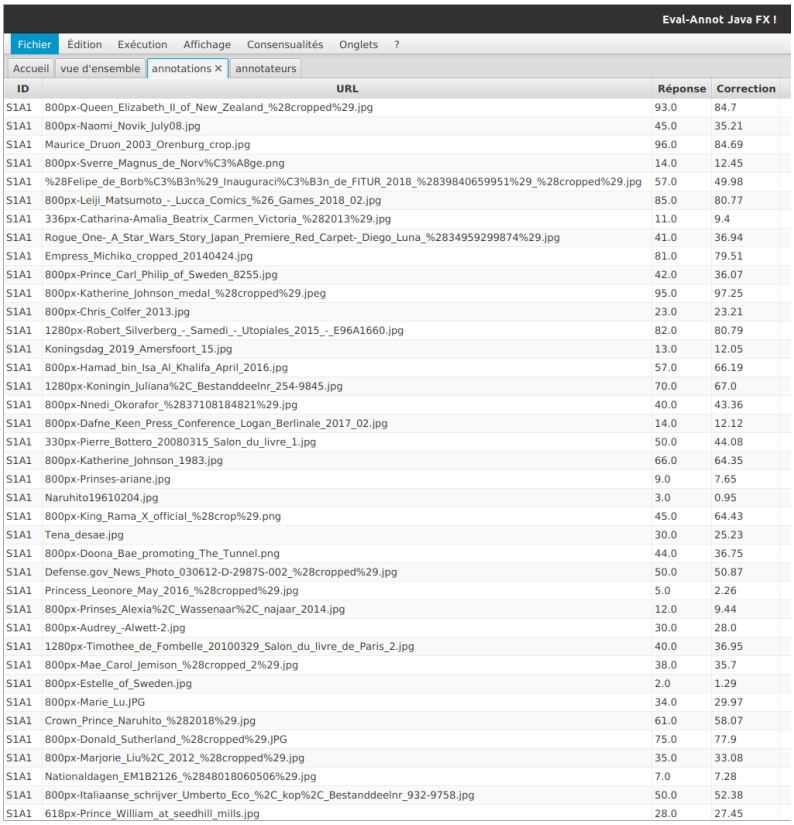
Toujours dans le cadre initial de la campagne « Portraits », l'application Éval-Annot a été développée par Jean-Luc Manguin et Christophe Couronne, au sein du service Développement et Déploiement d'Applications du GREYC. L'application est implémentée en Java et une interface graphique est disponible. Elle prend en entrée un fichier CSV contenant les identifiants des annotateurs et ceux des images, ainsi que les âges exacts des photographies et toutes les annotations. Si nous le souhaitons, nous pouvons ne charger que les annotateurs ayant estimé au moins un des portraits, voire n'ayant que des annotations exploitables ; dans ce dernier cas, pour la campagne « Portraits », il y a un total de 42 annotateurs, soit 4 200 annotations.

L'application a été produite avec pour objectif principal l'analyse de la corrélation entre la performance et la consensualité : dans ce but, de nombreux graphiques peuvent être générés mettant en exergue ces deux concepts, et plus particulièrement la différence entre les consensualités initiale et dynamique. Il est aussi important de signaler que l'application Éval-Annot sera proposée en code source ouvert. Les utilisateurs pourront implémenter leurs propres mesures d'accord, que l'on peut coder rapidement et les utiliser directement dans l'application sans avoir à la relancer. Toujours dans une optique de généralisation, le logiciel pourra être adaptable à différents jeux de données, voire types de données autres que numériques.



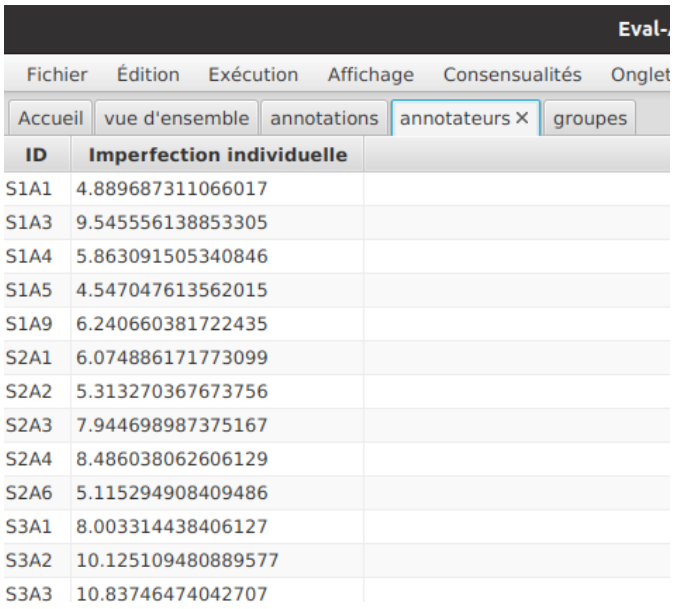
URL	Âge	Moyenne
800px-Queen_Elizabeth_II_of_New_Zealand_%28cropped%29.jpg	84.7	82.88095238095238
800px-Naomi_Novik_July08.jpg	35.21	40.07142857142857
Maurice_Druon_2003_Orenburg_crop.jpg	84.69	80.28571428571429
800px-Sverre_Magnus_de_Norv%C3%A8ge.png	12.45	13.523809523809524
%28Felipe_de_Borb%C3%B3n%29_Inauguraci%C3%B3n_de_FITUR_2018_%2839840659951%29_%28cropped%29.jpg	49.98	50.38095238095238
800px-Leiji_Matsumoto_-_Lucca_Comics_%26_Games_2018_02.jpg	80.77	72.21428571428571
336px-Catharina-Amalia_Beatrix_Carmen_Victoria_%282013%29.jpg	9.4	11.380952380952381
Rogue_One-_A_Star_Wars_Story_Japan_Premiere_Red_Carpet-_Diego_Luna_%2834959299874%29.jpg	36.94	35.166666666666664
Empress_Michiko_cropped_20140424.jpg	79.51	75.02380952380952
800px-Prince_Carl_Philip_of_Sweden_8255.jpg	36.07	36.785714285714285
800px-Katherine_Johnson_medal_%28cropped%29.jpeg	97.25	82.97619047619048
800px-Chris_Colfer_2013.jpg	23.21	22.595238095238095
1280px-Robert_Silverberg_-_Samedi_-_Utopiales_2015_-_E96A1660.jpg	80.79	68.0952380952381
Koningsdag_2019_Amersfoort_15.jpg	12.05	13.976190476190476
800px-Hamad_bin_Isa_Al_Khalifa_April_2016.jpg	66.19	57.0
1280px-Koningin_Juliana%2C_Bestanddeelnr_254-9845.jpg	67.0	69.64285714285714
800px-Nnedi_Okorafor_%2837108184821%29.jpg	43.36	38.857142857142854
800px-Dafne_Keen_Press_Conference_Logan_Berlinale_2017_02.jpg	12.12	14.30952380952381
330px-Pierre_Bottero_20080315_Salon_du_livre_1.jpg	44.08	49.523809523809526
800px-Katherine_Johnson_1983.jpg	64.35	62.666666666666664
800px-Prinses-ariane.jpg	7.65	8.880952380952381
Naruhito19610204.jpg	0.95	2.238095238095238
800px-King_Rama_X_official_%28crop%29.png	64.43	41.73809523809524
Tena_desae.jpg	25.23	26.904761904761905
800px-Doona_Bae_promoting_The_Tunnel.png	36.75	32.80952380952381
Defense.gov_News_Photo_030612-D-2987S-002_%28cropped%29.jpg	50.87	43.54761904761905
Princess_Leonore_May_2016_%28cropped%29.jpg	2.26	3.5476190476190474
800px-Prinses_Alexia%2C_Wassenaar%2C_najaar_2014.jpg	9.44	11.880952380952381
800px-Audrey_-Alwett-2.jpg	28.0	29.047619047619047
1280px-Timothee_de_Fombelle_20100329_Salon_du_livre_de_Paris_2.jpg	36.95	37.666666666666664
800px-Mae_Carol_Jemison_%28cropped_2%29.jpg	35.7	29.976190476190474
800px-Estelle_of_Sweden.jpg	1.29	1.880952380952381
800px-Marie_Lu.JPG	29.97	30.30952380952381
Crown_Prince_Naruhito_%282018%29.jpg	58.07	53.476190476190474
800px-Donald_Sutherland_%28cropped%29.JPG	77.9	71.85714285714286
800px-Marjorie_Liu%2C_2012_%28cropped%29.jpg	33.08	31.142857142857142
Nationaldagen_EM1B2126_%2848018060506%29.jpg	7.28	8.119047619047619
800px-Italiaanse_schrijver_Umberto_Eco_%2C_kop%2C_Bestanddeelnr_932-9758.jpg	52.38	50.166666666666664
618px-Prince_William_at_seedhill_mills.jpg	27.45	26.0

Figure 3.2 – Vue d'ensemble du corpus.



ID	URL	Réponse	Correction
S1A1	800px-Queen_Elizabeth_II_of_New_Zealand_%28cropped%29.jpg	93.0	84.7
S1A1	800px-Naomi_Novik_July08.jpg	45.0	35.21
S1A1	Maurice_Druon_2003_Orenburg_crop.jpg	96.0	84.69
S1A1	800px-Sverre_Magnus_de_Norv%C3%A8ge.png	14.0	12.45
S1A1	%28Felipe_de_Borb%C3%B3n%29_Inauguraci%C3%B3n_de_FITUR_2018_%2839840659951%29_%28cropped%29.jpg	57.0	49.98
S1A1	800px-Leiji_Matsumoto_-_Lucca_Comics_%26_Games_2018_02.jpg	85.0	80.77
S1A1	336px-Catharina-Amalia_Beatrix_Carmen_Victoria_%282013%29.jpg	11.0	9.4
S1A1	Rogue_One-_A_Star_Wars_Story_Japan_Premiere_Red_Carpet-_Diego_Luna_%2834959299874%29.jpg	41.0	36.94
S1A1	Empress_Michiko_cropped_20140424.jpg	81.0	79.51
S1A1	800px-Prince_Carl_Philip_of_Sweden_8255.jpg	42.0	36.07
S1A1	800px-Katherine_Johnson_medal_%28cropped%29.jpeg	95.0	97.25
S1A1	800px-Chris_Colfer_2013.jpg	23.0	23.21
S1A1	1280px-Robert_Silverberg_-_Samedi_-_Utopiales_2015_-_E96A1660.jpg	82.0	80.79
S1A1	Koningsdag_2019_Amersfoort_15.jpg	13.0	12.05
S1A1	800px-Hamad_bin_Isa_Al_Khalifa_April_2016.jpg	57.0	66.19
S1A1	1280px-Koningin_Juliana%2C_Bestanddeelnr_254-9845.jpg	70.0	67.0
S1A1	800px-Nnedi_Okorafor_%2837108184821%29.jpg	40.0	43.36
S1A1	800px-Dafne_Keen_Press_Conference_Logan_Berlinale_2017_02.jpg	14.0	12.12
S1A1	330px-Pierre_Bottero_20080315_Salon_du_livre_1.jpg	50.0	44.08
S1A1	800px-Katherine_Johnson_1983.jpg	66.0	64.35
S1A1	800px-Prinses-ariane.jpg	9.0	7.65
S1A1	Naruhito19610204.jpg	3.0	0.95
S1A1	800px-King_Rama_X_official_%28crop%29.png	45.0	64.43
S1A1	Tena_desae.jpg	30.0	25.23
S1A1	800px-Doona_Bae_promoting_The_Tunnel.png	44.0	36.75
S1A1	Defense.gov_News_Photo_030612-D-2987S-002_%28cropped%29.jpg	50.0	50.87
S1A1	Princess_Leonore_May_2016_%28cropped%29.jpg	5.0	2.26
S1A1	800px-Prinses_Alexia%2C_Wassenaar%2C_najaar_2014.jpg	12.0	9.44
S1A1	800px-Audrey_-Alwett-2.jpg	30.0	28.0
S1A1	1280px-Timothee_de_Fombelle_20100329_Salon_du_livre_de_Paris_2.jpg	40.0	36.95
S1A1	800px-Mae_Carol_Jemison_%28cropped_2%29.jpg	38.0	35.7
S1A1	800px-Estelle_of_Sweden.jpg	2.0	1.29
S1A1	800px-Marie_Lu.JPG	34.0	29.97
S1A1	Crown_Prince_Naruhito_%282018%29.jpg	61.0	58.07
S1A1	800px-Donald_Sutherland_%28cropped%29.JPG	75.0	77.9
S1A1	800px-Marjorie_Liu%2C_2012_%28cropped%29.jpg	35.0	33.08
S1A1	Nationaldagen_EM1B2126_%2848018060506%29.jpg	7.0	7.28
S1A1	800px-Italiaanse_schrijver_Umberto_Eco_%2C_kop%2C_Bestanddeelnr_932-9758.jpg	50.0	52.38
S1A1	618px-Prince_William_at_seedhill_mills.jpg	28.0	27.45

FIGURE 3.3 – Onglet Annotations, regroupant toutes les annotations.



ID	Imperfection individuelle
S1A1	4.889687311066017
S1A3	9.545556138853305
S1A4	5.863091505340846
S1A5	4.547047613562015
S1A9	6.240660381722435
S2A1	6.074886171773099
S2A2	5.313270367673756
S2A3	7.944698987375167
S2A4	8.486038062606129
S2A6	5.115294908409486
S3A1	8.003314438406127
S3A2	10.125109480889577
S3A3	10.83746474042707

FIGURE 3.4 – Onglet Annotateurs, permettant de voir l'imperfection individuelle.

3.3 Conclusion

La principale contribution de ce chapitre réside en la formalisation de ce que nous appelons les biais d'annotation. Pour rappel, ce sont des « phénomènes perturbateurs, de nature et d'origine variées, pouvant survenir à toutes les étapes du processus d'annotation et qui sont susceptibles d'en affecter le résultat ». Ces phénomènes sont variés, dépendant aussi bien de l'annotateur de façon isolée que d'un ensemble d'annotateurs, ou encore de la tâche d'annotation.

Nous avons proposé une première typologie des biais, reprenant les grandes étapes d'une campagne. Elle n'est pas exempte de défauts, car des biais peuvent avoir des impacts à différentes étapes, ou dépendre d'une étape et être problématiques lors d'une étape ultérieure. Par exemple, la constitution du corpus est une des toutes premières étapes, mais le véritable risque lié à un corpus non représentatif ne survient qu'au moment de la diffusion.

Chapitre 4

Campagne d'annotations « Portraits »

Sommaire

4.1 Présentation de la campagne 103

4.1.1 Constitution du corpus 103

4.1.2 Biais concernant l'estimation de l'âge 104

4.1.3 Scénarios 105

4.1.4 Déroulement de la campagne d'annotation 106

4.1.5 Une première approche des annotations récoltées : comparaison avec la référence 108

4.2 Analyse des consensualités 109

4.2.1 Rang de consensualité *versus* rang de performance 110

4.2.2 Retirer les annotateurs les moins consensuels 112

4.2.3 Distinguer les consensualités initiale et dynamique 114

4.2.4 Tester l'homogénéité de la consensualité 116

4.3 Influence de l'ordre des items 126

4.3.1 Avec un accès à la référence 127

4.3.2 Détecter un biais sans un accès à la référence 130

4.4 Résultats complémentaires 135

4.5 Conclusion 138

Pour la première expérience menée, en tenant compte des contraintes fixées présentées dans le chapitre en 3.2.1, nous avons choisi la tâche d'estimation des âges d'individus humains sur des photographies. Cette tâche allie en effet plusieurs critères évoqués :

— la référence est facile à constituer, en raison du fait qu'il est aisé d'avoir accès à l'âge exact de la personne au moment de la prise de la photographie ;
— l'estimation de l'âge relève d'une interprétation des caractéristiques visuelles du visage et les estimations peuvent beaucoup différer pour une même photographie ;
— il s'agit d'une tâche à laquelle nous sommes toujours confrontés, il n'est pas donc nécessaire d'avoir suivi un entraînement préalable.

De plus, le caractère numérique des âges ajoute deux plus-values intéressantes pour l'expérience :

Annotations scalaires : Les annotations numériques scalaires nous permettent de pouvoir quantifier finement l'écart à la référence, sans nous limiter à une évaluation binaire. Par ailleurs, nous voulons observer de plus près l'évaluation de telles annotations, encore peu étudiée.

Annotations agrégeables : Pour chaque item, des fonctions d'agrégation peuvent être simplement utilisées pour estimer la position du groupe, par exemple la moyenne.

Bien que l'objet de cette campagne soit éloigné des préoccupations langagières du T.A.L., cette expérience nous semble pertinente. En effet, la méthode suivie correspond à celle d'une campagne d'annotation traditionnelle en T.A.L. De plus, la facilité d'implémentation nous a permis de constituer rapidement un corpus de qualité. Enfin, la tâche à effectuer étant perçue relativement simple par les annotateurs, ces derniers se sont facilement prêtés au jeu. Nous avons donc pu récolter une quantité d'annotations raisonnable en peu temps. Nous pensons aussi que nos conclusions peuvent être utilisables et applicables sur l'annotation en général, et par conséquent sur des données textuelles.

Avec cette campagne, nous nous intéressons à trois aspects :
— le rapport entre l'accord inter-annotateurs et la validité de leurs annotations ;
— le biais de l'ordre de présentation des items.
— l'évaluation des annotations numériques.

Une partie du contenu de ce chapitre a fait l'objet d'une publication à la conférence Language Resources and Evaluation Conference (LREC) (Baledent et al.,

2022).

4.1 Présentation de la campagne

4.1.1 Constitution du corpus

La tâche et l'idée de corpus ainsi décidées, il convient aussi de réfléchir à ce qu'implique l'utilisation de portraits et à la source de ces derniers. Cette considération vaut d'autant plus pour les mineurs, pour lesquels il est souvent difficile d'avoir des photographies respectueuses de la loi et du droit à l'image. Bien qu'il existe déjà des corpus de photographies, ils sont néanmoins souvent non disponibles ou peu accessibles librement. Parfois, selon l'origine ou l'objectif de l'expérience pour laquelle ils ont été constitués, ils ne contiennent pas de photographies de personnes mineures ou les métadonnées nécessaires (par exemple l'âge de la personne). Pour ces raisons, nous avons décidé de ne pas utiliser un corpus de photographies déjà existant et constituer nous-même le corpus pour l'expérience.

La principale difficulté à laquelle nous avons été confrontée est la licence des photographies. Pour cela, nous avons choisi de privilégier des photographies de personnes publiques [1], comme des membres de familles royales, des personnalités liées au monde du cinéma ou des livres, disponibles sur le site du projet Wikimedia Commons. Prendre comme sujets des personnes publiques règle notamment les problèmes relatifs aux photographies de mineurs [2], pour lesquelles l'accord des responsables légaux est normalement nécessaire. Cela nous a aussi permis de calculer l'âge exact de la personne, ayant accès à la date de naissance et à la date de prise de la photographie.

En tout, le corpus est constitué de cent photographies de personnes âgées entre 3 mois et 97 ans. Le tableau 4.1 détaille la répartition des photographies selon les dix tranches d'âge que nous avons choisies. Le détail du corpus (code, liens hypertextes des photographies et âge) figure dans l'annexe A.

1. Considérer de tels sujets pour la tâche peut introduire un biais de données par rapport à des sujets anonymes : les photographies ont souvent été travaillées et les personnalités publiques adoptent souvent des postures les mettant à leur avantage, ce qui peut les vieillir ou les rajeunir — parfois fortement — selon l'image qu'ils préfèrent renvoyer

2. En effet, il nous semble important d'avoir un corpus avec le plus d'hétérogénéité et de représentativité possibles concernant les âges à estimer.

Catégorie	1	2	3	4	5	6	7	8	9	10
Tranche	[0;10[	[10;20[	[20;30[	[30;40[	[40;50[	[50;60[	[60;70[	[70;80[	[80;90[	[90;100[
Répartition	14	17	13	10	10	8	9	7	9	3

TABLE 4.1 – Répartition des photographies selon les tranches d'âge.

Un des aspects importants était en la création d'un corpus représentatif de la diversité, c'est-à-dire sélectionner des photos d'individus humains de genres et d'ethnies différents. Il s'agit avant tout d'un devoir symbolique et moral.

4.1.2 Biais concernant l'estimation de l'âge

CLIFFORD et al. (2018) évoquent deux biais courants dans les tâches d'estimation de l'âge à partir du visage : d'une part la dépendance sérielle, phénomène qui induit d'être influencé par l'item précédent, d'autre part le fait d'avoir tendance à vieillir les personnes aux visages jeunes et à rajeunir les personnes à l'apparence âgée. Ce biais est aussi décrit par VESTLUND et al. (2009).

La dépendance sérielle visuelle, comme c'est le cas dans notre tâche d'annotation, a déjà été abordée dans les travaux de FISCHER et WHITNEY (2014) ; PEGORS et al. (2015). Les auteurs décrivent deux formes de dépendance visuelle, biaisant le jugement de l'annotateur :

— une dépendance assimilative, qui ferait paraître l'image actuelle plus similaire à la précédente ;
— une dépendance contrastive, où l'image détonnerait plus de l'image précédente.

Ces biais peuvent donc affecter plusieurs jugements lors de la reconnaissance faciale, comme l'identification du genre de la personne ou son expression (LIBERMAN et al., 2014).

D'autres biais peuvent intervenir lors de l'estimation de l'âge : VOELKLE et al. (2012) évoquent l'habitude de côtoyer des personnes de notre âge qui nous permettrait de mieux estimer l'âge d'une personne issue de cette tranche d'âge. Enfin, dans WATSON et al. (2016), les auteurs remarquent une tendance à « rapprocher » l'âge de la personne au nôtre.

4.1.3 Scénarios

Un des objectifs principaux de notre expérience est d'étudier l'influence de l'ordre de présentation des items sur les annotations. Pour étudier ce biais, nous avons choisi de proposer différents scénarios en incluant des variations relatives à la chronologie. Ces variations concernent :

— l'ordre des tranches d'âge ;
— l'ordre des photographies au sein de ces tranches ;
— le facteur de présentation aléatoire, avec un aléatoire plus ou moins poussé.

Il y a sept scénarios :

Scénario 1 (S1) : Toutes les photographies sont présentées dans l'ordre croissant des âges.

Scénario 2 (S2) : Les tranches d'âge sont présentées dans l'ordre croissant et les photographies de chaque tranche sont présentées aléatoirement.

Scénario 3 (S3) : Toutes les photographies sont présentées dans un ordre aléatoire.

Scénario 4 (S4) : Les tranches d'âge sont présentées dans l'ordre croissant, sauf les tranches d'âge 4 et 5 qui ont été interverties, ainsi que les 8 et 9, et les photographies de chaque tranche sont présentées dans l'ordre croissant.

Scénario 5 (S5) : Les tranches d'âge sont présentées dans l'ordre croissant, sauf les tranches d'âge 4 et 5 qui ont été interverties ainsi que 8 et 9, et les photographies de chaque tranche sont présentées dans l'ordre aléatoire. L'ordre aléatoire est différent de celui du S2.

Scénario 6 (S6) : Les tranches d'âge 1, 2, 3, 4, 5, 9 et 10 sont mélangées et les photographies présentées dans l'ordre aléatoire mais les tranches 8, 7 et 6 sont présentées dans cet ordre à la fin du scénario et leurs photographies dans l'ordre croissant des âges.

Scénario 7 (S7) : Les tranches d'âge 1, 2, 3, 4, 5, 9 et 10 sont mélangées et les photographies présentées dans l'ordre aléatoire mais les tranches 8, 7 et 6 sont présentées dans cet ordre à la fin du scénario et leurs photographies aléatoirement.

4.1.4 Déroulement de la campagne d'annotation

4.1.4.1 Premier appel à participation

Initialement, la campagne s'est tenue dans le cadre d'un cours dispensé à l'ensemble d'une licence Humanités Numériques à l'université de Caen. Les annotateurs étaient donc des étudiants de licence, sans formation ou sensibilisation spécifiques à l'annotation. La campagne d'annotation s'est déroulée via la plateforme Moodle de l'université. Comme la plateforme ne permettait pas d'attribuer aléatoirement un scénario à un étudiant ou à un groupe, nous avons dû, avant le lancement de la campagne, nous avons réparti les étudiants dans sept groupes, de manière homogène, et chaque groupe s'est vu attribuer un scénario.

Le temps estimé de la tâche est de trente minutes, mais une heure entière était laissée aux annotateurs. Dans les instructions données aux annotateurs, il était demandé de saisir un nombre entier correspondant à l'estimation la plus précise de l'âge de la personne. Il était aussi indiqué aux annotateurs de ne s'appuyer que sur les informations visuelles données par la photographie elle-même, sans chercher à identifier la personne concernée ou à retrouver la photographie sur le web en quête d'indications complémentaires [3]. De plus, il avait été précisé de traiter les questions dans l'ordre dans lequel elles étaient présentées, sans revenir en arrière pour modifier une ancienne annotation. Les consignes exactes envoyées aux étudiants sont présentées dans l'annexe A.

Les étudiants n'ayant pas tous réalisé le questionnaire, le nombre d'annotateurs par scénario varie légèrement (voir le tableau 4.2). En tout, cinquante-deux annotateurs ont répondu à la campagne, avec une moyenne de sept annotateurs par scénario, et nous avons recueilli 4 850 annotations exploitables [4].

3. Évidemment, nous ne pouvons pas empêcher les annotateurs de reconnaître les personnes en photographie. Cette précision vaut surtout pour limiter les biais externes : cela peut permettre à l'annotateur d'estimer plus précisément l'âge, mais aussi induire des erreurs si l'annotateur se trompe lors de son observation.

4. Tous les annotateurs n'ont pas annoté les 100 photographies et certaines annotations n'étaient pas au format demandé.

Scénario	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**
Annotateurs	9	6	7	8	6	7	9
Annotations exploitables	738	599	700	702	533	698	880

TABLE 4.2 – Nombre d'annotateurs par scénario (vague 1)

4.1.4.2 Relance de la campagne

Un an après la première récolte d'annotations, nous avons relancé la campagne « Portraits ». Cette relance avait un objectif principal : le but était de tester et de se familiariser avec la plateforme de questionnaire LIMESURVEY, en vue de la campagne des « Erreurs » (présentée dans le chapitre 5). Nous voulions aussi tester l'hypothèse selon laquelle l'âge de l'annotateur peut influencer les estimations d'âges, en accord avec les résultats de CLIFFORD et al. (2018). Pour étudier ce dernier point, dont la réponse est optionelle, une question permet à l'annotateur de renseigner son âge au début du questionnaire.

L'ensemble du corpus et des scénarios a été repris de la première campagne. Il convient de noter, cependant, que notre corpus de photographies se révèle être une base documentaire, dans laquelle nous gardons les liens vers les images publiées sur WIKIMEDIA. Le corpus s'en trouve fluctuant : il arrive malheureusement que certaines images ne soient plus accessibles pendant la durée de l'expérience. Il nous a donc fallu trouver une solution pour les images ci-dessous :

1. **I008 :** la photographie représentait la princesse Estelle de Suède à 4 ans et 3 mois. Une image similaire a pu être trouvée.
2. **I036 :** la photographie représentait la princesse Mako d'Akishino à 25 ans et 2 mois. Ce cas-là a été plus problématique à résoudre, car aucune autre photographie comparable n'était libre de droit. La décision a été prise de changer de personnalité, et le choix s'est arrêté sur une photographie de Demi Lovato sur laquelle la jeune femme avait 25 ans et 1 mois.

Certains changements liés à l'outil, plus adapté à notre collecte que MOODLE, sont néanmoins à noter :

— le scénario est attribué aléatoirement et automatiquement à chaque annotateur ;
— désormais, les seules réponses acceptées sont les nombres entiers compris entre 0 et 101, pour éviter les réponses aberrantes ou dans un format incorrect telles que `<3`, `6 mois`, `1 an et demi`, `1000`, `0,33` ou encore `Moins d'un an` ;

— le retour en arrière n'est formellement plus possible, grâce au paramètre adéquat de LimeSurvey ;
— l'expérience n'est plus chronométrée, bien que lors de l'envoi du questionnaire il est précisé qu'elle a une durée approximative de trente minutes [5].

4.1.5 Une première approche des annotations récoltées : comparaison avec la référence

Dans un premier temps, nous voulons avoir un aperçu général des productions des annotateurs. Pour ce faire, nous avons simulé le comportement d'un annotateur moyen, et comparé ses annotations avec la référence – sans comparer les annotateurs entre eux. Le graphique 4.1 (qui regroupe les annotations des deux vagues) affiche ce premier calcul, simple. Il permet de voir l'écart par rapport à la référence, mais aussi si les annotateurs ont eu tendance à rajeunir ou à vieillir les individus. Il est constitué de la façon suivante : pour chaque photographie, dont l'âge réel est matérialisé par un point orange sur le graphique, nous avons calculé la moyenne de toutes les réponses des annotateurs, représentée par un point ainsi que les barres des écarts-types.

Sur ce premier graphique, nous observons déjà deux tendances : une première qui consiste à plutôt vieillir les personnes jeunes, et une deuxième à rajeunir les personnes âgées ; ce constat étaye les conclusions de Clifford et al. (2018). Nous remarquons aussi un autre phénomène intéressant : bien que les annotateurs vieillissent les enfants et adolescents, l'âge estimé reste assez proche de l'âge réel, alors que les estimations pour les personnes âgés ont tendance à être moins précises. Cette observation n'est guère étonnante : il est difficile de se tromper de plusieurs années sur l'âge des bébés ou des enfants, mais il est plus difficile d'estimer précisément l'âge de personnes âgées.

Les barres représentant les écarts-types vont en ce sens : ainsi, si les différentes annotations des premières photographies ont tendance à être plus regroupées autour de l'âge réel, les annotations semblent être plus disparates au fur et à mesure que les personnes présentées sur les photographies sont âgées.

5. Cette estimation a été émise par nous-même, Yann Mathet et Antoine Widlöcher.

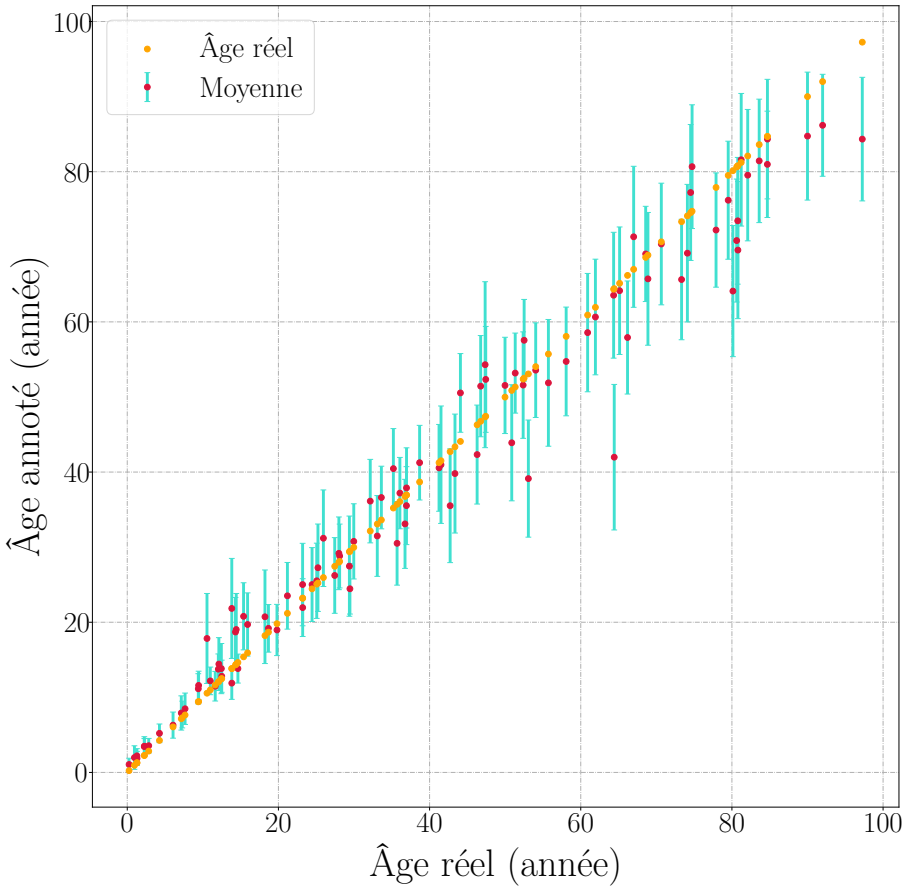


FIGURE 4.1 – Âge moyen trouvé par les annotateurs pour chaque photographie (ensemble des annotations).

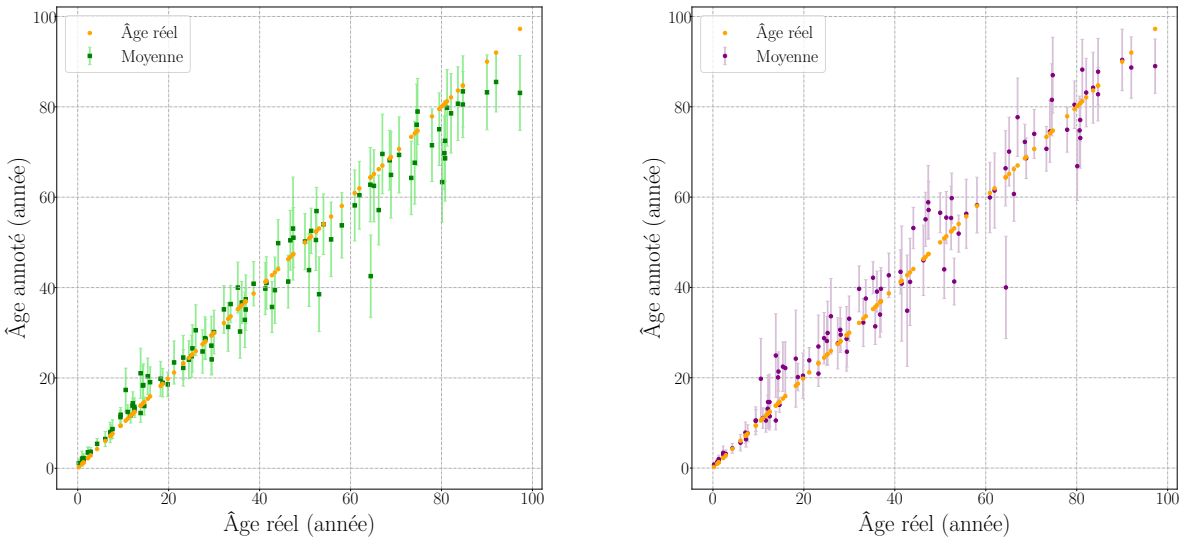
4.2 Analyse des consensualités

Deux observations principales ressortent du graphique de la figure 4.1 :

— il y a un écart, parfois important, entre l'âge de référence et l'annotation moyenne ;
— il y a de la variance dans les annotations, comme nous pouvons le voir grâce aux barres des écarts-types.

Il est à noter que les analyses menées dans cette partie portent sur les annotations de la première vague. Nous n'avons pas voulu regrouper les deux vagues, pour deux raisons principales :

— indépendamment, les deux groupes sont homogènes. Les rassembler induirait d'analyser un groupe hétérogène ;



(a) Vague 1 (b) Vague 2

FIGURE 4.2 – Âge moyen trouvé par les annotateurs pour chaque annotateur selon les vagues.

— l'environnement et le contexte d'annotation étaient différents selon les vagues, et nous ne voulions pas que ces paramètres perturbent et impactent les résultats obtenus.

Sauf indication contraire, les graphiques présentés se lisent de droite à gauche, et du haut vers le bas.

4.2.1 Rang de consensualité *versus* rang de performance

Dans un premier temps, nous avons calculé le rang de consensualité des annotateurs. Le graphique 4.3, généré grâce au logiciel ÉVAL-ANNOT, présente chaque annotateur (matérialisé par un point orange), disposé selon deux axes : sa performance est graduée sur l'axe des abscisses et sa consensualité sur l'axe des ordonnées. Ainsi, plus un annotateur est situé en haut à droite du graphique, moins il est performant et consensuel ; plus il est situé vers le bas et à gauche, plus il est performant et consensuel.

Idéalement, les points devraient former une ligne droite, descendant du haut-droite jusqu'au bas-gauche ; la consensualité et la performance seraient donc directement corrélées. Malheureusement, ce n'est pas le cas sur ce graphique. Si les annotateurs les moins consensuels sont aussi les moins performants (comme l'indiquent la dizaine de point situés

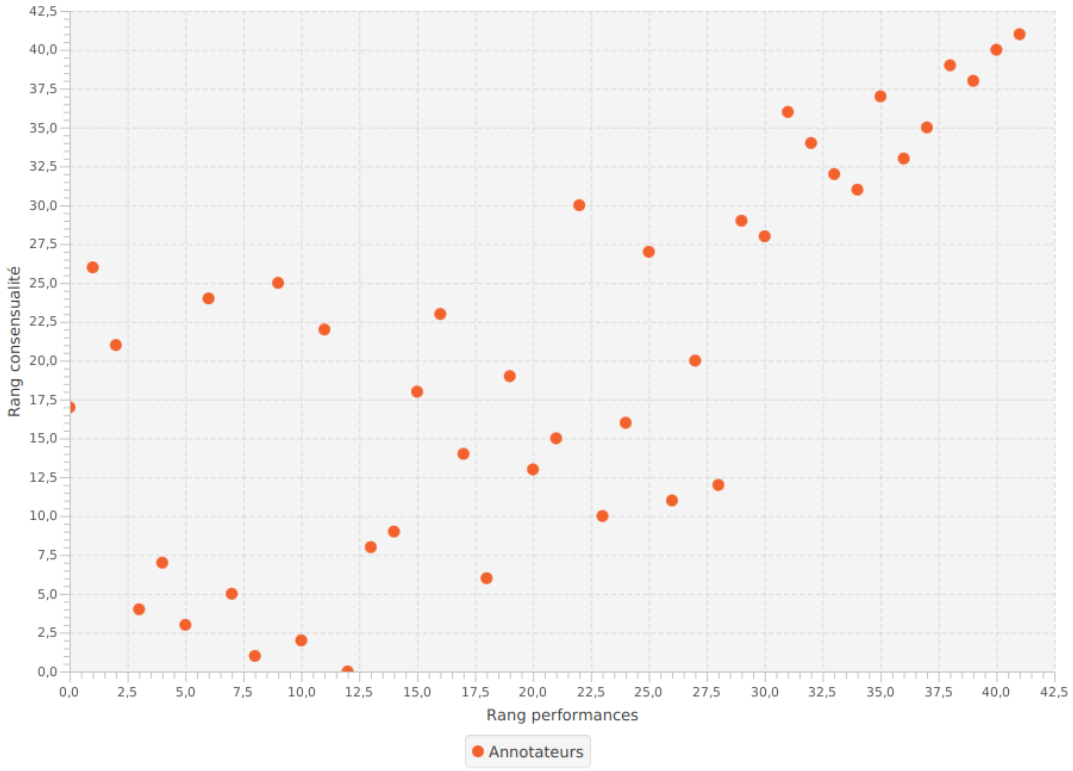


FIGURE 4.3 – Rangs des annotateurs selon leur performance et leur consensualité (*consensualité dynamique*)

dans le haut-droite du graphique), il nous est impossible de formuler une ligne claire pour le reste des annotateurs.

Les points commencent en effet à être plus dispersés : les meilleurs annotateurs ont une consensualité assez moyenne, tandis que des annotateurs moins performants se rapprochent du groupe constitué par les annotateurs les plus consensuels. Toutefois, parmi ce dernier groupe, nous remarquons que certains des annotateurs les plus consensuels ont des performances qui semblent tout à fait acceptables.

De ce graphique, nous ne pouvons donc tirer qu'une première conclusion : les annotateurs les moins consensuels seraient les moins performants.

4.2.2 Retirer les annotateurs les moins consensuels

À la suite du graphique 4.3, une question peut se poser : en l'absence de référence, si nous retirons les annotateurs les moins consensuels, la performance de groupe s'améliore-t-elle ? Dans cette partie, pour répondre à cette question, nous réalisons une série d'expériences pour comprendre ce qui se passe lorsque l'on retire progressivement des annotateurs en fonction de leur manque de consensualité avec les autres annotateurs.

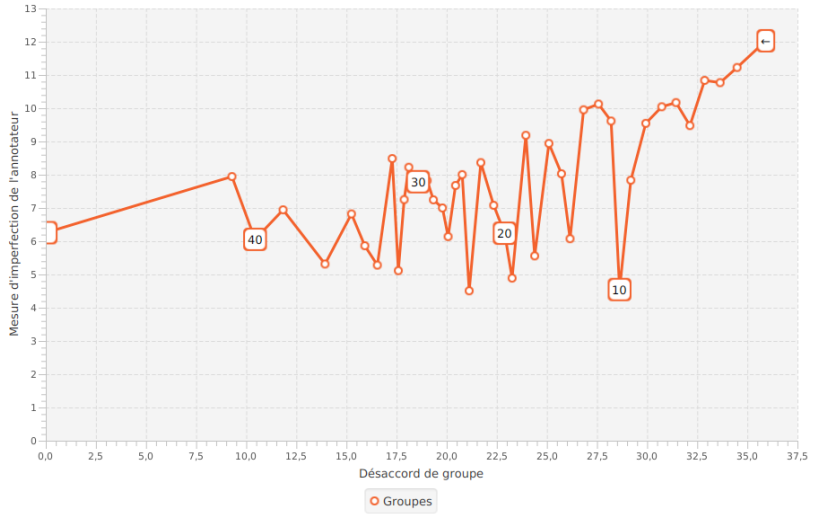
Sur les graphiques présentés dans cette partie, le premier point en haut à droite correspond au groupe comprenant tous les annotateurs, le deuxième point le même groupe auquel on a retiré l'annotateur jugé le moins consensuel, et ainsi que de suite, jusqu'au dernier point, tout à gauche, ne contenant que l'annotateur le plus consensuel.

4.2.2.1 Désaccord de groupe et évolution de la performance des annotateurs

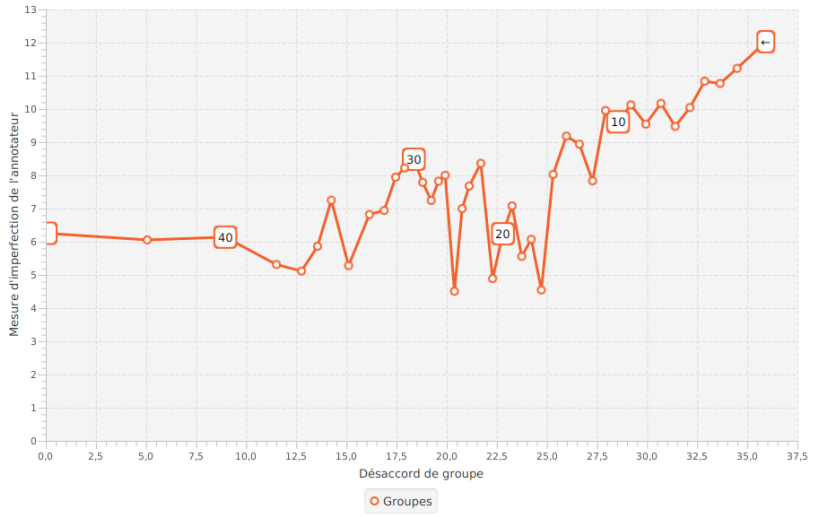
Dans un premier temps, nous nous intéressons à l'évolution du désaccord de groupe et de la performance de l'annotateur, ainsi qu'aux différences que les consensualités initiale et dynamique entraînent sur cette évolution. Les figures 4.4 montrent l'évolution du désaccord de groupe (sur l'axe des abscisses) et les performances individuelles de chaque annotateur (sur l'axe des ordonnées).

Sur les deux figures, nous observons que les annotateurs les moins performants (déjà repérés sur le graphique 4.3) sont retirés dès les premières itérations du calcul. Le désaccord de groupe s'améliore quelque peu. La courbe de la consensualité dynamique (graphique 4.4b) semble avoir un meilleur comportement. En effet, cette consensualité permet de retirer principalement les annotateurs les moins performants jusqu'au rang 15, alors que pour la consensualité initiale (graphique 4.4a) cette tendance s'arrête au rang 8. De plus, la courbe de la consensualité initiale est très peu monotone, cette consensualité retirant souvent un annotateur peu performant, puis un annotateur très performant.

Dans les deux cas, les annotateurs les plus performants sont retirés rapidement dans les itérations.



(a) En consensualité initiale



(b) En consensualité dynamique

FIGURE 4.4 – Imperfection de l'annotateur en retirant l'annotateur le moins consensuel à chaque fois

4.2.2.2 Désaccord de groupe et évolution de la performance de groupe

Le graphique 4.5 reprend le même principe que le graphique 4.4b, si ce n'est qu'au lieu d'afficher la performance individuelle, il affiche l'imperfection de groupe. Dans un premier temps, nous remarquons une lente baisse de l'imperfection de groupe : le retrait des annotateurs les moins consensuels permet de baisser l'imperfection générale. Toutefois, après le retrait du seizième annotateur (en partant de la droite), nous observons une

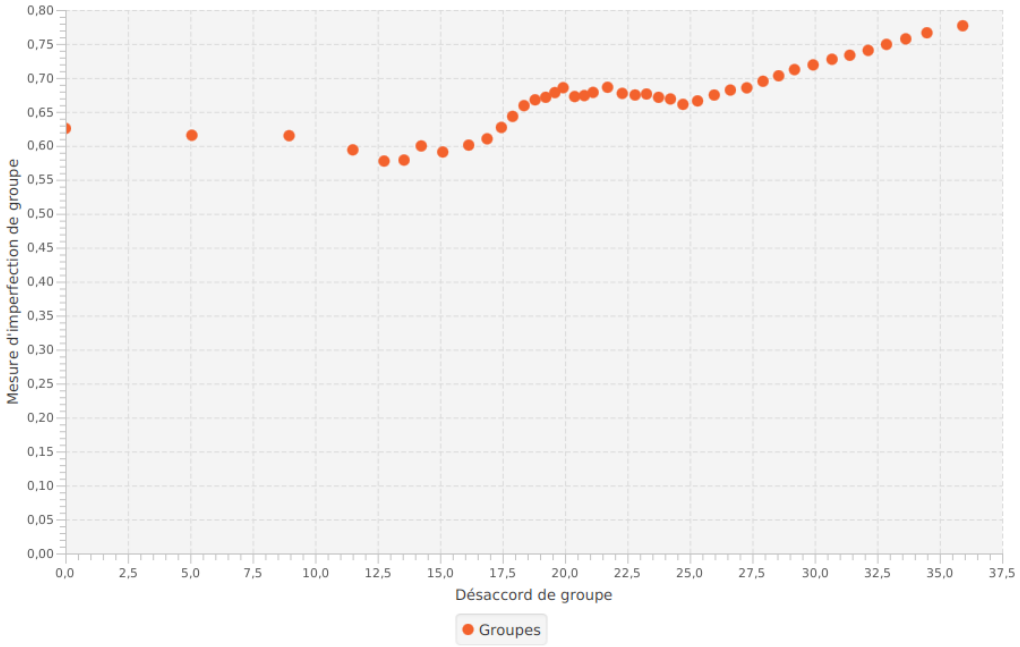


FIGURE 4.5 – Imperfection de groupe en retirant l'annotateur le moins consensuel à chaque fois (consensualité dynamique).

petite remontée de l'imperfection de groupe. D'autres remontées de l'imperfection sont à noter sur la courbe. À la lumière du graphique 4.4a, nous pouvons lier ces remontées de l'imperfection de groupe aux retraits des annotateurs les plus performants. Malgré ces retraits, l'imperfection de groupe des annotateurs les plus consensuels reste acceptable : nous pourrions envisager de ne garder que les 10 % des annotateurs les plus consensuels et obtenir tout de même de bonnes annotations.

4.2.3 Distinguer les consensualités initiale et dynamique

Pour étayer cette dernière hypothèse, nous avons voulu comparer l'imperfection moyenne des annotateurs les plus consensuels trouvés selon chaque type de consensualité. Cette comparaison se retrouve sur le graphique 4.6. Pour un pourcentage d'annotateurs les plus consensuels, nous avons calculé la moyenne de leur imperfection (autrement dit, plus leur moyenne tend vers 0, meilleure est leur performance de groupe).

Une première tendance se dégage immédiatement du graphique : les groupes d'annotateurs les plus consensuels selon la consensualité dynamique (courbe avec les carrés)

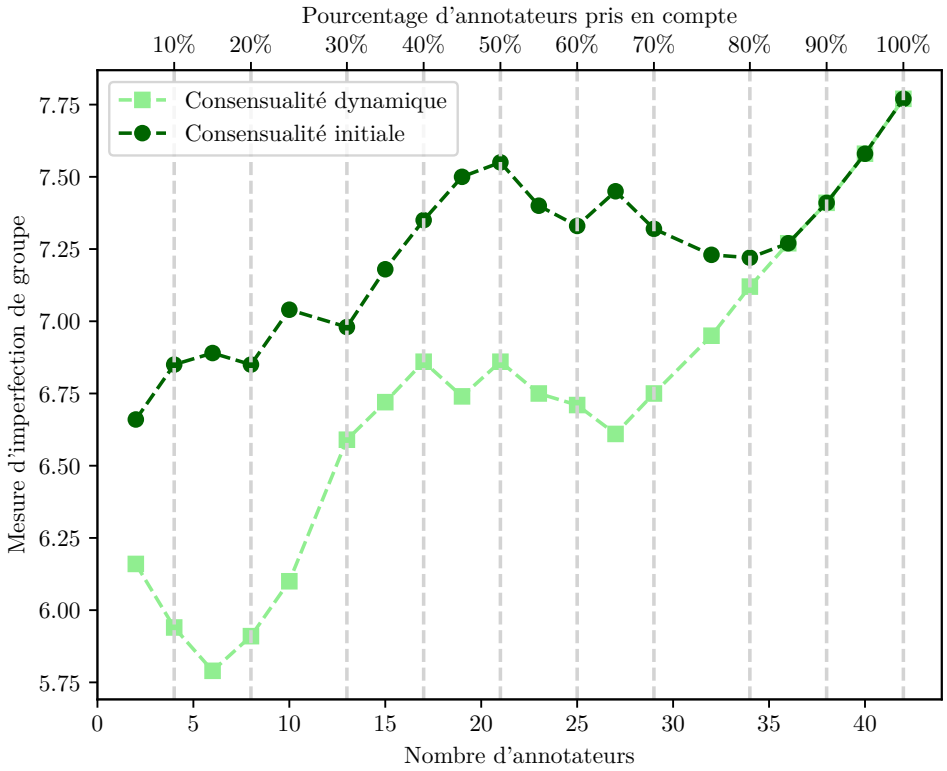


FIGURE 4.6 – Moyenne des imperfections pour les annotateurs les plus consensuels

obtiennent de meilleures moyennes que les groupes à nombre d'annotateurs équivalant trouvés avec la consensualité initiale. Bien que les courbes ne soient pas strictement monotones, nous observons aussi des meilleures performances de groupe lorsque le nombre d'annotateurs sélectionnés est moins élevé.

Ainsi, à la lumière des précédentes analyses et de ce graphique, il faudrait *a priori* privilégier un groupe restreint des annotateurs les plus consensuels, sélectionnés grâce à une consensualité dynamique, pour espérer recueillir des annotations de meilleure qualité. Bien entendu, d'autres expériences doivent être menées pour confirmer cette hypothèse.

Nous pouvons toutefois citer d'ores et déjà des expériences comparables menées dans le cadre de la myriadisation, où le fait de recueillir des annotations par une foule d'annotateurs variés conduit souvent à un questionnement quant à la qualité des données produites. Ainsi, PASSONNEAU et al. (2012) identifient les annotateurs dont les annotations se distinguent le plus des autres annotateurs pour ensuite les retirer de l'expérience ; l'accord inter-annotateurs s'en voit alors nettement amélioré, suffisamment pour indiquer

une annotation fiable. Dans Inel et al. (2014), les auteurs préfèrent étudier les désaccords d'annotations afin de repérer des items ambigus, mais les désaccords leur permettent aussi de repérer les annotateurs qui se détachent trop des autres.

4.2.4 Tester l'homogénéité de la consensualité

En approfondissant la notion de consensualité et de ces apports pour l'établissement d'une référence, une nouvelle question se pose : la consensualité est-elle homogène sur l'ensemble du corpus, ou varie-t-elle selon les parties du corpus ? La réponse à cette question est notamment intéressante lorsque nous avons à disposition une référence partielle, sur une partie du corpus. Si la consensualité est homogène et stable, nous pourrions alors envisager de compléter la référence sur la totalité du corpus.

Pour mesurer l'homogénéité de la consensualité, nous avons divisé le corpus en sous-corpus. Pour des soucis de représentativité des résultats, nous avons décidé de créer des sous-corpus de cinquante images chacun, tirées aléatoirement à chaque fois. Ensuite, nous avons calculé le degré de consensualité et le rang des annotateurs pour chaque sous-corpus. Idéalement, ces deux indices devront rester assez stables selon les sous-corpus, ce qui sous-entendrait que la consensualité est homogène.

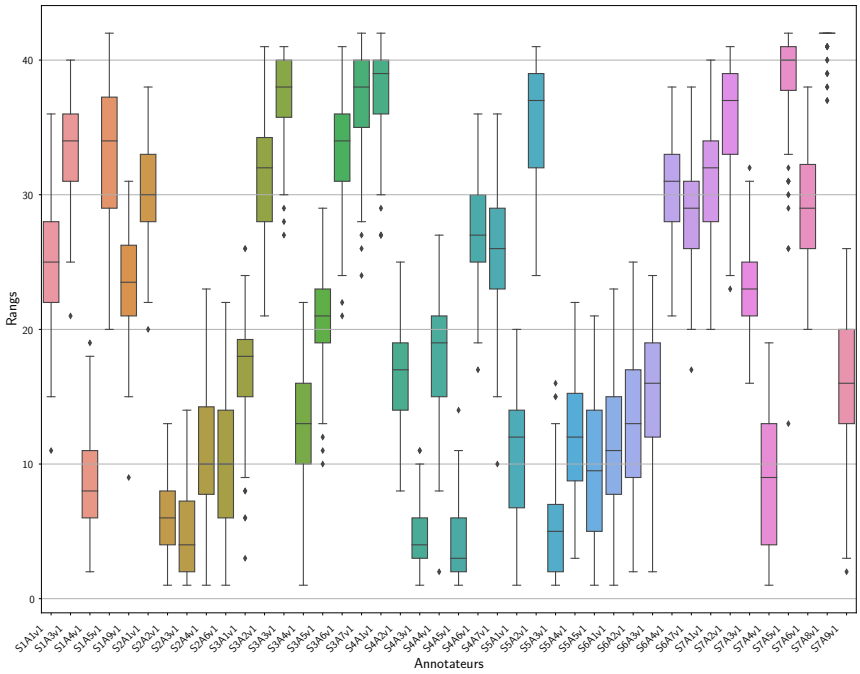
Concernant l'analyse, nous avons principalement étudié la distribution des rangs et des degrés selon les sous-corpus (c'est-à-dire le minimum, le maximum et la moyenne). Nous avons aussi utilisé la moyenne des variances pour chaque annotateur. Pour la partie visualisation de données, nous avons opté pour les boîtes à moustaches, permettant de rendre compte de l'essentiel des indices utilisés, tout en restant lisibles lorsque nous étudions tous les annotateurs. Quand nous nous focalisons sur quelques annotateurs, nous préférons alors des diagrammes en violon. Cette visualisation reprend le principe de la boîte à moustache, tout en permettant une compréhension plus fine de la distribution.

4.2.4.1 Étude de l'homogénéité de la consensualité

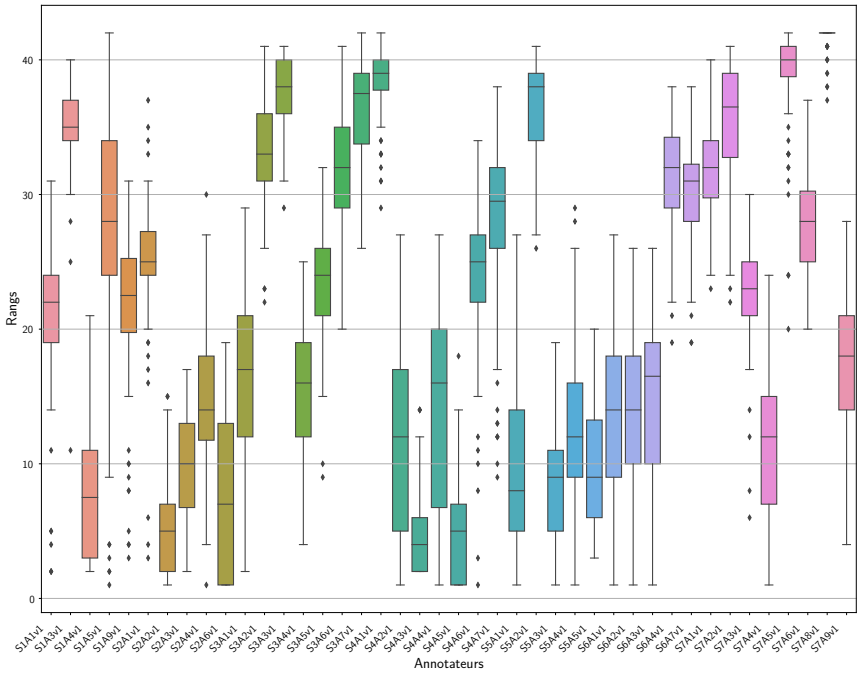
Analyse basée sur les rangs

Dans un premier temps, nous avons calculé les indicateurs de position de base (minimum, médiane, maximum) pour l'ensemble des annotateurs, selon les deux modalités de

la consensualité ; les boîtes à moustaches correspondantes sont sur les figures 4.7a et 4.7b.



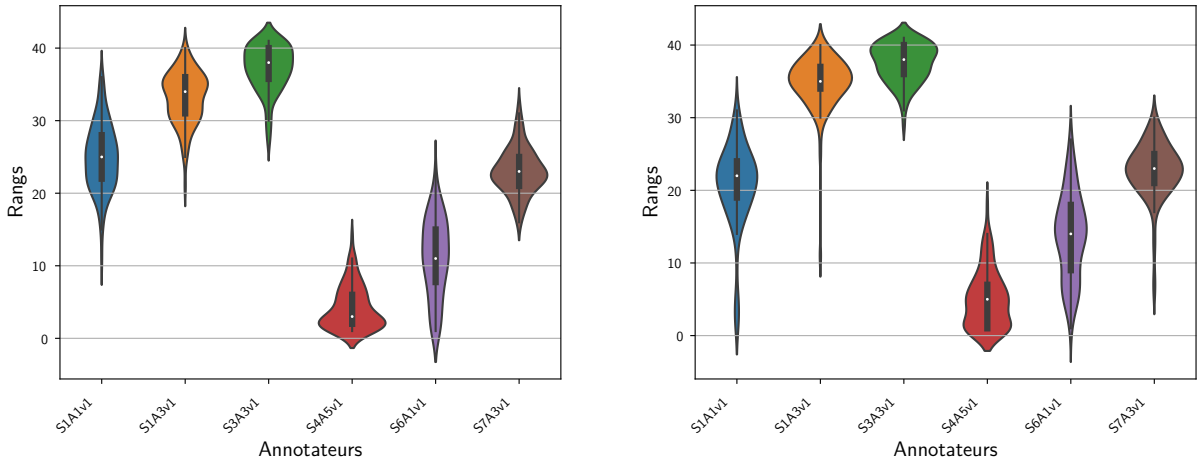
(a) Consensualité initiale



(b) Consensualité dynamique

FIGURE 4.7 – Boîtes à moustache représentant la distribution des rangs pour chaque annotateur, basée sur un échantillon de 100 sous-corpus de 50 photographies aléatoires.

La consensualité semble être assez homogène, particulièrement selon la consensualité initiale. En ce sens, pour confirmer cette première observation, nous avons calculé la moyenne des variances selon les types de consensualité : celle pour la consensualité initiale est à 19,32, tandis que celle de la dynamique est à 28,72. Certains annotateurs se distinguent néanmoins, soit par un rang de consensualité qui change peu selon les sous-corpus (par exemple `S4A3`, `S4A5`, `S7A3` ou encore `S7A8`), soit, à l'inverse, une grande disparité dans les rangs qu'ils peuvent atteindre (notamment `S3A2`).



(a) Consensualité initiale (b) Consensualité dynamique

FIGURE 4.8 – Exemple de diagrammes en violon représentant la distribution des rangs pour six annotateurs de distribution de la consensualité pour 100 sous-corpus de 50 photographies aléatoires.

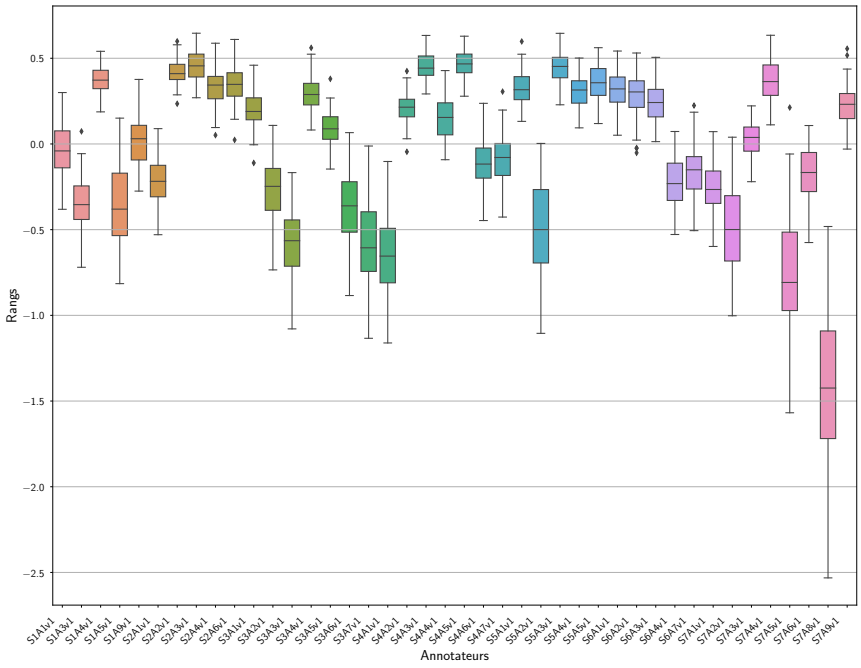
Plus spécifiquement, nous avons regardé la distribution pour six annotateurs, choisis après observation de l'ensemble des annotateurs et sélectionnés pour constituer un sous-ensemble représentatif de la diversité ; les diagrammes en violon pour leurs distributions se retrouvent sur les graphiques de la figure 4.8. Ces six annotateurs sont plus ou moins consensuels : deux ont un rang moyen entre [30;40[, deux entre [20;30[, un entre [10;20[et enfin un dont les rangs sont toujours en-dessous de 10.

Deux tendances ont l'air de se dégager, davantage marquées pour la consensualité dynamique. La première concerne les annotateurs extrêmes, soit très, soit très peu consensuels : ceux-là conservent des rangs généralement proches, quels que soient les sous-corpus. La deuxième observation concerne les annotateurs moyennement consensuels. La distribution de leurs rangs varie souvent selon les sous-corpus. Leur consensualité semblerait

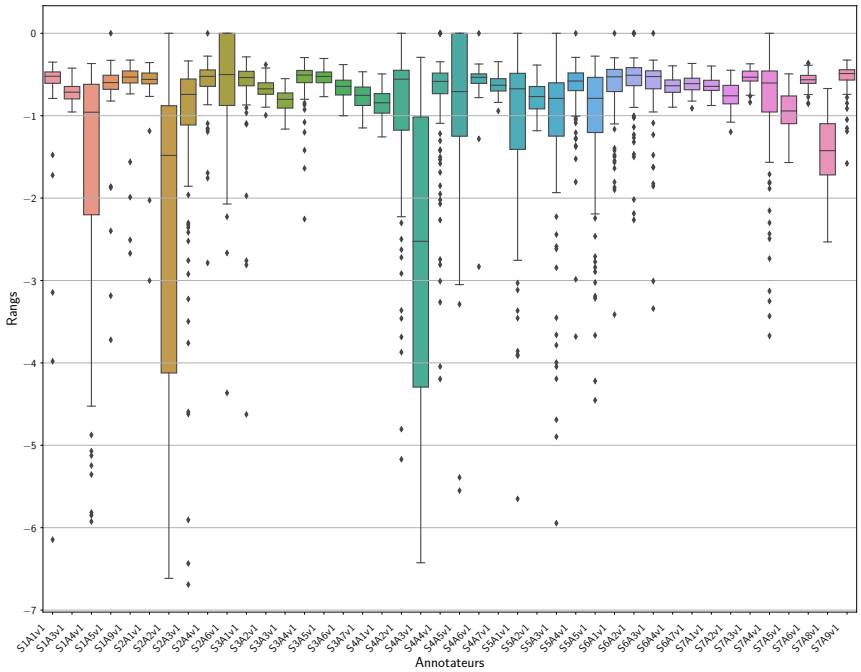
dépendre des sous-corpus, et des items pris en considération.

Analyse basée sur les degrés de consensualité

Nous avons ensuite calculé la dispersion des degrés de consensualité selon les sous-corpus, pour chaque type de consensualité et tous les annotateurs (figures 4.9a et 4.9b).



(a) Consensualité initiale



(b) Consensualité dynamique

FIGURE 4.9 – Boîtes à moustache représentant la distribution des degrés de consensualité pour chaque annotateur, basée sur un échantillon de 100 sous-corpus de 50 photographies aléatoires.

La première observation recouvre celle faite avec les rangs, c'est-à-dire que la consensualité initiale semble rester homogène selon les sous-corpus : cela se voit avec les boîtes symétriques et à la quasi-absence de points signifiant des valeurs dites aberrantes. À l'inverse, bien que les boîtes soient plus petites, et que la plupart paraissent symétriques, la consensualité dynamique semble plus sujette à la variabilité selon les sous-corpus. Notons que l'éparpillement (ou non) des boîtes est lié à la manière de calculer les deux types de consensualité : les valeurs pour la consensualité dynamique sont généralement dans les mêmes intervalles.

4.2.4.2 Quelle doit être la taille de l'échantillon de référence ?

Une fois la stabilité observée de la consensualité selon les sous-corpus, nous souhaitons à présent examiner une nouvelle question : à partir de quelle taille de sous-corpus la consensualité se stabilise-t-elle ? Ou, *a minima*, à partir de quel seuil la consensualité donne-t-elle déjà des informations utilisables/satisfaisantes ? Pour étudier cette question, nous procéderons en deux temps :

1. Distribution des rangs selon la taille des sous-corpus : regarder cette distribution permet d'observer à partir de quelle taille la consensualité se stabilise relativement bien.
2. Évolution de l'imperfection de groupe pour les annotateurs les plus consensuels : la deuxième expérience consiste à calculer l'imperfection de groupe pour les n% annotateurs les plus consensuels, en moyenne, selon la taille des sous-corpus, et observer ensuite l'évolution de cette imperfection.

Distribution des rangs selon la taille des sous-corpus

Pour cette première expérience, nous avons tiré 100 sous-corpus aléatoires, de tailles différentes (respectivement $\{10; 15; 20; 25; 30; 35; 40; 45; 50\}$). Ensuite, nous avons observé l'évolution de la distribution des rangs obtenus selon les tailles de sous-corpus. Pour chaque taille considérée, nous avons calculé la variance des rangs de chaque annotateur, puis calculé la moyenne de ces variances. Nous avons ensuite choisi de visualiser ces données sous forme d'un graphique en courbes, avec deux courbes, pour les types de consensualité. Le résultat se retrouve dans le graphique de la figure 4.10.

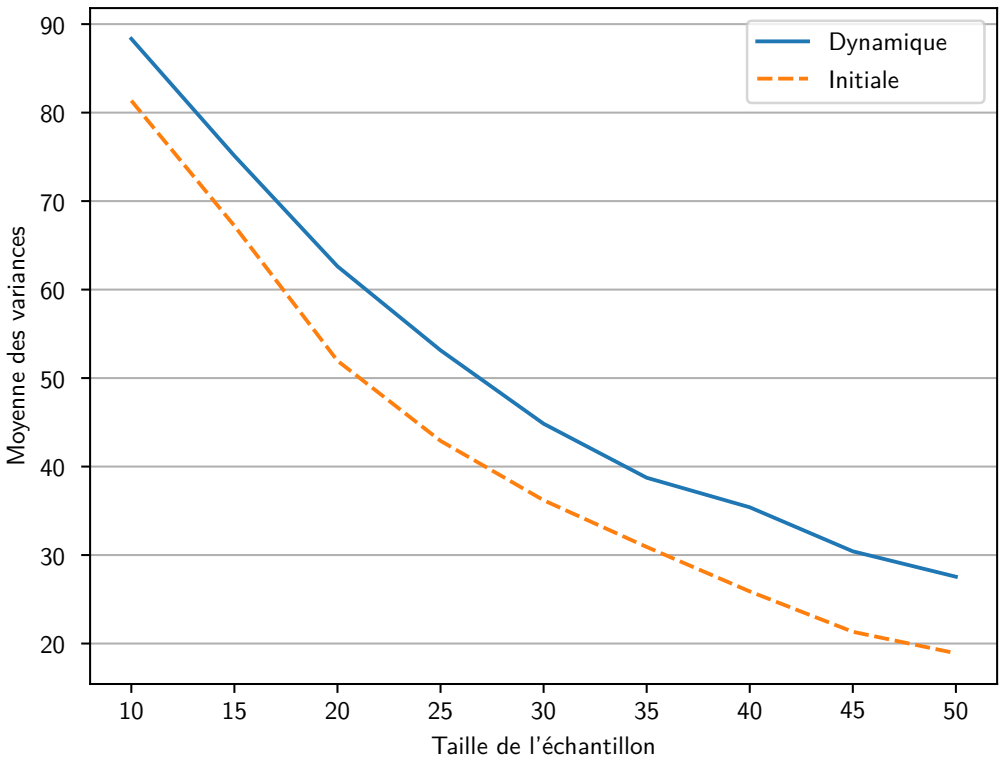
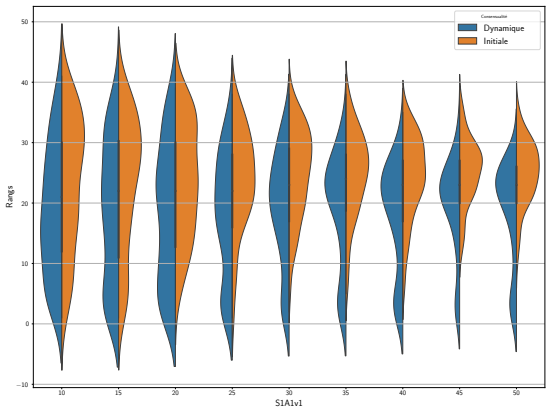


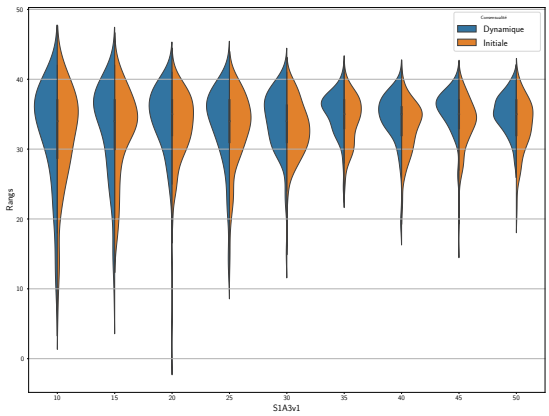
FIGURE 4.10 – Évolution de la variance des moyennes des rangs selon la taille des sous-corpus, en fonction des consensualités initiale et dynamique.

Le graphique est simple à analyser. Avec une taille d'échantillon de 10% du corpus global, la moyenne des variances est importante, entre 80 et 90. Cela sous-entend une grande variabilité dans les rangs : un annotateur peut être amené à être considéré comme très consensuel, mais aussi très peu consensuel. Par exemple, l'annotateur `S1A1` est parfois classé dans les cinq premiers les plus consensuels, mais dans aussi les 10 derniers. Dans cet état actuel des choses, la consensualité n'est donc pas utilisable. Puis la moyenne diminue progressivement, au fur et à mesure que la taille de l'échantillon considéré augmente. Les deux courbes suivent la même tendance. Ce graphique montre surtout que, plus la taille de l'échantillon augmente, plus le calcul des rangs de consensualité devient stable et exploitable.

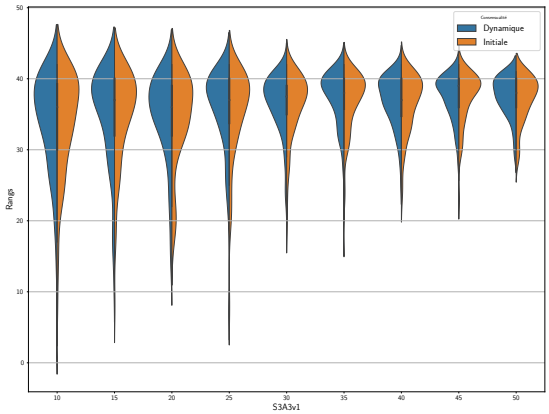
Plus spécifiquement, nous avons aussi regardé six annotateurs, les mêmes que dans la partie précédente, à savoir `S1A1`, `S1A3`, `S3A3`, `S4A5`, `S6A1` et `S7A3`. L'évolution de leur distribution se retrouve sur les graphiques en figure 4.11, où la distribution de leurs rangs est présentée selon les deux types de consensualités.



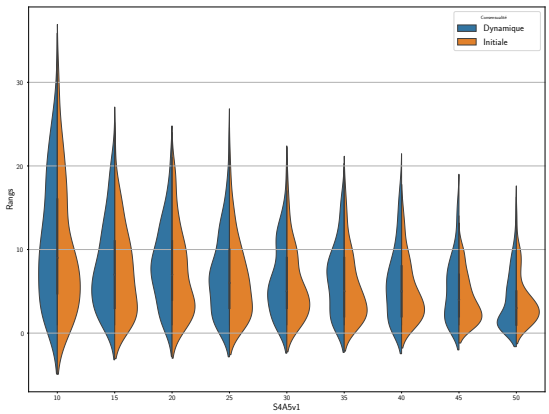
(a) Annotateur S1A1



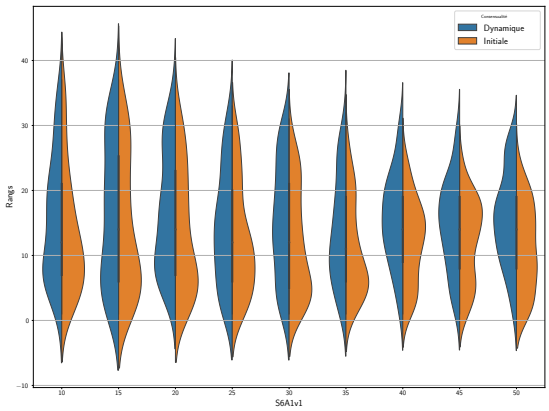
(b) Annotateur S1A3



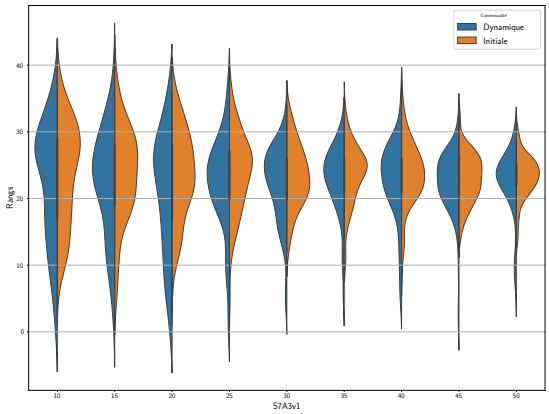
(c) Annotateur S3A3



(d) Annotateur S4A5



(e) Annotateur S6A1



(f) Annotateur S7A3

FIGURE 4.11 – Exemples d'évolution de la distribution des rangs de consensualité pour six annotateurs choisis.

La principale observation à souligner concerne les annotateurs les moins consensuels (comme `S1A3` et `S6A1`, voire `S7A3`) : ceux-ci se comportent d'une manière relativement similaire quels que soient les sous-corpus. Déterminer les annotateurs les moins consensuels peut donc être réalisé assez rapidement, même sur un petit échantillon du corpus. *A contrario*, pour les annotateurs *a priori* les plus consensuels (comme le `S4A5`), il est peu aisé de tirer des conclusions rapidement, même si une certaine stabilité semble apparaître aux alentours d'un sous-corpus de 30% du corpus global.

Pour les annotateurs plus dans la moyenne, comme `S1A1` ou `S6A1`, il est difficile de pouvoir tirer des conclusions, et ce même avec un échantillon de corpus grand. Nous remarquons toutefois des valeurs parfois aberrantes, comme en témoignent les longues traînes présentes sur les graphiques. Il est important d'avoir conscience que nous pouvons tomber sur des classements de consensualités qui ne reflètent pas totalement la réalité.

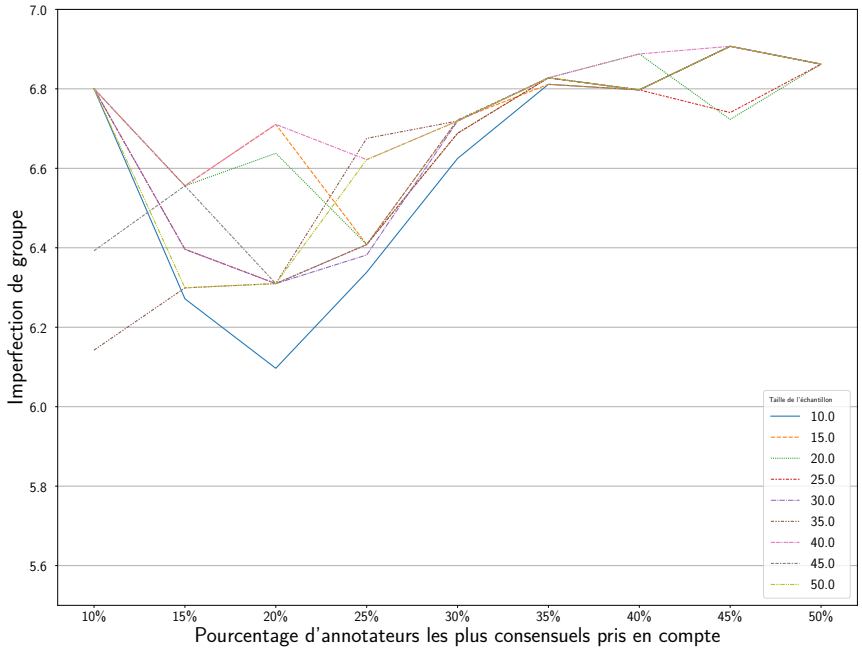
Évolution de l'imperfection de groupe pour les annotateurs les plus consensuels

Une fois que nous avons vérifié que la consensualité était une notion robuste et ne dépendait que peu des sous-corpus, nous avons souhaité regarder si les annotateurs jugés comme les plus consensuels au fil des itérations selon la taille des sous-corpus étaient aussi performants. Pour ce faire, nous avons réalisé les graphiques présentés dans la figure 4.12. Ils montrent, pour chaque taille de corpus (symbolisée par une courbe de couleur et style différents), l'évolution de l'imperfection de groupe pour les n % annotateurs les plus consensuels, en moyenne. À l'instar du graphique 4.6, plus la courbe est basse, meilleure est la performance.

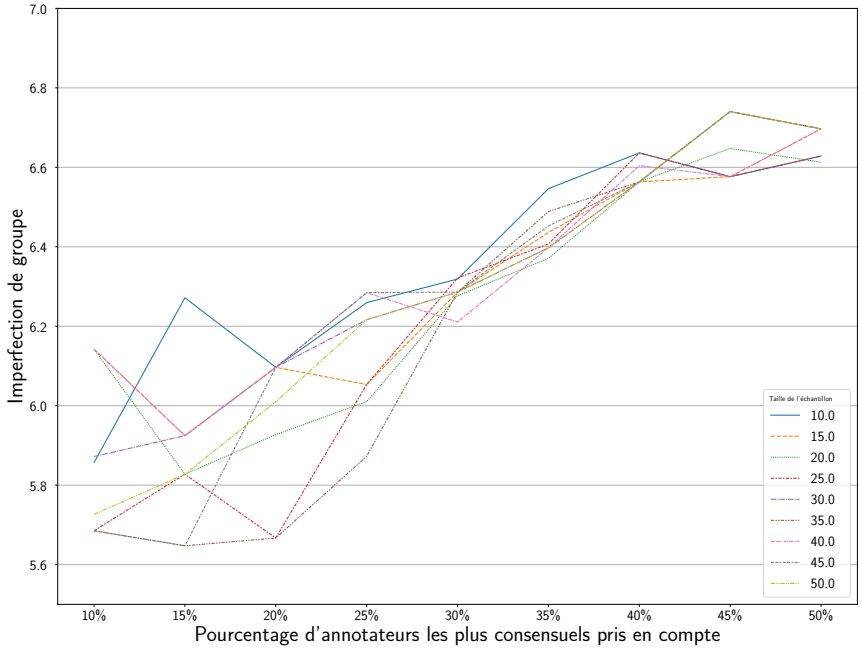
Ces graphiques sont intéressants à plusieurs égards. D'une part, ils sont cohérents avec les résultats de la partie précédente, notamment ceux issus de l'analyse de la figure 4.6 :

— privilégier un groupe restreint d'annotateurs contribue à obtenir une meilleure performance de groupe ;
— les annotateurs sélectionnés selon la consensualité dynamique sont, de manière générale, plus performants que ceux choisis selon l'initiale.

Ces tests confirment l'intérêt d'utiliser la consensualité dynamique pour déterminer les annotateurs qui sont les plus consensuels et dont la performance est bonne. La stabilité



(a) Consensualité initiale



(b) Consensualité dynamique

FIGURE 4.12 – Évolution de la performance de groupe pour les n% annotateurs les plus consensuels (en moyenne).

des rangs n'est pas gage de fiabilité dans le cas de la consensualité initiale.

Observations générales

Sans surprise, plus la taille de l'échantillon augmente, plus le calcul des rangs devient stable et, par extension, exploitable. Toutefois, si la consensualité initiale donne des résultats plus homogènes selon les tailles que la consensualité dynamique, la dernière expérience menée montre clairement que la dynamique reste encore à privilégier pour sélectionner les annotateurs les plus consensuels, qui sont aussi souvent performants.

En introduction de cette sous-partie, nous avons posé la problématique suivante : si la consensualité est homogène selon les sous-corpus, est-il pertinent d'établir une référence partielle sur une fraction du corpus, de déterminer les annotateurs les plus consensuels et de les sélectionner ? Les premiers résultats obtenus et présentés dans cette partie semble prometteurs pour répondre favorablement à cette question. Selon le graphique 4.12b, plus que la taille de l'échantillon sur laquelle est basée la sélection des annotateurs, c'est surtout le pourcentage d'annotateurs qui semble essentiel.

4.3 Influence de l'ordre des items

Dans cette section, nous présentons la démarche que nous avons suivie pour analyser la présence et l'influence du biais de l'ordre des items sur les annotations. Idéalement, avec un accès à la référence — comme cela a été notre cas pour cette expérience —, nous connaissons la performance précise des annotateurs. Sans référence, l'analyse sera un peu plus dégradée, moins fine ou plus limitée. Cette partie s'articule autour de l'accessibilité ou non à une référence :

— grâce à la référence, nous souhaitons **observer** la présence de variations possiblement dues au biais, et **quantifier** cet impact ;
— sans la référence, il convient de nous donner des moyens pour **détecter** la présence du biais, puis de le **corriger**.

4.3.1 Avec un accès à la référence

4.3.1.1 Observer à un niveau global

Dans un premier temps, nous calculons la moyenne des erreurs absolues pour chaque scénario. Il y a deux manières de voir cette erreur : soit de manière individuelle en faisant la moyenne de toutes les erreurs commises par tous les annotateurs, soit de façon collective en regardant l'erreur du groupe (obtenue à partir de la moyenne des annotateurs). Formellement :

— la moyenne de l'**erreur individuelle** (formule 4.1) est

$$EI = \frac{1}{N}\sum_{i=1}^{N}(\frac{1}{G}\sum_{a=1}^{G}|x_{i,a} - \mu_i|) \tag{4.1}$$

— la moyenne de l'**erreur collective** (formule 4.2) est

$$EC = \frac{1}{N}\sum_{i=1}^{N}|(\frac{1}{G}\sum_{a=1}^{G}x_{i,a}) - \mu_i| \tag{4.2}$$

Où N est le nombre total de photographies, G le nombre d'annotateurs, $x_{i,a}$ l'annotation de l'annotateur a sur l'image i, et μ_i la référence de cette même image. Ainsi, pour un portrait dont la référence est 50 ans, si un annotateur estime 45 ans et un autre 52, leur moyenne d'erreur individuelle serait 3,5, alors que la moyenne d'erreur collective serait 1,5.

Les résultats se retrouvent dans le tableau 4.3. Il y a peu de différence de moyennes entre les deux types d'erreur, sauf pour les scenarios **S1** et S5 — bien que cela ne change pas les observations que nous pouvons en tirer. Les modestes écarts entre les erreurs individuelles et collectives indiquent notamment que l'erreur faite par les annotateurs est sensiblement la même dans un groupe considéré. Les erreurs sembleraient liées à un groupe, à un potentiel biais que les annotateurs auraient tous en commun.

Plus spécifiquement, les annotateurs ayant eu le scénario **S1** ont eu tendance à estimer l'âge des personnes à plus ou moins 4 ans — la meilleure performance. À la lecture des autres valeurs du tableau, nous remarquons que les annotateurs ayant eu un scénario avec un aléatoire très présent (**S3**, **S6** et **S7**) ont moins bien estimé l'âge, se trompant d'au moins 6 ans. Les annotateurs des scénarios avec quelques variations (**S2**, S4 et S5) ont des performances moyennes.

Scénarios	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**
Description	Chrono.	Chrono. Aléa. dans tranche	Aléa.	Chrono. Échange tranches	**S2** + **S4**	(quasi-) Aléa.	(quasi-) Aléa. +
Erreur individuelle	4,11	4,62	6,16	5,06	5,33	6,02	6,52
Erreur collective	2,04	3,36	4,57	3,64	4,20	4,84	4,56

TABLE 4.3 – Moyennes des erreurs absolues en fonction du scénario (en année).

À la suite de ces premiers résultats, une première question se pose naturellement : ces différences de moyennes sont-elles réellement significatives ? Pour tenter de répondre à cette question, nous procédons de manière empirique [6] à un ré-échantillonnage aléatoire sur l'ensemble des groupes possibles pour calculer la significativité statistique des résultats.

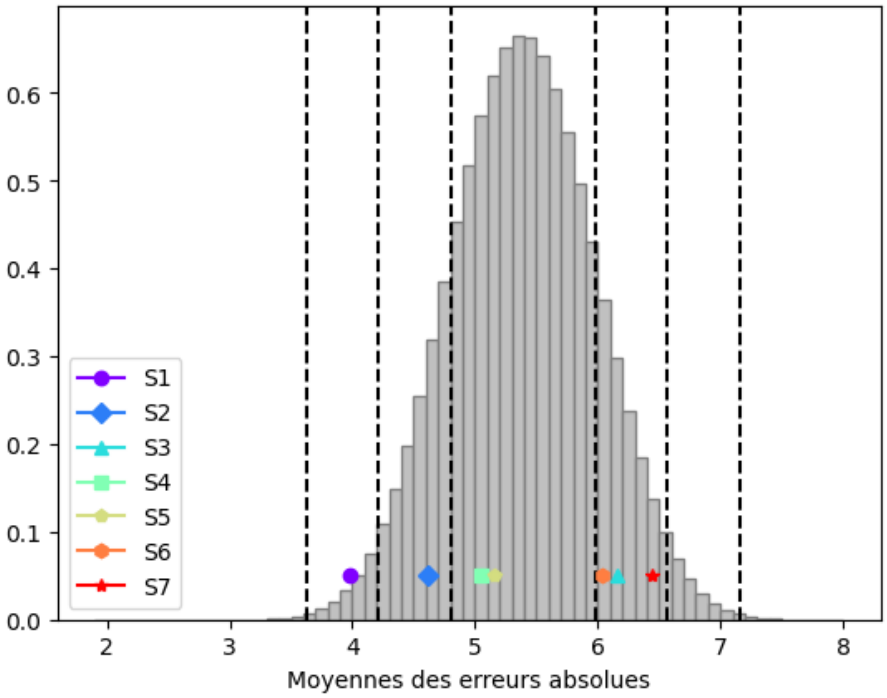


FIGURE 4.13 – Distribution des moyennes pour 6 parmi 52 groupes (erreur individuelle).

La distribution des moyennes obtenue se retrouve sur le graphique de la figure 4.13. La hauteur des marqueurs indiquant les différentes moyennes d'origine de chaque scénario n'est pas significative, c'est leur emplacement sur l'axe horizontal qui importe véritablement. La moyenne obtenue avec les annotateurs de **S1** semble significative : elle se situe

6. Il existe évidemment des outils statistiques pour vérifier cela de manière plus rigoureuse — sous réserve que le modèle corresponde —, mais nous nous contenterons ici de cette méthode qui donne des résultats déjà significatifs.

dans l'intervalle de confiance de 99%, c'est-à-dire qu'il est rare d'avoir cette moyenne selon l'échantillonnage considéré. Les moyennes des scénarios avec la part d'aléatoire la plus importante (à savoir le **S3**, **S6** et le **S7**) se situent aussi dans un intervalle de confiance qui indique que leur moyenne est peu commune. Il en va de même pour le scénario **S2**. Enfin, les moyennes des scénarios S4 et S5 sont dans la norme. Les écarts entre les différents scénarios ne sont donc pas dus au hasard et semblent indiquer un biais.

Grâce aux calculs des moyennes des écarts absolus entre les annotations et les références, nous avons pu déceler un premier impact du biais de l'ordre de présentation des items : les annotateurs ayant eu un scénario « facile », dans un ordre chronologique, estiment mieux les âges. À l'inverse, les annotateurs avec beaucoup d'aléatoire se trompent plus souvent. Cette méthode pour déceler un biais de façon générale ne peut s'appliquer qu'en présence d'une référence pour calculer les écarts à la « vérité ».

4.3.1.2 S'intéresser au local

Les observations précédentes ont été réalisées à un niveau global, c'est-à-dire en prenant en compte tous les scénarios et l'entièreté des tranches. Le biais peut néanmoins intervenir plus localement et avoir un impact plus discret sur la globalité. Une lecture plus attentive des différences d'annotations entre les scénarios s'impose alors. Nous voulons particulièrement examiner les inversions de tranches ayant eu lieu dans les scénarios S4 et S5.

Nous étudions l'impact du biais sur les tranches inversées dans les scénarios S4 et S5. Pour ce faire, nous les comparons avec, respectivement, **S1** et **S2**, car ces scénarios sont presque identiques : les tranches d'âge et les photographies dans ces dernières sont présentées dans l'ordre chronologique, sauf lorsqu'il s'agit des tranches inversées — pour **S2** et S5. Nous rappelons toutefois que l'ordre de présentation est aléatoire dans les tranches. Il n'y aurait donc que le biais de l'ordre des items qui viendrait perturber localement l'annotateur. Le tableau 4.4 présente les moyennes des erreurs absolues, individuelles et collectives, pour les tranches considérées (à savoir [30;40[, [40;50[, [70;80[, [80;90[).

Pour la tranche [40;50[, que S4 et S5 ont eu juste après la tranche [20;30[, nous observons une erreur très significativement supérieure pour le scénario S4 par rapport à l'ensemble des autres scénarios, et en particulier du **S1**. De façon surprenante, ce biais n'est pas visible lorsque nous comparons **S2** et S5, mais ceci peut s'expliquer par le fait

		Erreur individuelle				Erreur collective			
		S1	**S4**	**S2**	**S5**	**S1**	**S4**	**S2**	**S5**
Tranche	[30;40[	4,7	5,09	3,89	4,31	1,78	2,81	2,96	2,74
	[40;50[	5,79	6,73	5,87	5,83	2,82	4,49	3,68	3,49
	[70;80[	5,82	6,62	6,04	8,96	2,51	4,46	3,33	5,9
	[80;90[	5,73	9,16	6,96	11,5	2,83	8,67	4,68	11,4
	Autres	3,34	4,05	4,05	4,15	1,8	2,84	3,19	3,34

TABLE 4.4 – Moyennes des erreurs absolues sur les tranches inversées, selon le scénario (en année).

que les portraits sont présentés dans l'ordre aléatoire par tranche et donc que la rupture de régularité n'est pas significative. Pour la tranche suivante [30;40[, nous constatons un retour à la normale.

Pour la seconde paire de tranches échangées, la différence entre **S1** et S4 est identique à ce qui est décrit dans le paragraphe précédent. Pour **S2** et S5, nous observons *a contrario* une nette augmentation des erreurs. Nous pouvons essayer d'avancer que le biais a l'air d'impacter très fortement les annotateurs lorsque ceux-ci ont plus d'hésitations.

Le procédé présenté plus haut a besoin, pour être généralisé, que soient pré-identifiées les tranches susceptibles d'apporter un biais. Pouvoir comparer deux (ou plus) groupes d'annotateurs est aussi primordial.

4.3.2 Détecter un biais sans un accès à la référence

La très grande majorité des campagnes d'annotation menées sont réalisées sans accès à la référence — une des questions fondamentales en annotation étant justement de déterminer si l'établissement d'une référence est toujours possible. Dans ce cadre, nous nous interrogeons sur les méthodes que nous pouvons employer pour détecter d'éventuels biais et leur impact sans pour autant être capable de mesurer l'écart à la vérité (du moins, celle qui fait autorité). Si les analyses décrites dans cette partie s'appuient sur une campagne menée spécialement pour détecter un biais, elles peuvent aussi s'appliquer à toute autre campagne.

4.3.2.1 Simuler une référence

À défaut d'avoir une référence avérée, nous pouvons construire une référence à partir des annotations recueillies. Nous considérons alors les annotateurs dans leur(s) ensemble(s), en modélisant un annotateur-type, collectif. Dans le cadre de notre expérience, nous calculons la moyenne des annotations pour chaque image et nous la prenons pour référence. Les analyses réalisées en 4.3.1.1 peuvent donc être reproduites. Les résultats de ces analyses se trouvent dans les tableaux 4.5 et 4.6, présentés ci-dessous :

Scénarios	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**
Erreur individuelle	4,36	3,77	4,62	4,18	3,64	4,02	4,89
Erreur collective	2,25	1,83	2,04	1,89	1,99	1,87	1,7

TABLE 4.5 – Moyennes des erreurs absolues par rapport à l'annotation moyenne en fonction du scénario (en année).

		Erreur individuelle				**Erreur collective**			
		S1	**S4**	**S2**	**S5**	**S1**	**S4**	**S2**	**S5**
Tranche	[30;40[	4,74	4,55	3,48	3,84	1,49	1,16	2,16	1,55
	[40;50[	5,41	5,76	5,32	5,01	1,55	3,12	2,74	2,7
	[70;80[	5,91	6,64	5,3	7,29	2,83	3,46	1,81	2,63
	[80;90[	7,6	5,93	6,07	5,88	5,42	2,52	3,13	4,67
	Autres	4,36	4,18	3,77	3,64	1,98	1,56	1,46	1,5

TABLE 4.6 – Moyennes des erreurs absolues par rapport à l'annotation moyenne sur les tranches inversées, selon le scénario (en année).

Les analyses effectuées par cette méthode donnent des résultats peu satisfaisants. Dans l'ensemble, les moyennes des erreurs absolues par scénarios sont assez similaires entre elles, sauf pour les scénarios **S2** et **S5**. Concernant les erreurs absolues dans les tranches inversées pour les quatre scénarios considérés, aucune tendance ne se dégage véritablement. Simuler une référence ne semble donc pas être une méthode efficace pour détecter un biais, du moins pour cette expérience, car les personnes ayant subi un biais ont en moyenne une erreur relative à leur groupe équivalente à celles n'en ayant pas subi.

4.3.2.2 Comparer des groupes

Même si cette première expérience sans référence n'est pas concluante, nous pouvons toujours regarder plus précisément les différences d'estimation qu'il y a entre **S1** et **S4**.

Ainsi, nous calculons, pour chaque image, la moyenne de l'âge que les annotateurs du scénario S4 ont estimé et nous soustrayons la moyenne du **S1**. Ainsi, si les annotateurs du S4 ont eu tendance à rajeunir les personnes par rapport à **S1**, la valeur obtenue serait négative ; à l'inverse, s'ils ont eu tendance à vieillir, la valeur serait positive.

Le graphique de la figure 4.14 présente les résultats. Le lecteur ne s'étonnera pas de la tendance présente sur le graphique qui consiste à se tromper de façon plus importante sur les personnes âgées : il s'agit d'un biais couramment repéré dans de telles expériences Clifford et al., 2018 ; Vestlund et al., 2009 [7]. Toutefois, quelle que soit l'amplitude des erreurs, il s'agit surtout de la tendance des erreurs à n'être que d'un seul côté de l'axe horizontal qui sera importante.

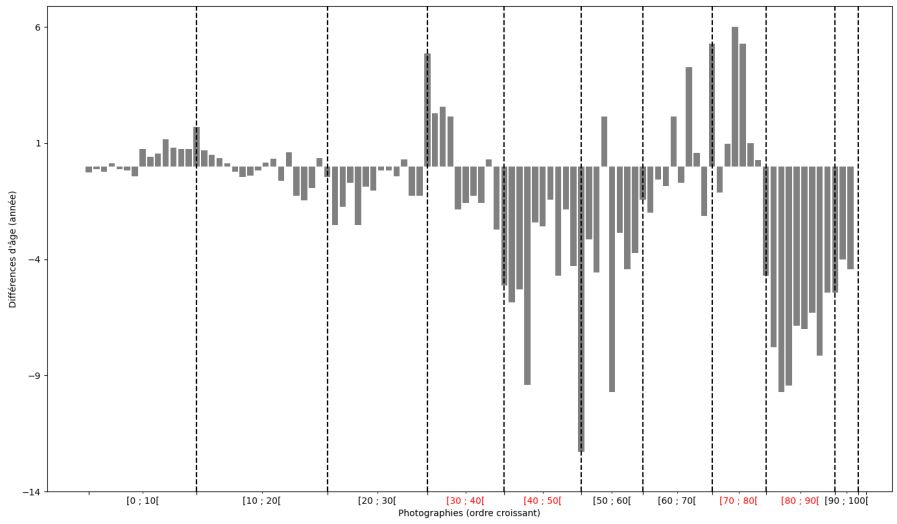


Figure 4.14 – Différences des moyennes attribuées par les annotateurs des scénarios **S1** et S4.

Pour rappel, les tranches qui ont été inversées lors de la présentation du corpus aux annotateurs du S4 sont les tranches [30;40[et [40;50[, puis [70;80[et [80;90[. Les annotateurs ont donc eu des photographies de personnes âgées entre 20 et 30 ans, puis entre 40 et 50 ans, ensuite entre 30 et 40 ans, et enfin 50 et 60 ans. Sur le graphique de la figure 4.14, nous pouvons voir une tendance nette à rajeunir au début de la tranche [40;50[, sans

7. Dans l'optique de traiter finement ces annotations, il serait nécessaire d'intégrer une méthode de pondération afin de compenser cet autre biais.

que cette tendance disparaisse. Les premières photographies de la tranche [30;40[ont, quant à elles, été vieillies (par rapport aux annotateurs du scénarios **S1**). Les annotateurs semblent toutefois encore impactés par l'inversion pour la tranche [50;60[, car les estimations rajeunissent (presque) toutes les personnes dans cette tranche. Ces observations sont sensiblement les mêmes avec les tranches [70;80[et [80;90[.

Il semble donc bien y avoir un impact fort de l'échange des tranches sur les annotations produites par les annotateurs ayant eu le scénario S4. Cette conclusion fait écho à la dépendance sérielle visuelle assimilative évoquée par Fischer et Whitney (2014) et Pegors et al. (2015), qui ferait paraître l'image actuelle plus similaire à la précédente. Si les annotateurs sont confrontés à une série de portraits qui semble suivre une régularité, ils auraient plus tendance à annoter en respectant cette régularité.

Pour cette méthode d'analyse, il est nécessaire de savoir *a priori* quelles cohortes ont subi le biais. C'est plus une méthode de validation que de détection.

4.3.2.3 Mesurer la dispersion des annotations

Si les annotateurs ont subi un biais, nous pouvons émettre l'hypothèse qu'ils seront moins consensuels dans leurs annotations produites et proposer plus de variation dans leurs réponses que des annotateurs non biaisés. L'emploi de l'écart-type est alors tout à fait pertinent pour détecter cela : nous pouvons observer l'écart-type au sein de chaque scénario pour mesurer la dispersion des annotations. Attention toutefois : la façon dont le biais influence les annotations, c'est-à-dire si les annotateurs ont eu tendance à rajeunir ou vieillir, ne peut pas être détectée à partir de cette méthode.

Néanmoins, si le biais impacte de façon uniforme tous les annotateurs d'un groupe donné, il ne sera pas possible de le détecter en comparant l'écart-type moyen selon les scénarios — car le biais est intrinsèque. Dans ce cas, la comparaison en prenant l'écart-type de deux groupes regroupés nous donnera plus d'information : si un des sous-groupes suit une tendance divergente de l'autre groupe, cela va augmenter la disparité avec les annotations de l'autre sous-groupe, et donc l'écart-type sera plus grand. C'est pourquoi, sur les tableaux 4.7 et 4.8, nous présentons les écarts-types moyens pour les scénarios individuels, mais aussi les écarts-types pour des scénarios regroupés. Les scénarios regroupés se basent sur l'analyse menée en 4.3.1.2 : il s'agit des scénarios **S1**–S4 et **S2**–S5.

Scénarios	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S1 et S4**	**S2 et S5**
Écart-type	4,28	4,04	5,05	4,42	3,7	4,05	5,62	4,65	4,35

TABLE 4.7 – Écart-type moyen selon le scénario.

Les écarts-types moyens présentés dans le tableau 4.7 n'indiquent pas de grande disparité au sein d'un scénario, sauf pour les scénarios à très fort caractère aléatoire (**S3** et **S7**). Les annotateurs du **S5** seraient aussi assez consensuels au sein de leur groupe, ayant seulement 3,7 d'écart-type ; la majorité se situe entre 4 et 4,5. Pour les écarts-types des scénarios regroupés, leur regroupement dégrade un peu l'écart-type mais rien de très significatif. Le biais, s'il y en a un, se répercuterait donc de manière homogène sur l'ensemble des annotateurs d'un groupe. Il ne serait toutefois pas assez impactant pour être détecté au niveau global en comparant deux sous-groupes d'annotateurs, l'un étant influencé par ce biais et l'autre non.

		Scénario individuel				**Scénarios regroupés**	
		S1	**S4**	**S2**	**S5**	**S1 et S4**	**S2 et S5**
Tranche	[30;40[	5,11	5,18	3,54	4,25	5,35	4,19
	[40;50[	5,96	5,72	5,56	5,11	6,37	6
	[70;80[	6,39	6,65	5,71	8,12	6,82	7,32
	[80;90[	6,37	7,5	6,59	5,14	7,96	7,34
	Autres	3,37	3,42	3,33	2,71	3,58	3,37

TABLE 4.8 – Écart-type moyen sur les tranches inversées

À présent, de manière plus locale, nous nous intéressons aux écarts-types sur les tranches inversées dans les scénarios **S4** et **S5**, toujours en comparant avec les scénarios **S1** et **S2**. Si nous regardons les écarts-types des scénarios individuels, nous ne remarquons pas de grandes disparités, si ce n'est pour la tranche [80;90[de **S4** (écart-type à 7,5) et celle de [70;80[de **S5** (écart-type à 8,12). Toutefois, ces observations sont trop ponctuelles pour révéler un biais.

Concernant les écarts-types des scénarios regroupés, ceux des premières tranches inversées ([40;50[et [80;90[) sont relativement supérieurs à la moyenne des scénarios pris individuellement. Les annotations des sous-groupes seraient donc assez hétérogènes pour qu'un impact soit visible sur l'écart-type. Toutefois, cette tendance semble se lisser dans la seconde tranche inversée.

De même que l'expérience précédente, il convient de connaître quel groupe d'annota-

teurs peut avoir été impacté par le biais et d'avoir identifié les tranches susceptibles de le provoquer. Néanmoins, de façon plus générale, cette méthode peut être utile pour repérer là où les annotations divergent et prêter une attention plus particulière à ces annotations.

4.4 Résultats complémentaires

Un deuxième objectif de la campagne des « Portraits » consiste en l'exploration de méthodes pour analyser et évaluer des annotations numériques. En ce sens, nous avons mené dans un premier temps des analyses statistiques sur les annotations recueillies. Ensuite, nous nous sommes penchés sur la question de l'évaluation de telles annotations, et comment prévenir les biais possibles liés à ce type d'annotation. Enfin, nous avons voulu comparer avec d'autres types de données scalaires, notamment issues du domaine sportif.

4.4.0.1 Correction de l'écart entre l'annotation et la référence

Nous définissons la formule de l'écart absolu moyen comme étant la moyenne des moyennes des écarts absolus entre la référence et les annotations de chaque image. La moyenne absolue des écarts par rapport aux références est de $5,49$ ans, toutes vagues confondues. Nous nous sommes dans un premier temps demandée s'il était possible de réduire cette moyenne. Ainsi, à partir du graphique , nous avons proposé une première formule de correction des écarts des annotations par rapport à la référence :

$$\frac{|x-a|}{\frac{a}{10}+1} \tag{4.3}$$

Nous obtenons alors un écart absolu moyen de $1,14$ ans : la formule semble bel et bien améliorer les écarts. Nous avons recalculé une régression linéaire (graphique 4.15), permettant toujours de mieux visualiser la tendance générale des écarts, et la droite ainsi tracée se rapproche d'un écart nul.

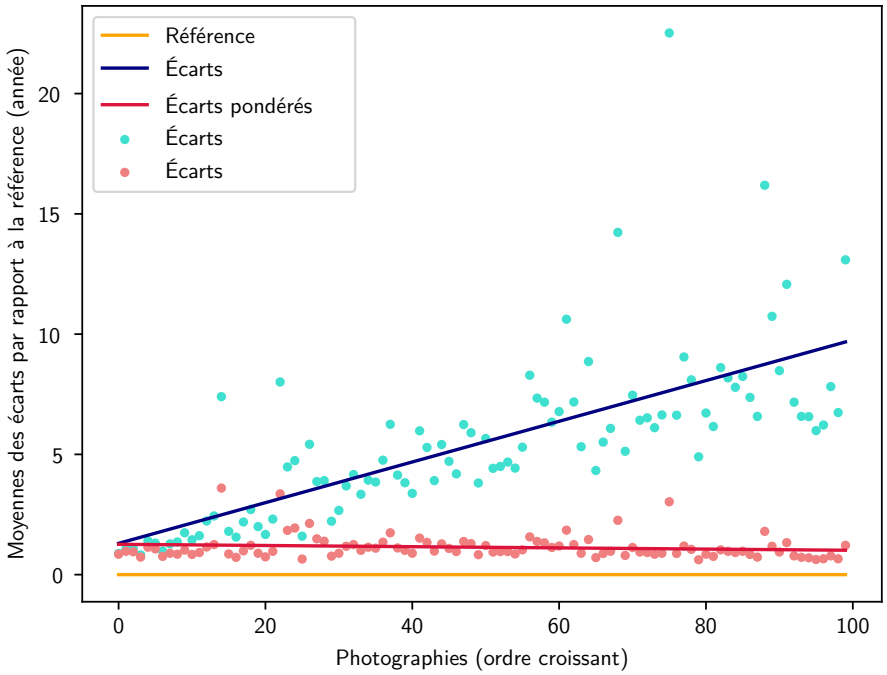


FIGURE 4.15 – Régression linéaire synthétisant l'évolution des écarts par rapport à la référence

4.4.0.2 Correction directe des annotations

Dans cette partie, nous utilisons la formule de l'erreur globale de groupe qui est la moyenne des écarts entre les annotations d'une image et la référence de cette même image. Nous définissons cette formule ainsi :

$$\frac{1}{n}\sum_{i}^{n}|x_i - a_i| \tag{4.4}$$

Une de nos premières suppositions est de s'inspirer de la formule 4.3, en ajoutant le résultat à chacune des estimations :

$$\frac{a-x}{\frac{a}{10}+1} \tag{4.5}$$

L'utilisation de cette formule n'a pas été concluante, à deux égards :

— La moyenne des annotations pour chaque image s'en trouve détériorée et augmente les écarts-types ;

— Le but est d'essayer de s'affranchir de l'accès à la référence, or cette formule contient encore une comparaison avec la référence.

Nous avons alors cherché à établir une formule de correction d'annotation, se basant sur l'existence d'un âge pivot p, en-dessous du quel nous aurions tendance à vieillir, et rajeunir au-dessus, modulé aussi par un facteur arbitraire f :

$$x' = \begin{cases} x & \text{si } x < 10 \\ x + (p - x) \times f & \text{sinon} \end{cases} \tag{4.6}$$

Pour chaque combinaison de $p \in [20; 40]$ et $f \in [0{,}1; 0{,}30]$, nous avons donc corrigé les annotations et calculé l'erreur globale de groupe 4.4. Le graphique 4.16 affiche les erreurs globales : les résultats sont peu concluants. En effet, sans la correction des annotations, l'erreur globale est à 3,43 ans, alors qu'avec la meilleure combinaison ($p \in [38; 40]$ et $f = 0{,}1$), nous obtenons 4,62 ans d'erreur globale.

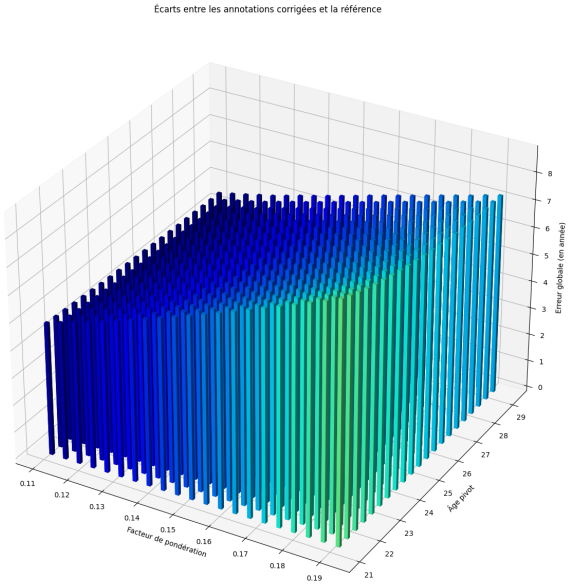


FIGURE 4.16 – Erreurs globales pour chaque combinaison de $p \in [20; 40]$ et $f \in [0{,}1; 0{,}30]$

À la suite de cette seconde expérience peu concluante, nous nous sommes interrogée sur une autre manière de corriger de telles annotations. Au lieu d'appliquer la même correction pour toutes les tranches d'âge, nous avons décidé de trouver une correction plus adaptable selon les tranches d'âge. Une première piste a été de définir deux âges pivots, correspondant à un âge pivot inférieur (i) et à un âge pivot supérieur (s), et

d'ajouter ou retirer un an à l'annotation si cette dernière est comprise entre ces âges pivots :

$$x' = \begin{cases} x & \text{si } x < i \\ x+1 & \text{si } x > s \\ x-1 & \text{sinon} \end{cases} \tag{4.7}$$

Le but était ensuite de faire varier la valeur de ces âges pivots et de comparer les erreurs globales obtenues ainsi. Les premiers résultats semblent être plus convaincants : pour chaque combinaison de $i \in [10; 25]$ et $s \in [40; 50]$, l'erreur globale est comprise entre [3,23; 3,32].

Toutefois, nous nous demandons si nous pouvons améliorer davantage ces résultats, en affinant la formule 4.7 : varier la valeur à ajouter ou retirer à l'annotation selon des tranches d'âge plus fines. Par exemple, cela peut être ajouter 3 ans à une annotation estimant la personne entre 70 et 80 ans et, *a contrario*, retirer 2 ans lorsque l'annotateur a estimé que l'individu paraissait avoir entre 20 et 30 ans.

4.5 Conclusion

Dans ce chapitre, nous avons présenté la campagne « Portraits ». Les objectifs initiaux de cette campagne étaient multiples :

— étudier le lien entre la performance et l'accord ;
— mieux comprendre l'évaluation d'annotations numériques ;
— détecter et quantifier la présence d'un biais, plus particulièrement celui de l'ordre de présentation des items.

L'étude du lien entre la performance et l'accord a permis d'introduire une nouvelle notion : celle de la consensualité. Cette mesure sert notamment à appréhender à quel point un annotateur se distingue d'un groupe d'annotateurs. Grâce à cette mesure, nous avons pu mettre en exergue un premier rapport entre la performance et la consensualité des annotateurs. En effet, nous avons pu voir grâce à nos analyses que les annotateurs les moins consensuels sont aussi les moins performants.

Pour poursuivre l'approfondissement, nous avons aussi distingué deux types de consensualité : la consensualité initiale, calculée à partir du groupe d'annotateurs de base, et la

consensualité dynamique, où les annotateurs sont retirés au fur et à mesure du calcul. La consensualité dynamique semble permettre un meilleur filtrage des annotateurs, gardant ceux étant à la fois consensuels et performants. Enfin, nous avons aussi montré que la consensualité restait relativement homogène sur l'ensemble du corpus.

Les premières expérimentations sur l'évaluation des annotations numériques sont, pour l'instant, à l'état expérimental. Grâce à la modélisation du comportement d'un annotateur type, effectuée avec la régression linéaire, nous avons pu voir que les annotateurs avaient eu tendance à légèrement surestimer l'âge des personnes jeunes et à sous-estimer celui des personnes âgés. Nous avons cherché à corriger directement les annotations, toutefois les résultats actuels ne sont pas probants. Il reste à effectuer une comparaison avec d'autres types de données scalaires, afin de mieux saisir les particularités de l'annotation et de l'évaluation de chaque type d'erreur qui puisse exister.

La détection et la quantification de la présence d'un biais a permis de poser les premiers jalons pour l'étude d'un biais. Plus particulièrement, nous avons étudié si l'ordre dans lequel sont présentés les items aux annotateurs influence leurs annotations et, si oui, dans quelle mesure. Avec un accès à la référence, il a été possible d'observer tant au niveau global qu'au niveau local la présence et l'impact du biais. Il reste néanmoins encore un travail à fournir concernant la détection des biais sans un accès à la référence, car notre expérience n'a pas permis de mettre en valeur de méthode probante.

Les méthodes d'analyses présentées dans ce chapitre ont été conçues pour des annotations numériques, scalaires. Ce type d'annotation reste toutefois assez rare en T.A.L., à la différence des annotations catégorielles ou nominales qui sont extrêmement fréquentes. L'adaptation des méthodes à ces annotations se révèle donc une nécessité. À la manière présentée dans Artstein et Poesio (2008), nous pourrions nous inspirer de la formule de la variance et en abstraire une notion de distance entre des catégories nominales, pour pouvoir appliquer certaines méthodes décrites dans le présent chapitre.

Aux yeux de certains, les expérimentations menées au cours de ce chapitre peuvent paraître vaines pour le T.A.L., car menées à partir de données numériques. Or, notre objectif principal est de fournir des méthodes pour l'annotation en général, et nous pensons que les méthodes présentées dans ce chapitre sont utilisables sur des données textuelles. Quant à l'ordre de présentation de items, la question se pose aussi en T.A.L. D'une part, lorsque nous annotons un texte, il s'agit d'un continuum, d'un flux textuel, et un phénomène linguistique n'est pas forcément présent équitablement tout au long d'un flux.

D'autre part, même dans le cas où les items sont présentés dans un ordre aléatoire, cet ordre n'est pas exempt d'avoir, somme toute, une certaine récurrence.

Chapitre 5

Campagne d'annotation « Erreurs »

Sommaire

5.1 Typologie des erreurs et corpus disponibles 143
5.1.1 Typologie des erreurs 143
5.1.2 Corpus d'erreurs de français disponibles 144
5.2 Présentation de la campagne 145
5.2.1 Objet annoté et liens entre les items 146
5.2.2 Modalité d'interaction et de saisie : le retour arrière 148
5.2.3 Modalité de présentation : ordre de présentation 149
5.2.4 Déroulement de la campagne 150
5.2.5 Première approche des résultats 153
5.3 Traiter deux cohortes hétérogènes ? 154
5.3.1 Étude des scores 154
5.3.2 Comment traiter une telle disparité ? 157
5.3.3 Réflexions et discussions : motivation et volition des annotateurs 158
5.4 Utiliser la consensualité pour une annotation catégorielle binaire 159
5.4.1 Adaptation des formules 159
5.4.2 Étude globale de la consensualité et de l'imperfection 160
5.4.3 Consensualité des phrases 164
5.5 Retour arrière possible et paires 167
5.5.1 La possibilité du retour arrière influence-t-elle les annotations ? 167
5.5.2 Paires d'énoncés 168
5.6 Résultats complémentaires 172
5.6.1 Niveau d'expertise attribué par les annotateurs 172
5.6.2 Taux de réponses correctes par énoncé 174
5.6.3 Outil non adapté pour l'analyse des biais ? 176

5.7 Conclusion 177

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

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

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.

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 qu'ils contiennent ou non une erreur de français.

Cette tâche satisfait les deux pré-requis supplémentaires que nous nous étions fixés pour l'étude d'un biais. D'une part, cette tâche ne requiert pas d'entraînement et tout le monde est confronté régulièrement à l'acceptabilité d'un énoncé lorsqu'il s'exprime. D'autre part, le jugement de l'acceptabilité d'un énoncé peut varier d'une personne à une autre. Cette variation est principalement due à la connaissance des règles de français et à notre positionnement par rapport à la *norme* (ce qui est prescrit, par l'Académie Française par exemple) et à l'*usage* (ce que nous observons des pratiques langagières). Ainsi, un locuteur peut juger qu'un énoncé est correct s'il se réfère à un usage répandu, voire même rejeter les normes. De plus, les annotateurs peuvent se tromper sur l'application des règles ou ne connaissent pas toutes les règles, *a fortiori* rares ou complexes.

Une partie du contenu de ce chapitre a fait l'objet d'une publication à la conférence Rencontres Etudiants Chercheurs en Informatique pour le TAL (RÉCITAL) (Baledent, 2022).

5.1 Typologie des erreurs et corpus disponibles

5.1.1 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), elles-mêmes inspirées de Anxionnaz (2015) ; RO et Ledegen (2012) ; Roubaud (2014), ont montré que certains types d'erreurs reviennent plus souvent et peuvent être classifiés ainsi :

Erreur orthographique : L'énoncé présente une erreur liée à l'orthographe. 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*), ou encore un problème d'homophonie. Exemples :
- *proffesseur* au lieu de *professeur* ;
- *dicernement* au lieu de *discernement* ;
- *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. Exemples :
- *La jeune fille est tombé* au lieu de *La jeune fille est tombée* ;
- *J'ai assister au cours* au lieu de *J'ai assisté au cours* ;
- *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 préposition employée, etc.
- *Tu te rappelles du film d'hier ?* au lieu de *Tu te rappelles le film d'hier* ;
- *Je t'ai pas entendu arriver* au lieu de *Je ne t'ai pas entendu arriver* ;
- *Le chat que je parle est roux* au lieu de *Le chat dont je parle est roux*.

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 ne sont forcément pas étanches : une erreur de grammaire peut découler d'une erreur d'orthographe ou de lexique. Par ailleurs, il est parfois difficile de savoir précisément pourquoi un locuteur se trompe : commet-il une erreur de grammaire

parce qu'il ne connaît pas la règle, parce qu'il confond plusieurs règles, parce qu'il ne connaît pas les exceptions… ?

5.1.2 Corpus d'erreurs de français disponibles

Nous souhaitons chercher un corpus regroupant des erreurs de français. Cette recherche a pour but principal de trouver un *support* pour construire notre expérience, et non pour analyser la campagne d'annotation. Ce support doit idéalement présenter des phrases courtes contenant une seule erreur, proposant une certaine variété dans les erreurs et avec une difficulté variable [1].

Des corpus regroupant des erreurs de français existent déjà. Nous pouvons notamment citer le corpus WiCoPaCo [2], 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 (voir la figure 5.1). Le corpus TRACE [3] (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 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 l'expérience que nous souhaitons mener.Néanmoins, quelques-uns de ces corpus sont disponibles dans un format textuel,comme c'est le cas des corpus É-Calm [4] (Ho-Dac et al., 2020), EMA [5] (Boré & Elalouf, 2017), Dire Autrement [6] (Hamel & Milicevic, 2007) ou encore CEFLE [7] (Ågren, 2008).

Toutefois, ces corpus ne correspondent pas à ce que nous souhaitons pour notre expé-

1. 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.
2. https://wicopaco.limsi.fr/.
3. https://anrtrace.limsi.fr/.
4. http://e-calm.huma-num.fr/.
5. https://www.ortolang.fr/market/corpora/ema-ecrits-scolaires-1.
6. http://web5.uottawa.ca/direautrement/index.html.
7. https://projekt.ht.lu.se/cefle.

Correction	
Normalizations	***Son 2ème disque → Son deuxième disque***
Non-Word Error Corrections	c'est-à-dire la ***dernrière → dernière*** année avant l'ère chrétienne
Diacritics Error Corrections	la jeune Natascha Kampusch, ***agée → âgée*** de 18 ans
Real-Word Error Corrections	dans le but de ***sensibilisé → sensibiliser*** sur les changements
Reformulation	
Close Meaning	Le tritium ***existe dans la nature . Il est produit → se forme naturellement*** dans l' atmosphère
	"Gimme Gimme Gimme" et "I Have A Dream" ***contribueront au gigantesque succès de → viendront alimenter la gloire d'*** Abba
Different Meaning	alors ***que l' ordinateur → qu'un processeur de la famille x86*** reconnaîtra ce que l' instruction machine
	Le principal du collège M. Desdouets → Un de ses professeurs dit de lui
	Des opérations de base sont disponibles dans ***tous les → la plupart des*** jeux d' instructions
Spam	
Obvious Agrammatical Spamming	Süleyman Ier s' ***empare de l' Arabie et fait entrer dans l' → emp kikoo c moi ca va loll*** ' empire ottoman Médine et La Mecque
Subtle Grammatical Spamming	pour promouvoir la justice , la solidarité et ***la paix → l'apéro*** dans le monde

FIGURE 5.1 – Exemples tirés du corpus WICOPACO

rience. 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.

5.2 Présentation de la campagne

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, en accord avec les trois critères évoqués en introduction, la construction du corpus que nous avons décidé

de créer, ainsi que la manière dont nous avons pensé l'expérience. Pour rappel, ces trois aspects sont :

— la constitution du corpus et les relations entre les items à annoter ;
— la modalité d'interaction ou de saisie : pour notre expérience, nous nous interrogeons sur le fait que la possibilité du retour en arrière ou non influence la qualité des annotations ;
— la modalité de présentation : en l'occurrence, il s'agit de l'ordre des questions.

5.2.1 Objet annoté et liens entre les items

La tâche d'annotation choisie, le repérage d'erreurs de français, peut avoir trois significations différentes :

a) localiser l'erreur ;
b) repérer si une phrase contient une erreur ou non ;
c) catégoriser le type d'erreur.

Ces trois exemples de tâches sont assimilables à la différence qui existe entre l'*unitizing*, tel que défini par Krippendorff (1995) et décrite dans le 2, 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/-grammaticale/etc.). Des combinaisons, par exemple localiser l'erreur et la catégoriser, sont possibles.

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 avec une référence ou d'autres annotations. 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 des 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 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 souligné par Di Eugenio et Glass (2004).

Nous avons produit en collaboration avec Yann Mathet et Antoine Widlöcher le corpus. Pour ce faire, nous avons créé des énoncés, d'une longueur d'une phrase ; certains exemples sont visibles dans le tableau 5.1. 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.

Code	**Énoncé**
SF007	Je vous renvoie à la discussion de la semaine dernière pour éclairer ce point.
SF011D	Vous devez indiquer les phrases comportant des erreurs.
SF020J	C'est un film que je me rappelle très précisément.
AF006	Cette proposition n'est ni raisonable, ni rationnelle.
AF011E	Le comportement de cet individu laisse pensé qu'il ignore le règlement intérieur.
AF015	Ce livre est un espèce de mélange entre la fantasy et la science-fiction.

Table 5.1 – Exemples d'énoncés extraits du corpus.

Enfin, nous désirons aussi analyser le comportement des annotateurs lorsqu'ils rencontrent des occurrences presque identiques à annoter : ont-ils tendance à tenir compte de l'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 des « 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.

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

Comme évoqué dans le chapitre 1, le choix de l'outil d'annotation n'est pas à sous-estimer. 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 à la possibilité de voir et modifier ses annotations précédentes. Une grande majorité des outils d'annotation propose un retour arrière pour corriger ses annotations, ou du moins pour voir 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 Fort et al. (2014) ; Poesio et al. (2013).

Par ailleurs, il faut aussi distinguer la *possibilité* et l'*incitation* au retour arrière. En effet, la possibilité technique du retour arrière peut être disponible (par exemple la présentation de tous les items sur une seule et même page), et l'annotateur devine implicitement qu'il peut revenir sur ses anciennes annotations. Cela est différent que si nous invitons explicitement (en le spécifiant dans les consignes données ou à l'aide d'un bouton visible) l'annotateur à utiliser la fonctionnalité du retour arrière, que cela soit pour voir ou corriger ses anciennes annotations si l'annotateur le souhaite.

Il convient alors de s'interroger si la possibilité 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.

5.2.3 Modalité de présentation : ordre de présentation

La modalité de présentation, c'est-à-dire ici 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 (Fort, François, Galibert et al., 2012 ; Mathet & Widlöcher, 2016), il n'y a pas, à notre connaissance, d'étude sur une distribution inégale observée non à l'échelle du corpus mais de parties de celui-ci.

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 désirons notamment pouvoir observer quelque chose comme une 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 concerne le fait que les paires doivent être directement contiguës ou non ; nous souhaitons aussi savoir si la distance entre les membres de la paire a un effet mesure, par exemple si le premier item a été oublié au moment de la consultation du second si la distance est trop importante ;
— l'ordre interne des paires : nous désirons aussi regarder si le fait de présenter les énoncés des paires *Sans erreur*/*Avec erreur* ou *Avec erreur*/*Sans erreur* influençait

les annotations.

Finalement, nous nous sommes arrêtée sur quatre scénarios différents[8], dont la description est donnée ci-dessous :

Scénario 1 : Nous présentons 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 contiguës. L'ordre de présentation *Sans erreur* et *Avec erreur* est préservé. 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[9] et l'ordre de présentation de ces dernières (*Sans erreur*/*Avec erreur* ou *Avec erreur*/*Sans erreur*) varie selon les cas.

Nous rappelons que chaque scénario est proposé dans deux versions, une sans retour arrière possible, une autre avec retour arrière possible. Du point de vue de l'implémentation logicielle, les énoncés étaient, dans ce cas-là, affichés sur une seule et même page. Les captures d'écran de la figure 5.4 illustrent cette différence de présentation.

5.2.4 Déroulement de la campagne

5.2.4.1 Première vague

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

8. Nous avons réfléchi à d'autres scénarios, notamment pour observer plus finement le lien entre les items à annoter. 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.

9. Il est à noter que l'éloignement de 33 phrases a été choisi arbitrairement.

FIGURE 5.2 – **Avec** retour arrière possible

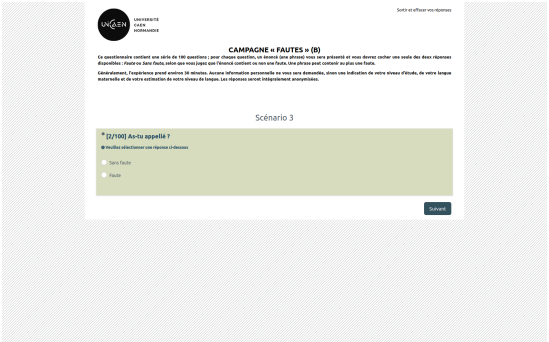


FIGURE 5.3 – **Sans** retour arrière

FIGURE 5.4 – Retour arrière : différence de présentation.

différents scénarios. À cette fin, 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.

Les origines des annotateurs sont donc multiples. Pour cette raison, 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 :

1. le niveau d'études ;
2. l'auto-évaluation du niveau de français, sur une échelle de 1 (Très mauvais) à 5 (Excellent) ;
3. si le français est la langue maternelle.

Par ailleurs, afin d'inciter les personnes à participer plus volontiers à la campagne, nous avons aussi voulu rendre l'expérience aussi pédagogique que possible. 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 en fin de session d'un document explicatif des réponses et des règles abordées durant l'expérience. Ce document peut être retrouvé en annexe B.

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

	S1	S2	S3	S4	Total
ARA	6	10	8	4	28
SRA	2	11	8	5	26

TABLE 5.2 – Nombre d'annotateurs par scénario à l'issue du premier appel à participation. **ARA** signifie *Avec retour arrière*, **SRA** signifie *Sans retour arrière*.

5.2.4.2 Deuxième vague

Quelque temps après la première vague, nous avons voulu relancer la campagne. Cette relance avait notamment pour origine la modification de quelques énoncés. En effet, certains énoncés comprenaient parfois deux erreurs, ce qui entraînait une incertitude quant à l'identification. Pour d'autres phrases, il manquait aussi des points, ce qui était perçu comme une erreur pour certains annotateurs — alors que la catégorie de la phrase était *Sans Erreur*. Les énoncés ayant changé sont présentés dans le tableau 5.3.

Code	Origine	Correction
SF030	Il y a quelque 2000 personnes qui se sont rendues à la manifestation	Il y a quelque 2000 personnes qui se sont rendues à la manifestation**.**
SF035	Elle gardait une place dans ses pensées quoiqu'il advînt.	Elle gardait une place dans ses pensées **quoi qu'il** advînt.
SF043	Leur moment était celui de ce battement de cœur raté	Leur moment était celui de ce battement de cœur raté**.**
SF044	L'adolescente porte une robe à carreaux rouge et blanc.	L'adolescente porte une robe **bleu marine**.
AF032	Aujourd'hui dans cette vidéo, je vous partage mon expérience pour réussir vos cheese-cakes	Aujourd'hui dans cette vidéo, je vous partage mon expérience pour réussir vos cheese-cakes**.**
AF016	Parmis ces figures de style, lesquels vous semblent être des métonymies ?	Parmis ces figures de style, **lesquelles** vous semblent être des métonymies ?

TABLE 5.3 – Énoncés corrigés entre les deux vagues. Les correction apparaissent en **gras**. Pour les énoncés SF030, SF043 et AF032, un point final a été ajouté.

Nous y avons principalement fait participer des collégiens, d'un niveau quatrième–troisième, par l'intermédiaire de leur professeur de français.

Si nous avions réussi à obtenir une répartition équilibrée entre les deux modalités de

	S1	S2	S3	S4	Total
ARA	11	19	18	23	71
SRA	3	6	9	6	24

TABLE 5.4 – Nombre d'annotateurs par scénario pour la vague 2.

la campagne (**ARA** et **SRA**) pendant la vague 1, la répartition est inégale pour cette vague 2 — nous reviendrons sur ce problème. Les analyses qui suivent vont devoir tenir compte de cette perturbation.

5.2.5 Première approche des résultats

Pour observer les principales tendances, nous pouvons commencer par analyser la dispersion des scores, c'est-à-dire le pourcentage de réponses correctes au questionnaire, présentée sur le graphique de la figure 5.5, en séparant les deux vagues.

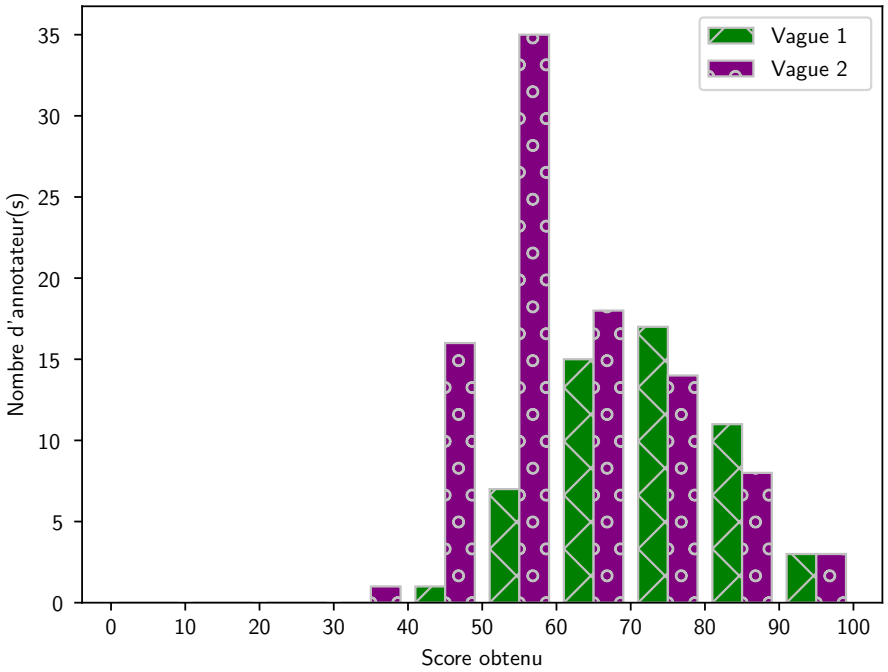


FIGURE 5.5 – Dispersion des scores.

Les scores individuels de la vague 1 sont assez regroupés : les annotateurs ont généralement un score aux alentours de 70, même si plusieurs ont plutôt des scores hétérogènes

entre 55 et 85. En comparaison, la vague 2 est plus hétérogène, avec une prédominance des scores entre 50 et 60.

Concernant la distribution effective des catégories *Sans erreur* et *Avec erreur*, affectées par les annotateurs aux items, indépendamment de leur catégorie réelle, elle est présentée dans le tableau 5.5.

Vague	***Sans erreur***	***Avec erreur***
Vague 1	54% (2730)	46% (2270)
Vague 2	58% (5540)	42% (3960)

TABLE 5.5 – Distribution des catégories *Sans erreur* et *Avec erreur*, selon les deux vagues.

Les énoncés étant répartis équitablement entre les deux catégories, nous aurions pu nous attendre à retrouver une distribution parfaitement partagée entre *Sans erreur* et *Avec erreur* ; ce n'est pas le cas. En effet, nous remarquons une préférence des annotateurs à ne pas repérer d'erreur, car il y a plus d'annotations *Sans erreur*, et ce, pour les deux vagues. Les annotateurs, lorsqu'ils ne détecteraient pas d'erreur ou n'auraient pas une idée claire d'erreur à détecter, répondraient, dans le doute, que la phrase est correcte. Cette préférence pourrait n'être aussi que l'incapacité à repérer certains types d'erreurs peut-être plus difficiles.

5.3 Traiter deux cohortes hétérogènes ?

Comme nous l'avons expliqué, les deux vagues n'ont pas été réalisées dans les mêmes conditions. De plus, nous pouvons voir, grâce au graphique de la figure 5.5, une hétérogénéité dans les scores obtenus par les annotateurs selon les deux vagues.

5.3.1 Étude des scores

Dans un premier temps, nous avons souhaité regarder les scores obtenus selon le niveau d'études, grâce à la question préliminaire prévue pour distinguer les cohortes. Nous pensons, en effet, que cette première approche peut nous aiguiller sur les possibles raisons de la disparité des scores. Il convient de mettre en perspective le graphique de la figure 5.6 avec l'effectif pour chaque vague et chaque niveau d'étude (affiché dans le tableau 5.6).

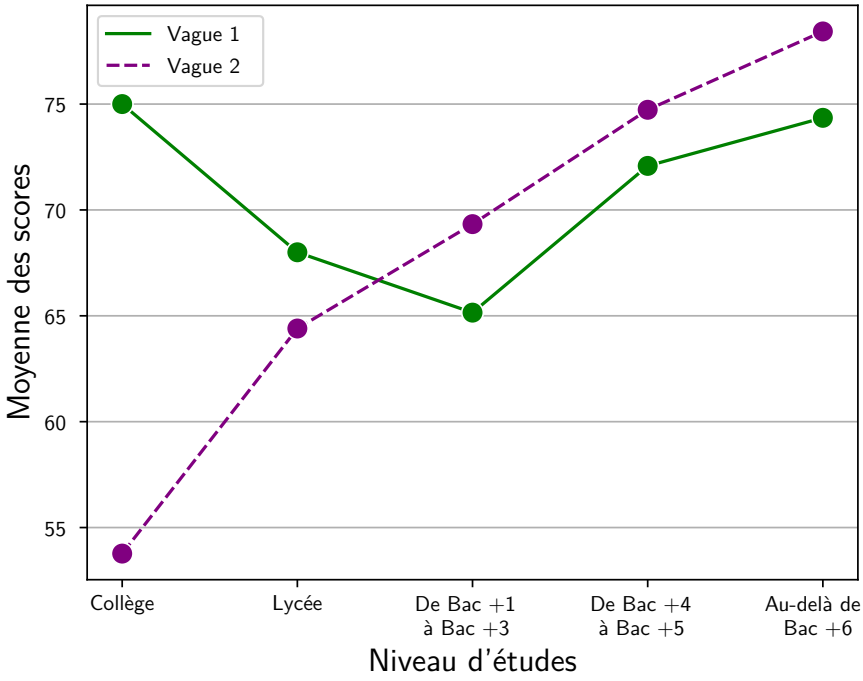


FIGURE 5.6 – Moyenne des scores selon les niveaux d'études des participants.

Vague	Collège	Lycée	De Bac +1 à Bac +3	De Bac +4 à Bac +5	Au-delà de Bac+6
Vague 1	1	2	13	12	26
Vague 2	57	5	15	11	7

TABLE 5.6 – Nombre de participants par vague et par niveau d'études.

Le nombre d'annotateurs par niveau d'étude est globalement le même selon les vagues. Il y a deux exceptions :

— niveau d'étude « Collège » : il y avait un seul participant lors de la vague 1, tandis qu'il y en a 57 pour la vague 2 :

— niveau d'étude « Au-delà de Bac+6 » : il y avait 26 participants pour la vague, et seulement 7 pour la vague 2.

Les scores moyens présentent peu de disparités selon les deux vagues, sauf lorsqu'il s'agit des collégiens. Ce contraste se retrouve à deux niveaux : le score moyen entre la vague 1 et la vague 2 [10] et entre ce niveau d'étude particulier et les autres.

10. Soulignons toutefois que l'annotateur collégien de la vague 1 est peut-être justement l'anormalité.

Nous avons voulu savoir si cela se répercutait sur l'ensemble de la vague 2, et si oui, dans quelle mesure. Ainsi, nous avons regardé la dispersion des scores selon plusieurs sous-groupes d'annotateurs : l'ensemble de la vague 2, seulement les collégiens de la vague 2, les annotateurs des autres niveaux. Nous avons aussi comparé avec la globalité de la vague 1, ainsi qu'avec la vague 1 et les annotateurs de la vague 2 qui n'étaient pas des collégiens. Les résultats apparaissent sur le graphique de la figure 5.7.

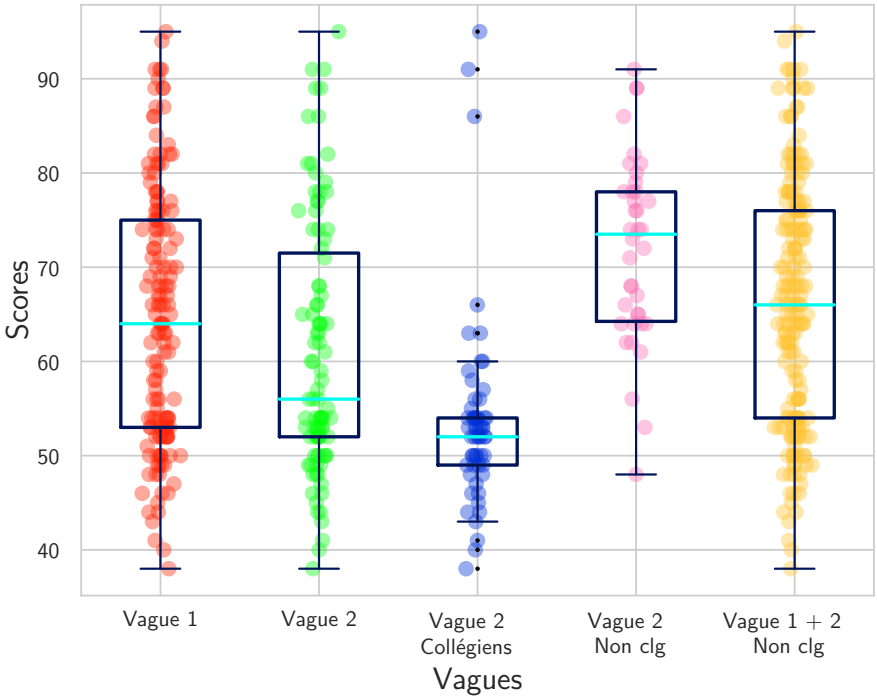


FIGURE 5.7 – Dispersion des scores.

Les valeurs de la vague 1 sont disparates, mais leur distribution reste homogène. Pour la vague 2, la dispersion des scores semble être la même que la vague 1. Néanmoins, nous voyons aussi une accumulation de scores aux alentours de 50. Pour la cohorte des collégiens de la vague 2, nous remarquons aussi trois scores jugés comme « aberrants » (selon le principe des boîtes à moustaches) dans la vague 2 – Collégiens : en effet, trois collégiens de cette vague ont eu des scores de 95, 91 et 86.

La vague 2 est aussi marquée par une autre difficulté. En effet, un fort déséquilibre apparaît entre les deux modalités de la campagne. Pour rappel, il y a presque le triple des participants pour la modalité **ARA** que **SRA**.

5.3.2 Comment traiter une telle disparité ?

Face à ces deux disparités, nous devons réfléchir sur la manière de traiter intelligemment les annotations de ces deux vagues. Plusieurs solutions sont possibles, dont par exemple celles listées ci-dessous :

— **Regrouper les deux vagues** indépendamment de ces problèmes ;
— **Continuer de ne traiter que la vague 1**, sans utiliser les annotations récoltées au cours de la vague 2 ;
— **Traiter uniquement la vague 2** ;
— **Traiter uniquement les collégiens de la vague 2** ;
— **Deux cohortes Non collégiens et Collégiens** : regrouper les données de la vague 1 avec celles de la vague 2 sans les collégiens, et traiter les données des collégiens séparément.

Avant de choisir, nous avons réalisé des analyses, notamment liées à la consensualité (partie 5.4). Nous avons vu que les annotations des Collégiens seuls sont délicates à interpréter, et donc à prendre en compte. Ainsi, nous avons pris la décision de créer deux cohortes : une avec des collégiens, une autre sans les collégiens. Si pour la consensualité, nous avons choisi d'étudier les deux vagues séparément, nous nous sommes concentrée sur la cohorte des Non collégiens pour ce qui est relatif à la possibilité de retour en arrière et aux paires de phrases.

En ne traitant que la cohorte des Non collégiens, nous obtenons la répartition suivante les scénarios et les modalités — à 5 annotateurs près, la répartition est équitable entre ARA et SRA :

	S1	**S2**	**S3**	**S4**	**Total**
ARA	9	14	11	14	48
SRA	5	17	14	7	43

TABLE 5.7 – Nombre d'annotateurs par scénario pour la cohorte des Non Collégiens.

5.3.3 Réflexions et discussions : motivation et volition des annotateurs

Nous pouvons nous interroger sur les potentielles causes de la disparité des scores moyens. Un motif qui semble évident est, qu'effectivement le niveau de français des collégiens est moins bon que ceux d'autres niveaux. Certaines règles abordées durant le questionnaire se révélaient difficiles, surtout pour des collégiens. Certaines remarques en fin de questionnaire vont d'ailleurs dans ce sens :

Remarques
`C'etait dur pour un niveau de collège.`
`TROP DIFFICILE`
`oui car les question ne veulent rien dire`
`difficile`
`Les séries de question m'a sembler assez difficile`

TABLE 5.8 – Exemples de remarques écrites par les collégiens de la vague 2.

Couplée au niveau de difficulté, la volition des collégiens est une autre potentielle cause. Pour rappel, ces derniers participaient sous la suggestion de leur professeur de français, dans un cadre scolaire. Cette participation sur la base du volontariat a peut-être favorisé le fait de faire à la hâte le questionnaire, plus encore vers la fin. *A contrario*, les participants de la vague 1 ont répondu de leur « propre volonté », limitant ainsi les participations aux personnes aimant participer à ce genre d'exercice et avec un bon niveau de français estimé.

D'autres causes sont encore possibles, notamment, la durée du questionnaire. Nous avons en effet estimé la durée du test à trente minutes. Toutefois, cette estimation était sûrement mal calibrée, au moins pour des annotateurs du niveau collège. Une des manières pour tester cette hypothèse est d'observer les scores obtenus en début et en fin de session. Notamment, nous pouvons nous attendre à des dégradations de scores à la fin du questionnaire.

Quant au déséquilibre entre les modalités, la méthode pour répartir les futurs participants n'était sûrement ni la meilleure ni la plus adaptée. En effet, la plateforme LIMESURVEY ne permettait pas de choisir scénario par scénario la manière dont étaient affichées les questions. Ainsi, nous avions dû prévoir deux questionnaires distincts, et rediriger les

participants vers l'un ou l'autre via un script PHP prenant une connexion sur deux. Cette méthode est donc sensible au nombre de clics sur le lien et ne peut pas garantir la répartition entre les deux modalités, même si elle aurait dû en théorie donner une répartition équitable. Faute de temps, nous n'avons pas approfondi les raisons de ce déséquilibre.

5.4 Utiliser la consensualité pour une annotation catégorielle binaire

5.4.1 Adaptation des formules

L'annotation des âges requérant des annotations numériques, il était aisé d'utiliser des fonctions d'agrégation. Par exemple, pour l'imperfection d'un annotateur, nous nous étions inspirée de la formule de l'écart-type, en calculant la racine carré de la moyenne des écarts entre l'âge de référence et l'âge donné. Cette méthode de calcul nous permettait de mieux percevoir les différences de performance entre les annotateurs. Ici, pour une tâche de catégorisation binaire, il convient de réaliser quelques adaptations.

Avant de préciser les changements effectués pour utiliser la consensualité pour la campagne « Erreurs », nous devons poser certaines notations. Le score est le nombre de réponses correctes. N est le nombre total d'énoncés. a est un annotateur. Soit un sous-groupe d'annotateurs a_n, $G = \{a_1, ..., a_n\}$, $|G|$ représente la cardinalité de ce groupe et $\sigma_e(G)$ la variance de ce groupe pour l'énoncé e.

Pour cette campagne « Erreurs », nous avons appliqué la formule 3.1 en l'adaptant à la tâche. Cela consiste simplement à calculer le nombre de réponses incorrectes et à en prendre la racine carrée :

$$\text{imperfection}_{\text{erreur}}(a) = \sqrt{N - score} \tag{5.1}$$

Quand cette valeur est nulle, les annotations de l'annotateur sont toutes valides.

L'intégration ou non de la racine a été source de discussion. Pour les annotations des âges, la racine carrée avait une raison d'être car elle permettait de revenir à des nombres normaux, après l'élévation au carré des écarts entre la référence et l'annotation. Dans le

cadre de cette campagne, en revanche, nous n'utilisons pas de carré. Finalement, nous avons préféré garder la racine carré pour deux raisons principales, à savoir :

— pour un souci de cohérence ;

— pour pénaliser l'imperfection.

Pour l'imperfection de groupe, nous reprenons la même formule que celle utilisée lors de la campagne « Portraits », à savoir :

$$\text{imperfection}_{\text{erreur}}(G) = \frac{1}{|G|} \sum_{j=1}^{|G|} \text{imperfection}_{\text{erreur}}(a_j) \tag{5.2}$$

Pour les calculs liés à la consensualité — le désaccord de groupe et le degré de consensualité —, nous n'avons pas modifié les formules originales. En effet, la variance prend déjà en compte le fait d'avoir des distributions très partagées (par exemple, une répartition d'annotations de 50% *Sans Erreur* et 50% de *Avec Erreur*). Ainsi, les formules utilisées sont les suivantes :

$$\text{desaccord}_{\text{erreur}}(G) = \frac{1}{N} \sum_{e=1}^{N} \sigma_e(G) \tag{5.3}$$

$$\text{consensualite}_{\text{erreur}}(a, G) = \text{desaccord}_{\text{erreur}}(G \setminus a) - \text{desaccord}_{\text{erreur}}(G) \tag{5.4}$$

Les considérations mathématiques réalisées, nous pouvons maintenant passer à l'analyse de la consensualité pour la campagne « Erreurs ».

5.4.2 Étude globale de la consensualité et de l'imperfection

En premier lieu, nous reprenons le principe de la figure 4.3. Cette figure permet de suivre les évolutions des imperfections des annotateurs selon leur rang de consensualité dynamique. Dans un premier temps, nous avons affiché tous les niveaux d'études (en différentes couleurs), ainsi que les deux vagues.

Nous observons un nuage de points en haut à droite, correspondant aux annotateurs les plus moins consensuels et les moins performants. Il est intéressant de relever que ces points, à quatre ou cinq exceptions près, sont des collégiens de la vague 2, ce qui corrobore

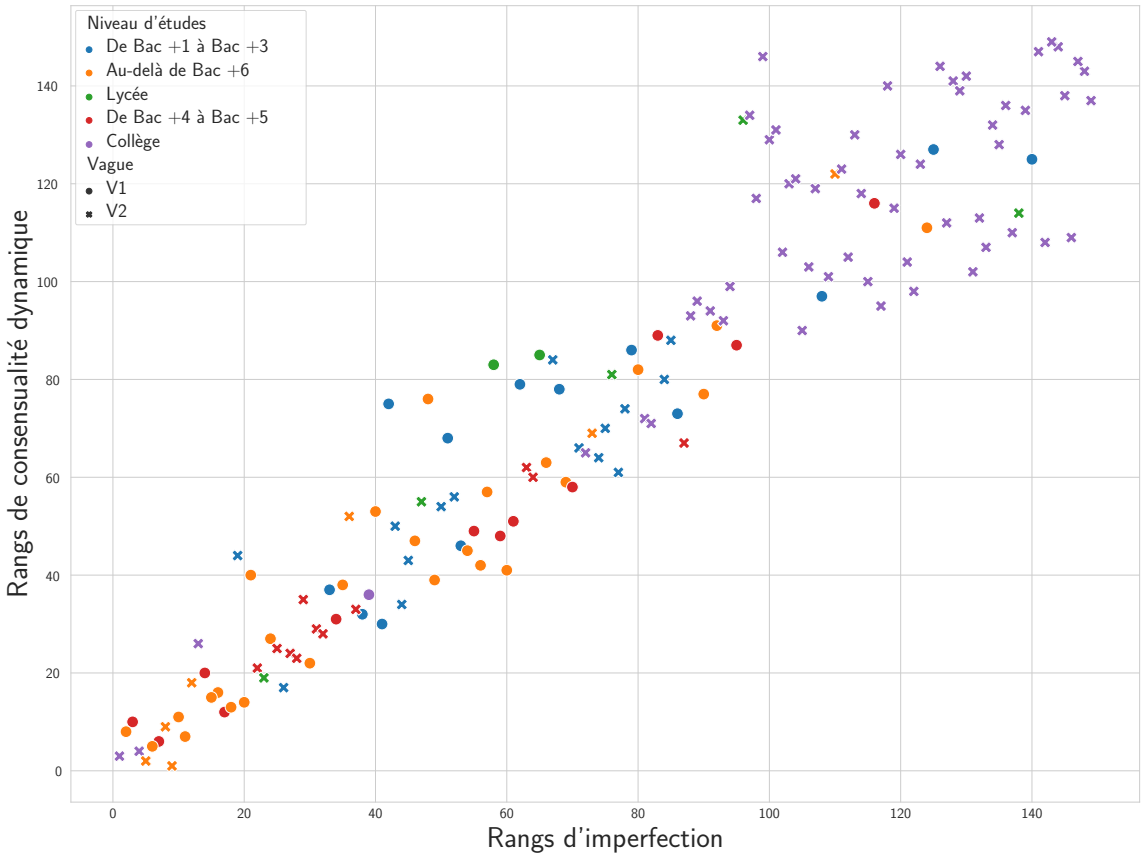
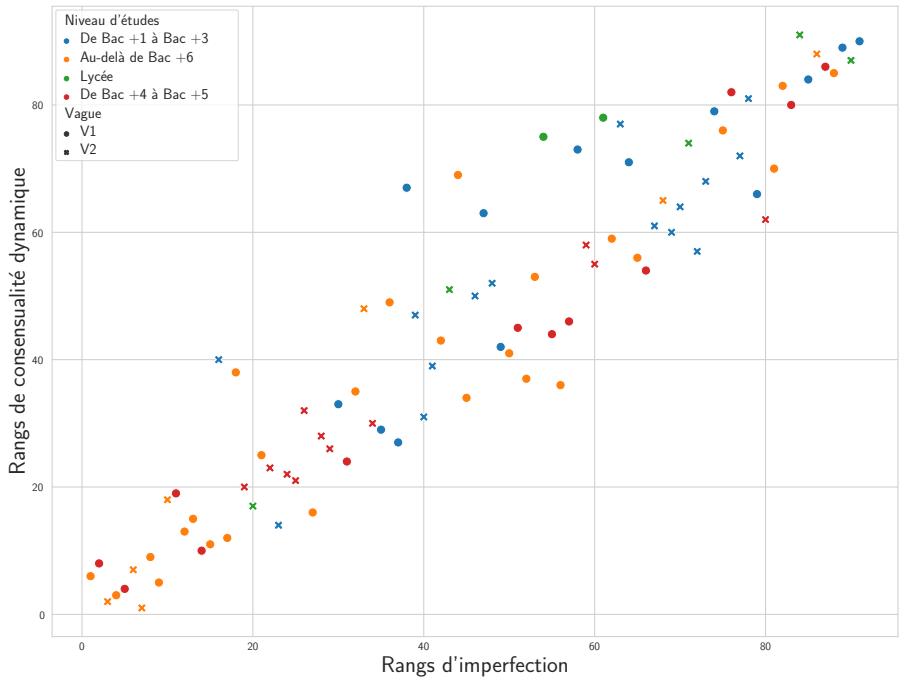


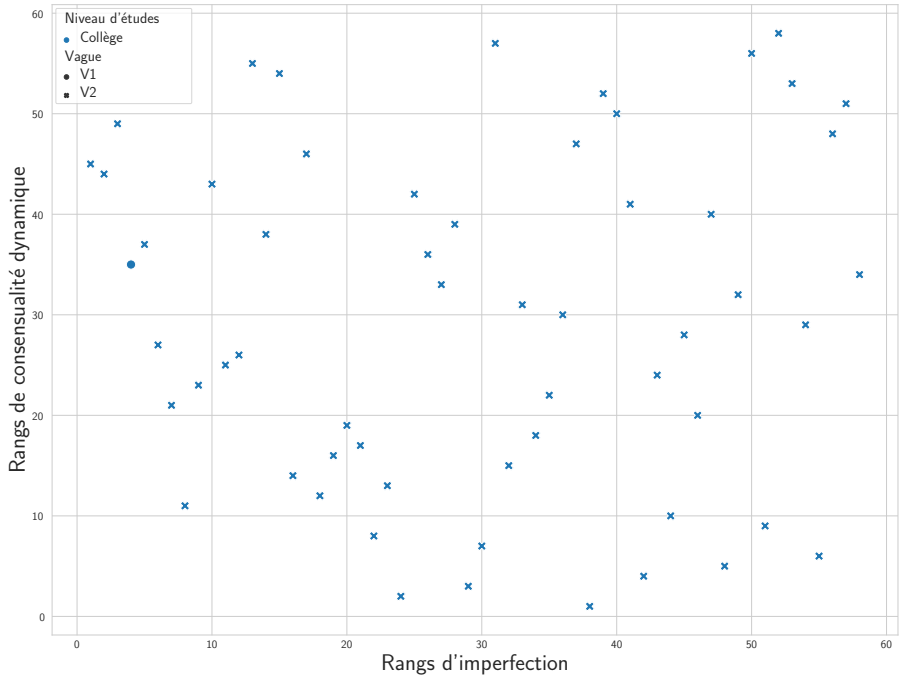
FIGURE 5.8 – Rangs des annotateurs selon leur performance et leur consensualité (consensualité dynamique).

le fait de devoir traiter séparément les cohortes des collégiens et des non-collégiens. Par la suite, les points forment plus ou moins une droite, même si un petit groupe d'annotateurs se détache de cette droite (aux alentours des rangs d'imperfection 40 et 70). Nous notons aussi que les annotateurs des deux vagues se comportent de façon similaire.

Nous nous sommes demandée si le fait que les collégiens soient regroupés parmi les moins consensuels avait un impact sur la consensualité globale, si le degré de consensualité des autres annotateurs changerait de façon significative sans les annotateurs collégiens. Nous avons alors séparé les deux groupes et recalculé leur rang de consensualité. Les résultats de cette expérience sont exposés sur les graphiques de la figure 5.9.

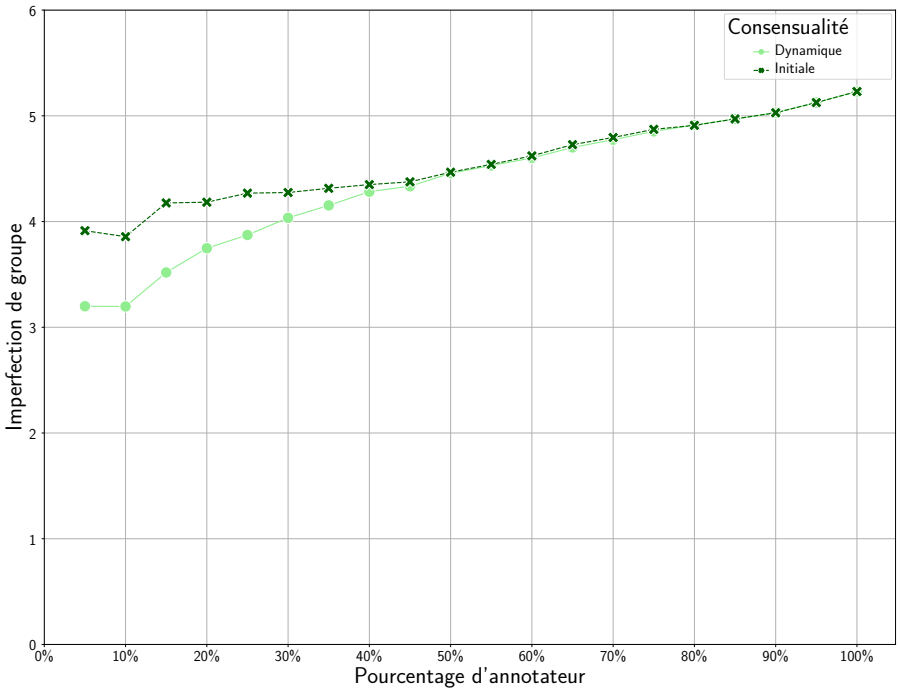


(a) Tous les annotateurs sauf les collégiens.

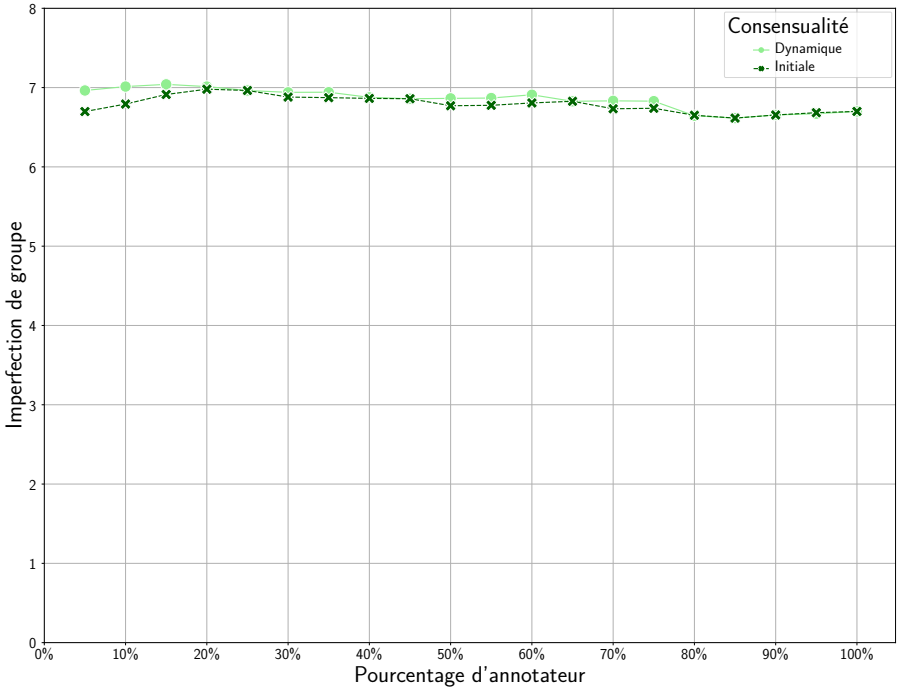


(b) Uniquement les collégiens.

FIGURE 5.9 – Rangs des annotateurs selon leur performance et leur consensualité (consensualité dynamique), sans ou seulement avec les collégiens.



(a) Tous les annotateurs sauf les collégiens.



(b) Uniquement les collégiens.

FIGURE 5.10 – Imperfection de groupe en fonction les n% annotateurs les plus consensuels, pour les deux types de consensualité.

En retirant la cohorte des collégiens de celles des autres niveaux, nous observons une évolution linéaire entre l'augmentation de la performance et celle du rang de consensualité. Ainsi, plus qu'avec la campagne « Portraits », nous voyons ici un rapport entre consensualité et performance. *A contrario*, lorsque nous nous penchons sur le graphique dédié exclusivement aux collégiens, l'analyse est plus délicate à réaliser. En effet, la position des annotateurs est plus chaotique, rendant pour l'heure difficile de tirer des conclusions.

Les graphiques de la figure 5.10 permettent de voir l'évolution de l'imperfection de groupe au fur à et mesure du retrait des annotateurs les moins consensuels, selon les deux types de consensualité.

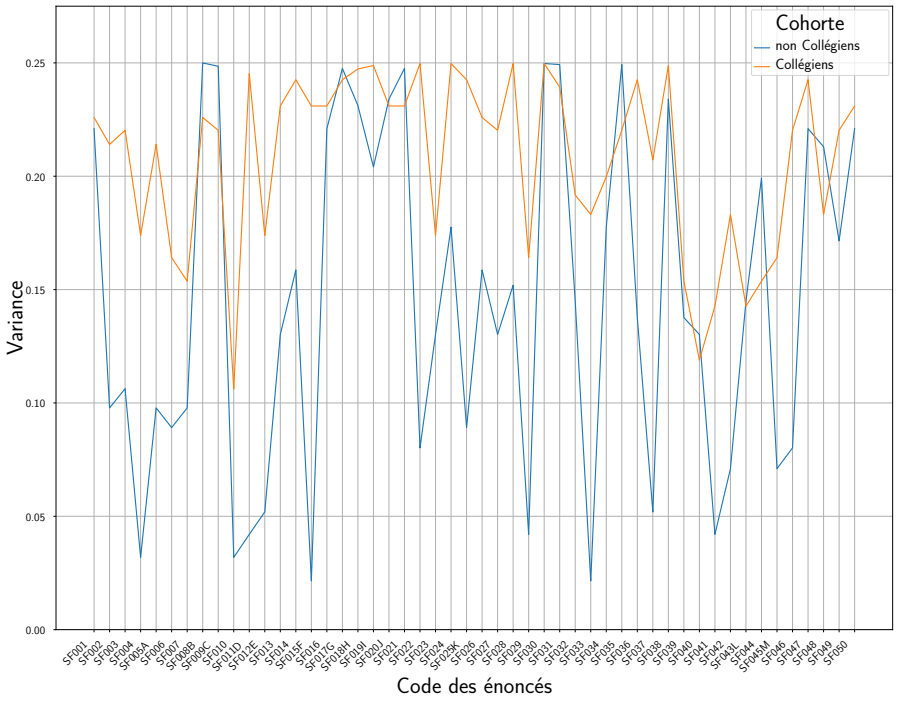
Ainsi, pour la cohorte des annotateurs non collégiens, les courbes sont strictement décroissantes jusqu'à 10 %, ce qui suggère une meilleure performance des annotateurs gardés. Nous observons aussi que les annotateurs les moins consensuels sont les mêmes selon les deux types de consensualités, les deux courbes se confondant jusqu'aux 45 % d'annotateurs restant. Ensuite, une nouvelle fois, la consensualité dynamique permet une meilleure sélection des annotateurs les plus performants. Pour la cohorte des collégiens, en revanche, les courbes sont stables, ré-augmentant même au fil des retraits des annotateurs jugés les plus consensuels.

Ces expériences ont permis d'étendre la consensualité à un autre type de campagne que celle des « Portraits ». Notamment, elles confirment le fait que les annotateurs les moins consensuels sont aussi les moins performants. La consensualité a aussi mis en exergue le fait que les annotations des collégiens étaient délicates à analyser et à utiliser, justifiant ainsi notre décision de séparer les cohortes des collégiens et des non collégiens.

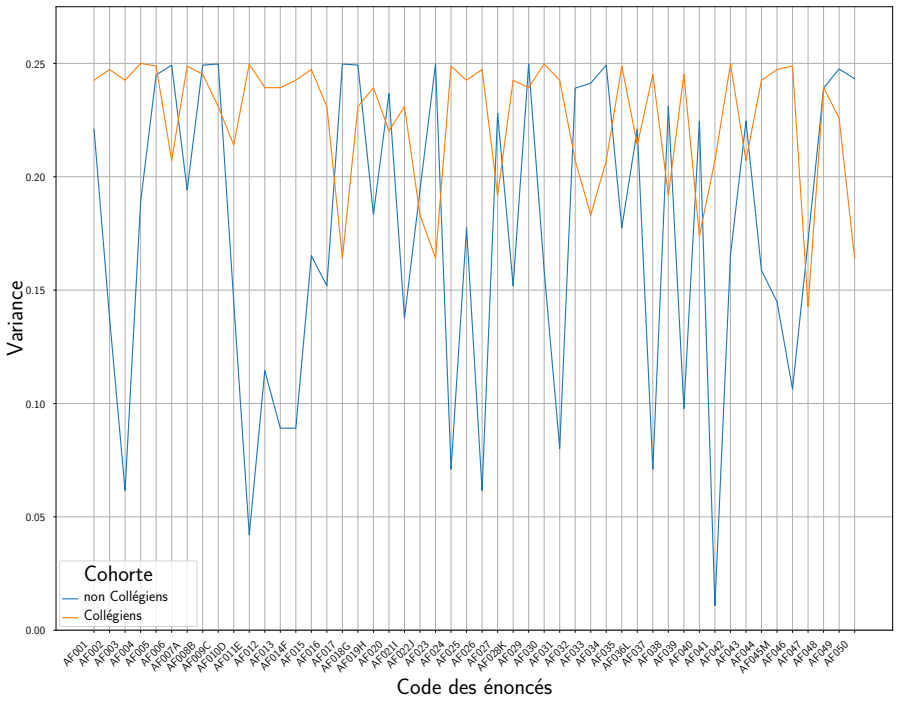
5.4.3 Consensualité des phrases

Nous avons tenu à regarder, pour chaque énoncé, la variance des annotations. Les résultats sont visibles sur les graphiques des figures 5.11a et 5.11b.

Pour rappel, la tâche d'annotation étant une catégorisation binaire, la variance pour une question ne peut pas excéder la valeur de $0,25$, si les annotations sont équitablement réparties, et donc, peu consensuelles. À l'inverse, si les annotateurs sont consensuels, la variance tend vers 0. Nous pouvons voir que, généralement, les annotateurs non collégiens étaient assez consensuels entre eux, autant pour les énoncés *Sans Erreur* que *Avec Erreur*.



(a) Énoncés **Sans Erreur**.



(b) Énoncés **Avec Erreur**

FIGURE 5.11 – Variance pour chaque énoncé, selon la cohorte considérée.

A contrario, les collégiens semblent beaucoup moins consensuels, phénomène davantage visible avec les énoncés *Avec Erreur*.

Code	Énoncé	Variance
AF041	La jeune fille aide à ranger les affaires tombé.	0,010 868
SF010	L'adresse de l'éditeur doit impérativement figurer dans chaque notice bibliographique.	0,106 124
SF033	L'étudiante attend que le chat errant mange.	0,021 495
SF040	Le beau temps a permis de faire une balade à vélo.	0,118 906
SF015L	Cette interprétation du texte mériterait d'être davantage diffusée.	0,021 495
SF043L	Le festival est censé se dérouler au printemps.	0,142 687
SF010	L'adresse de l'éditeur doit impérativement figurer dans chaque notice bibliographique.	0,031 880
SF041	Le jeune homme regarda les éclairs déchirer le ciel nocturne.	0,142 687
SF004	Le dysfonctionnement observé est probablement lié à une incompréhension des instructions.	0,031 880
AF047	La plupart de ces phrases est fausse.	0,142 687
SF030	Il y a quelque 2000 personnes qui se sont rendues à la manifestation	0,249 728
SF022	L'épidémie dégrade les conditions d'enseignement, voire décourage complètement certains étudiants.	0,249 703
AF009C	Les poètes qu'il a entendu chanter en Grèce lui ont donné le sens de la prosodie.	0,249 728
SF028	Sophie, il l'a envoyée sur les roses !	0,249 703
AF017	Cette situation est moins pire que l'autre.	0,249 728
SF030	Il y a quelque 2000 personnes qui se sont rendues à la manifestation	0,249 703
AF029	C'est entre autre pour cette raison qu'elle s'est inscrite en retard.	0,249 728
SF024	Ce manteau coûte quatre cent cinquante euros.	0,249 703
SF008B	Les poésies qu'il a entendu chanter en Grèce lui ont donné le sens de la prosodie.	0,249 970
AF004	Le recours à des procédés réthoriques ne dispense pas du respect de la logique.	0,250 000

TABLE 5.9 – Énoncés dont les réponses sont les plus consensuelles (partie haut du tableau) et les moins consensuelles (partie basse). Les lignes sur fond gris correspondent à la cohorte des non collégiens.

Le tableau présente les énoncés pour lesquels les annotateurs étaient relativement consensuels (ou à l'inverse, les moins consensuels) et ceci pour les deux cohortes. Le premier élément que nous remarquons est la différence entre les variances des phrases les

plus consensuelles selon les deux cohortes : ainsi, si les annotateurs non collégiens ont une variance relativement faible (entre 0,01 et 0,03), la variance des collégiens est déjà conséquente. L'énoncé qui fait le plus consensus parmi cette cohorte atteint déjà 0,11 (soit une répartition de 7 réponses *Avec Erreur* sur 58). Nous notons aussi que les cohortes ne sont pas consensuelles sur les mêmes énoncés, sauf pour la `SF010`. Il en va de même pour les phrases les moins consensuelles, où seule la `SF030` est présente dans les deux groupes. Grâce aux graphiques des figures 5.11a et 5.11b, nous pouvons voir que les phrases `AF009C`, `AF029` et `SF008B` sont toutefois assez peu consensuelles au sein des collégiens.

L'analyse de la variance pour chaque question nous a permis de mettre en lumière certains items pour lesquels les annotateurs sont soit très consensuels, soit pas du tout. Nous avons aussi vu que les collégiens sont assez peu consensuels en comparaison avec les non collégiens.

5.5 Retour arrière possible et paires

5.5.1 La possibilité du retour arrière influence-t-elle les annotations ?

La possibilité, pour les annotateurs, de retourner sur leurs anciennes annotations et de les modifier se révèle un paramètre à prendre en compte pour les fonctionnalités proposées par un logiciel d'annotation. Comme nous l'avons vu en 5.2.2, 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 5.10.

Nous n'observons pas de différence notable 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. Il y a toutefois des écarts-types plus importants avec la modalité SRA.Au regard des écarts-types, nous voyons aussi que les scores obtenus sont disparates au sein des scénarios.

La possibilité du retour arrière ne semble donc pas avoir un impact sur les scores globaux et généraux. Nous regarderons, dans la section suivante, si le retour arrière peut

	ARA	SRA	
S1	72,22 (± 9,66)	74,8 (± 7,98)	73,14 (± 9,18)
S2	70,07 (± 8,18)	73,12 (± 13,22)	71,74 (± 11,33)
S3	69 (± 7,06)	73,86 (± 10,9)	71,72 (± 9,71)
S4	71,07 (± 9,8)	68,29 (± 12,52)	70,14 (± 10,86)
	70,52 (± 8,81)	72,77 (± 12,04)	

TABLE 5.10 – Moyenne et écarts-types des scores pour chaque scénario, pour la cohorte des non collégiens.

plutôt avoir un impact au niveau local, sur des phrases en particulier, notamment grâce aux paires d'énoncés.

5.5.2 Paires d'énoncés

Comme vu dans la partie 5.5.1, la possibilité du retour arrière ne semble pas donner lieu à une annotation de meilleure qualité au niveau global. Nous nous demandons, alors, si l'impact ne serait pas observable à un niveau plus local. Dans cette partie, pour tester cette hypothèse, nous nous concentrons sur les paires d'énoncés. En repérant les paires, les annotateurs auront peut-être tendance à répondre différemment pour chaque énoncé d'une paire. Nous pouvons imaginer par exemple qu'un annotateur ait d'abord répondu *Sans erreur* au premier énoncé d'une paire de phrases ; en voyant la seconde phrase, dans laquelle il ne repère pas non plus d'erreurs, il peut vouloir reconsidérer sa précédente annotation.

Pour tester cette hypothèse, nous calculons, pour chaque annotateur d'un 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 5.11. Par exemple, la case contenant 4,65% 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 (56,64% et 25,95%). Nous remarquons aussi une propension des annotateurs à ne pas repérer d'erreurs.

Énoncé sans / **Énoncé avec**	**Réponse incorrecte**	**Réponse correcte**
Réponse correcte	4,65	56,64
Réponse incorrecte	25,95	12,76

TABLE 5.11 – Comparaison des réponses pour chaque paire d'énoncés (en pourcentage).

5.5.2.1 Par modalité

Nous raffinons ensuite le tableau précédent en distinguant les deux types de campagnes : ARA en 5.12a et SRA en 5.12b. 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 légère augmentation des couples ayant deux réponses différentes (27,4% pour **ARA** et 24,33% pour **SRA**) au détriment de ceux ayant répondu *Sans erreur* aux deux énoncés d'une paire (cela passe de 14,67% pour **SRA** à 11,06% pour **ARA**). Nous supposons que ce phénomène se produit, lorsque après avoir annoté 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 étaient présentées.

Énoncé sans / **Énoncé avec**	**Réponse incorrecte**	**Réponse correcte**
Réponse correcte	4,65	56,89
Réponse incorrecte	27,4	11,06

(a) ARA

Énoncé sans / **Énoncé avec**	**Réponse incorrecte**	**Réponse correcte**
Réponse correcte	4,65	56,35
Réponse incorrecte	24,33	14,67

(b) SRA

TABLE 5.12 – Comparaison des réponses pour chaque paire d'énoncés (en pourcentage) selon le type de campagne.

5.5.2.2 Par scénario

Un autre angle d'attaque est de regarder l'influence de la distance entre les énoncés 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 5.13 ; 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 26,3% et 31,08% des paires). À l'inverse, si les paires sont trop distantes, comme c'est le cas pour les scénarios S1 et S4, les annotateurs ont moins tendance à différencier les paires.

Énoncé avec \ Énoncé sans	Réponse incorrecte	Réponse correcte
Réponse correcte	9,89	55,49
Réponse incorrecte	20,33	14,29

(a) S1

Énoncé avec \ Énoncé sans	Réponse incorrecte	Réponse correcte
Réponse correcte	3,23	59,8
Réponse incorrecte	26,3	10,67

(b) S2

Énoncé avec \ Énoncé sans	Réponse incorrecte	Réponse correcte
Réponse correcte	2,46	57,54
Réponse incorrecte	31,08	8,92

(c) S3

Énoncé avec \ Énoncé sans	Réponse incorrecte	Réponse correcte
Réponse correcte	5,86	51,65
Réponse incorrecte	23,08	19,41

(d) S4

TABLE 5.13 – Comparaison des réponses pour chaque paire d'énoncés (en pourcentage) selon le scénario.

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. Nous devons néanmoins préciser que dans les campagnes réelles, nous avons rarement des énoncés aussi proches que dans nos paires. Il peut, en revanche, y avoir des items ayant des propriétés voisines.

5.5.2.3 Par modalité et par scénario

Pour approfondir davantage cette étude sur les paires de phrases, ainsi que les biais possibles liés à la modalité de présentation, il nous faut analyser les scénarios en regard avec les deux modalités. Dans un premier temps, nous avons étudié plus en détail les scénarios S3 et S4. En effet, l'ordre des énoncés dans ces scénarios étant aléatoire, nous pouvons supposer que les annotateurs étaient moins influencés par le contexte d'apparition des énoncés (les énoncés *Avec erreur* des paires ne sont pas au milieu des autres énoncés *Sans erreur*).

Énoncé avec \ Énoncé sans	Réponse incorrecte	Réponse correcte
Réponse correcte	0,7	60,14
Réponse incorrecte	35,66	3,5

(a) ARA–S3

Énoncé avec \ Énoncé sans	Réponse incorrecte	Réponse correcte
Réponse correcte	3,85	55,49
Réponse incorrecte	27,47	13,19

(b) SRA–S3

Énoncé avec \ Énoncé sans	Réponse incorrecte	Réponse correcte
Réponse correcte	4,95	52,2
Réponse incorrecte	23,63	19,23

(c) ARA–S4

Énoncé avec \ Énoncé sans	Réponse incorrecte	Réponse correcte
Réponse correcte	7,69	50,55
Réponse incorrecte	21,98	19,78

(d) SRA–S4

TABLE 5.14 – Comparaison des réponses pour chaque paire d'énoncés (en pourcentage) pour les scénarios S3 et S4 selon la modalité.

L'observation principale que nous pouvons émettre est que les annotateurs du ARA–S3 ont pratiquement répondu distinctement pour chaque paire d'énoncés quand ils en rencontraient une (dans 95,8% des cas). Cette tendance n'est pas aussi visible lorsque les annotateurs du S3 n'ont pas eu le droit de revenir sur leurs anciennes annotations, et est pratiquement inexistante pour le S4, quelle que soit la modalité. Dans ce dernier cas, la possibilité ou non du retour arrière semble bien avoir un impact sur les annotations.

Grâce aux tableaux 5.15, nous voyons que l'inclination à répondre distinctement aux deux items des paires de phrases se retrouve chez les annotateurs du ARA–S2 — cette proportion atteint effectivement 92,85%. Elle est moins prononcée chez les annotateurs de la modalité SRA.

En revanche, nous remarquons une irrégularité lorsqu'il s'agit d'étudier les pourcentages pour le S1. En effet, les tendances entre les modalités s'inversent : ce sont les annotateurs sans accès au retour arrière qui annotent distinctement les paires de phrases.

Énoncé avec \ Énoncé sans	Réponse incorrecte	Réponse correcte
Réponse correcte	2,75	63,19
Réponse incorrecte	29,67	4,4

(a) ARA–S2

Énoncé avec \ Énoncé sans	Réponse incorrecte	Réponse correcte
Réponse correcte	3,62	57,01
Réponse incorrecte	23,53	15,84

(b) SRA–S2

Énoncé avec \ Énoncé sans	Réponse incorrecte	Réponse correcte
Réponse correcte	11,97	50,43
Réponse incorrecte	19,66	17,95

(c) ARA–S1

Énoncé avec \ Énoncé sans	Réponse incorrecte	Réponse correcte
Réponse correcte	6,15	64,62
Réponse incorrecte	21,54	7,69

(d) SRA–S1

TABLE 5.15 – Comparaison des réponses pour chaque paire d'énoncés (en pourcentage) pour les scénarios S2 et S1 selon la modalité.

Le contexte d'apparition des énoncés peut rentrer en ligne de compte, car, pour rappel, les paires de phrases étaient séparées de 50 énoncés, et au milieu d'autres phrases de la même catégorie (*Sans/Avec erreur*). Nous pouvons aussi émettre une autre raison, liée au nombre d'annotateurs : au total, pour les deux modalités, il n'y a que de 14 annotateurs (9 pour le ARA–S1, 5 pour le SRA–S1). En tout état de cause, les résultats présentés ici ne sont peut-être pas encore stabilisés pour conclure à quelque chose de définitif et nous invitons à traiter les valeurs obtenues et analysées dans cette partie avec prudence.

5.6 Résultats complémentaires

Ces analyses complémentaires ont été réalisées uniquement sur les annotations récoltées lors de la vague 1.

5.6.1 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 de français des participants. Ils devaient aussi indiquer leur niveau d'études — cette question nous permet surtout de retrouver la cohorte d'origine de l'annotateur.

Cette auto-évaluation du niveau de français rejoint la pratique courante, en annota-

tion, de demander leur indice de confiance aux annotateurs, souvent au niveau de l'item. 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 au niveau d'annotations de meilleure qualité. Pour vérifier cette hypothèse, nous avons calculé la moyenne des scores obtenus par les annotateurs selon le niveau de français qu'ils s'étaient attribué, et leur niveau d'études ; le tableau 5.16 expose ces moyennes.

Études \ Français	**1**	**2**	**3**	**4**	**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
	46	60	68,19	71,58	76,43	

TABLE 5.16 – Moyennes des scores selon le niveau d'études et le niveau de français estimé. Les nombres entre parenthèses correspondent au nombre d'annotateurs dans ce sous-ensemble.

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-attribuée 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 a une corrélation entre le niveau estimé de français et le score effectivement obtenu par l'annotateur. Quant au niveau d'études des participants, nous ne voyons pas de corrélation entre le niveau d'études et un score typique de tel ou tel niveau.

L'étape suivante de cette expérience consiste à observer à un niveau plus local, celui de la question, s'il est notamment possible d'identifier les points sur lesquels les annotateurs ont des difficultés et les lier à l'estimation personnelle de leur niveau. Pour ce faire, à la manière de l'expérience suivante 5.6.2, 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.

5.6.2 Taux de réponses correctes par énoncé

Le score moyen tourne autour de 70, comme cela a pu être observé grâce au tableau 5.10. Une question intéressante est de regarder si le taux de réponses correctes est uniforme sur toutes les questions ou si nous pouvons identifier des phénomènes expliquant une disparité de réussite à certaines questions.

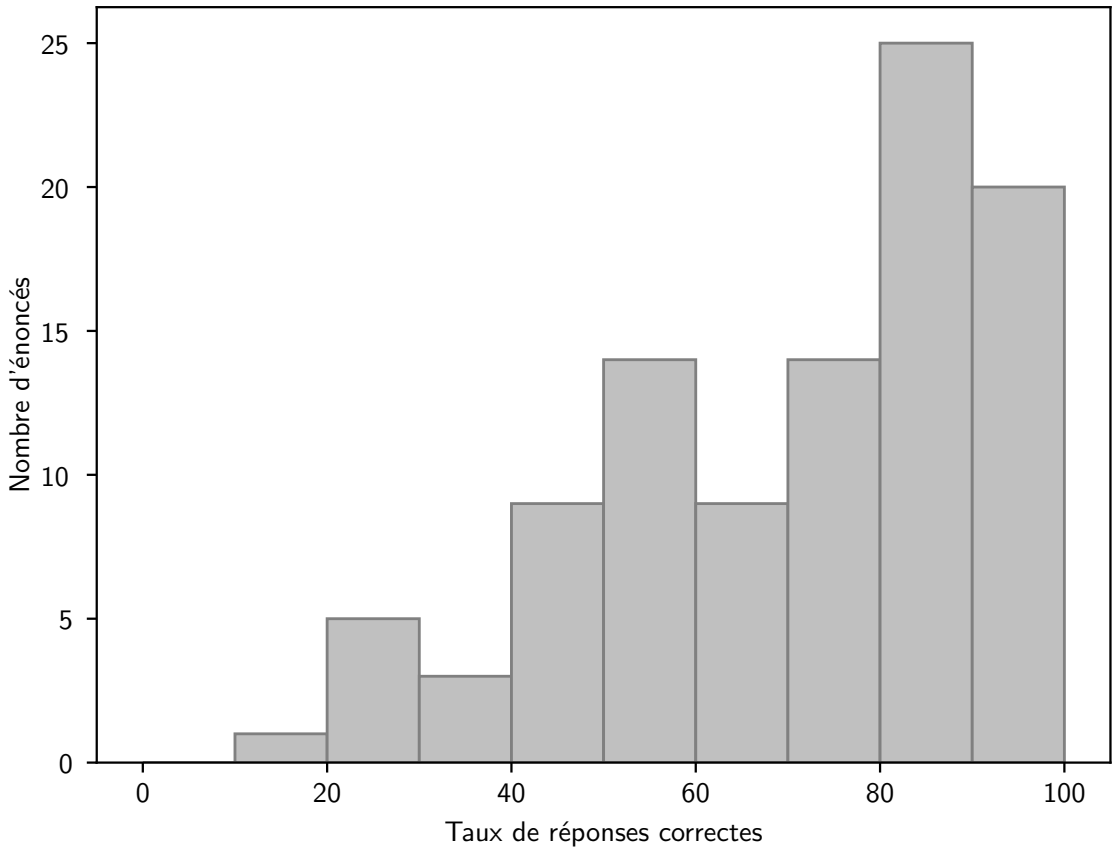


FIGURE 5.12 – Taux de réponses correctes pour chaque énoncé.

Une première approche est de classer les questions en fonction de leur taux de réponses correctes. Nous obtenons alors l'histogramme de la figure 5.12. 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 %.

Au regard de ces groupes, nous pouvons classer les énoncés en trois catégories selon le score observé :

— Plus de 80 % : questions considérées comme faciles ;
— Entre 40 % et 80 % : questions utilisant des règles plus complexes de français ou du vocabulaire plus soutenus ;
— moins de 40 % : questions « pièges » ou énoncés pour lesquels l'usage ne correspond pas à la norme.

Nous avons mis quelques exemples d'énoncés significatifs dans le tableau 5.17.

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 5.17 – Exemples d'énoncés, avec leur catégorie et le taux de bonnes réponses. **AE** correspond à *Avec erreur*, **SE** à *Sans erreur*.

Nous souhaitons à présent regarder plus finement le taux de bonnes réponses pour chaque question, 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 5.13.

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 5.5.2, 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.

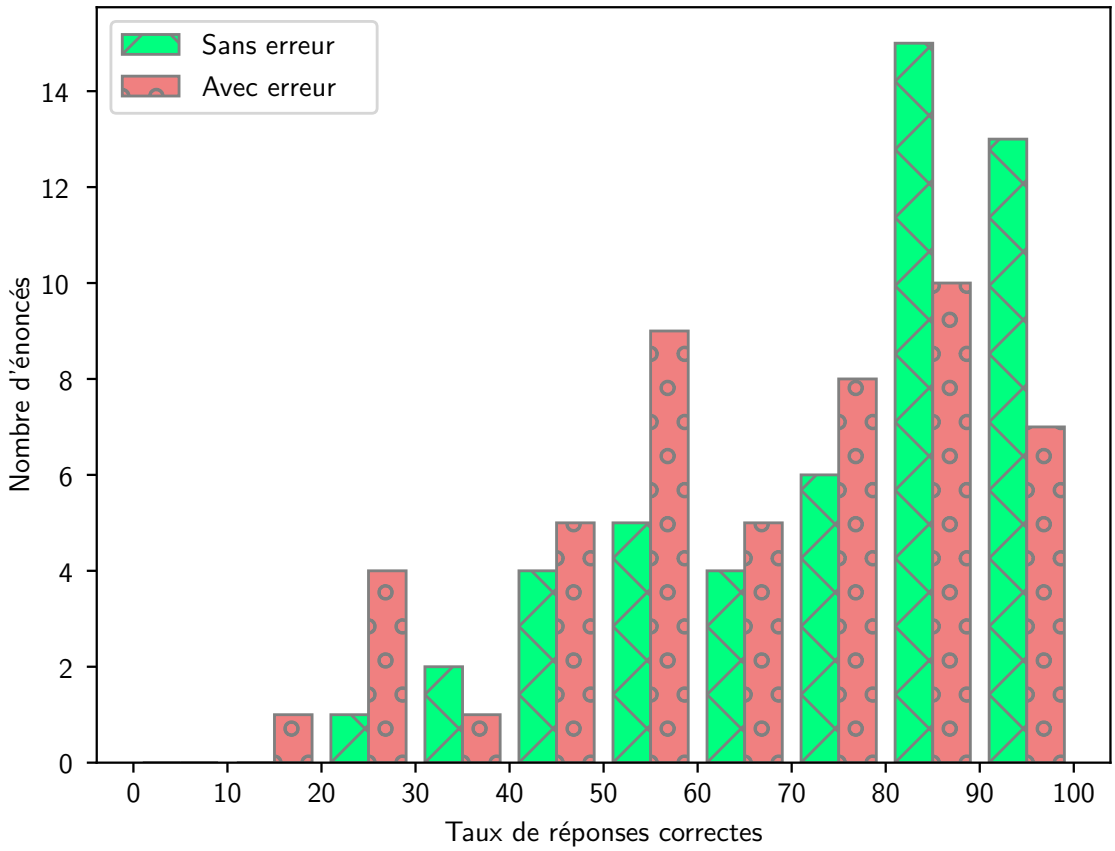


FIGURE 5.13 – Taux de réponses correctes pour chaque énoncé, selon la catégorie.

5.6.3 Outil non adapté pour l'analyse des biais ?

À la suite de la campagne des « Portraits », nous nous étions interrogée sur la plateforme à utiliser pour notre prochaine campagne. Pour rappel, la première plateforme utilisée était le MOODLE de l'université. Elle n'était pas ergonomique pour les annotateurs et ne permettait pas de diffuser facilement et aisément la campagne d'annotation au-delà de la communauté universitaire.

Nous nous sommes tournée vers LIMESURVEY. Cette plateforme possède certains avantages qui nous avait manqués pendant la campagne « Portraits ». L'avantage principal est de fournir un cadre davantage contraint et ergonomique pour la procédure d'annotation : réponses normées, meilleur affichage des questions, paramètres disponibles (retour arrière, question obligatoire, temps de réponse chronométré, etc.). Le questionnaire est accessible à toute personne disposant d'un lien, réglant ainsi le problème d'une diffusion restreinte.

Enfin, il y avait des avantages techniques : gérer les préférences d'affichage, l'attribution automatique du scénario à un annotateur ou encore un meilleur système d'exportation des annotations.

Malheureusement, nous nous sommes aperçue, au fil de nos expériences et analyses, que cette plateforme n'est pas idéale pour l'annotation et nos expérimentations. Nous avons notamment regretté de ne pas avoir accès aux traces d'utilisation de l'annotateur. Cette information nous aurait permis, entre autres, de vérifier si les annotateurs ayant eu les énoncés sur une seule et même page avaient véritablement utilisé cette fonctionnalité, et donc d'étudier un réel impact du retour arrière sur les annotations.

Pour les expériences futures, il conviendrait donc de réfléchir à un environnement dédié à l'annotation, tout en permettant de contrôler et vérifier certains paramètres et comportements de l'annotateur.

5.7 Conclusion

La campagne « Erreurs » s'est révélée fructueuse. Au départ conçue pour mesurer le biais introduit par la possibilité du retour arrière, elle nous a amenée à nous interroger sur la manière de traiter les annotations issues de cohortes hétérogènes. Elle nous a aussi permis d'utiliser la consensualité avec un jeu de données différent, et de confirmer des observations. Enfin, nous avons pu observer un autre biais via les paires d'énoncés.

Une difficulté à laquelle nous avons été confronté au cours de cette campagne est celle d'avoir des annotations issues de cohortes hétérogènes et présentant des disparités. Nous nous sommes demandée quelle serait la meilleure façon de traiter l'ensemble des données, sans pour autant fragiliser les résultats. Nous nous sommes aussi interrogée sur les possibles causes qui ont amené ces annotations disparates.

Cette campagne a été aussi l'occasion de tester la consensualité sur un jeu de données d'un type différent de celui pour lequel la consensualité a été initialement prévue. Ces nouveaux tests de la consensualité ont permis de confirmer les premiers résultats obtenus, à savoir que les annotateurs les moins consensuels sont aussi les moins performants. La consensualité dynamique, une nouvelle fois, démontre qu'elle sélectionne les annotateurs les plus performants.

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. Ceci est particulièrement visible lorsque les deux items des paires sont contiguës. Grâce à ces mêmes paires de phrases, nous avons pu mettre en évidence un phénomène intéressant : le biais provenant du fait que lorsque nous sommes confrontés à deux énoncés presque identiques, nous cherchons instinctivement à les différencier dans une catégorisation binaire[11].

11. 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 ».

Synthèse

Conclusions et perspectives

Au cours de ce mémoire, nous avons dressé un état de l'art des pratiques actuelles pour une campagne d'annotation pour un phénomène linguistique. Ces phénomènes ayant de multiples facettes, nous avons aussi mis en lumière certains points de vigilance, pour alerter les responsables de campagne, et avertir la communauté sur des angles morts. En effet, une campagne regorge de pièges qui chacun, pris individuellement, peut être résolu parfois aisément, mais dont l'absence de prise en considération peut avoir des conséquences néfastes.

En observant une approche insistant sur l'annotateur, qui est irrémédiablement au cœur du processus d'annotation, nous nous sommes interrogée sur la manière d'améliorer la qualité et la fiabilité des données produites, même en ayant peu de données disponibles par exemple si des contraintes limitent leur acquisition. Nous avons mené deux campagnes d'annotation, aux caractéristiques différentes. Chaque campagne était guidée par l'anticipation d'un problème potentiel que nous cherchions à observer finement en essayant de minimiser tous les autres paramètres. La méthodologie utilisée, bien que perfectible, a déjà fait ses preuves, en démontrant par exemple certaines influences sur les annotations.

Le biais : ce qui peut influencer les annotateurs

La colonne vertébrale de ce mémoire est l'étude des biais. Nous avons notamment défini cette notion de biais comme des « phénomènes perturbateurs, de nature et d'origine variées, pouvant survenir à toutes les étapes du processus d'annotation et qui sont susceptibles d'en affecter le résultat ». Nous avons ainsi pu proposer une première classification thématique, selon les grandes étapes d'une campagne d'annotation. Grâce aux deux campagnes d'annotation que nous avons menées, nous nous sommes particulièrement intéressée aux biais impactant les annotateurs, à savoir :

— l'influence de l'ordre des questions, s'il respecte un ordre particulier pouvant influencer l'annotateur ;
— l'influence des caractéristiques de l'annotateur, par exemple un lien potentiel entre l'âge de l'annotateur et ses estimations réalisées lors de la campagne « Portraits » ;
— l'influence de certaines modalités d'interaction de saisie, spécifiquement la possibilité du retour arrière et son impact sur les annotations ;

— l'influence des items proches par leur contenu, et particulièrement grâce à l'inclusion de paires de phrases.

Les travaux engagés sur ces biais nous paraissent encourageants. Les fondamentaux de la méthodologie pour observer l'existence, ainsi que l'impact, d'un biais ont été posés. Ces bases ont permis, notamment, de mettre en exergue le fait que tous les biais n'ont pas tous le même impact, et certains ont une relative importance. Il est difficile de les isoler et de les observer, car ils sont souvent interdépendants et ont des effets parfois légers. Il est donc nécessaire de mener des campagnes dédiées.

Ce travail se révèle principalement être une invitation, lors d'autres campagnes d'annotation futures, à mettre en place des moyens et des procédures pour faire attention à des potentiels biais. Ainsi, nous avons proposé une méthodologie pour de futures campagnes d'annotation, centrées ou non sur l'étude d'un biais ; nous soulignons aussi que l'examen des biais possibles devrait être un préalable à toute campagne. Certaines méthodes peuvent aussi s'appliquer sur d'anciennes campagnes, par exemple pour examiner plus en détail des endroits où les annotations divergent fortement ou pour comparer des cohortes d'annotateurs.

Avec ce que nous avons obtenu, nous avons une première intuition que parmi les biais à prendre particulièrement en compte, se trouve le biais de l'ordre des questions. Ce biais ne peut pas être anticipé dès la conception de la campagne : en effet, il nécessiterait de déjà connaître les annotations attribuées à chaque item. Cependant, il semble pertinent de s'interroger *a posteriori*, lorsque nous observons par exemple qu'un annotateur semble annoter d'une manière régulière présentant un motif.

Un exemple de biais partagé par les annotateurs et les concepteurs de campagne, réside dans les paires de phrases. C'est un biais que nous avons observé, notamment grâce aux paires de phrases, qui comportaient forcément une réponse correcte et une réponse non correcte (voir la section 5.5.2). Il sera intéressant de mener d'autres campagnes avec des catégories binaires et dans lesquelles des paires de phrases sont dans la même catégorie, ou de s'intéresser au cas où la catégorisation n'est pas seulement binaire.

La perspective la plus évidente de ce travail est l'étude de davantage de biais. La création d'un environnement contrôlé d'annotation spécialement conçu pour étudier les biais (leur détection et leurs impacts) serait un atout pour réaliser cela. Idéalement, le logiciel, reprenant le principe d'un outil d'annotation, devrait pouvoir gérer l'annotation

de tout type d'ancrage et de tout type de caractérisation : nous pensons en effet que les biais impactent différemment l'annotation selon l'ancrage et la caractérisation.

Nous attendons d'un tel logiciel qu'il puisse aussi gérer la gestion de scénarios, c'est-à-dire contrôler les paramètres sur lesquels se basera l'approche contrastive (ordre de présentation, présence d'items particuliers, autre modalité, etc.). Lorsque les annotations seront recueillies, il faudra aussi permettre aux responsables de procéder à des analyses élémentaires (moyenne des annotations si cela est possible, variance pour chaque question, calcul de la consensualité, etc.). Les utilisateurs pourraient ajouter des modules complémentaires, s'il y a une omission du logiciel concernant un type d'annotation non pris en charge, ou s'ils veulent accomplir des analyses plus poussées concernant l'influence des biais. Enfin, une collaboration avec des chercheurs en sociologie, en sciences cognitives ou en psychologie est nécessaire : ces disciplines sont en effet habituées aux questions des biais, des questionnaires et des tests en environnement contrôlé. Leur aide apportera une qualité indéniable au projet, afin de créer un environnement neutre et d'identifier les possibles problèmes (biais de sélection, variable sociologique…).

L'élimination complète des biais nous semble impossible, tant le processus d'annotation est complexe et les biais si imbriqués. En revanche, nous pouvons tenter de mieux quantifier leur influence. En effet, au cours de ce mémoire, nous nous sommes restreint à l'étude des biais sur les annotations produites. Une des suites naturelles de cette thèse est de chercher à quantifier l'impact de ces biais sur les chaînes de traitement et des performances des systèmes, à la manière de Millour et al. (2022).

Situer un annotateur par rapport à un groupe

Pendant l'étude des biais, lorsque nous cherchons à traquer les écarts entre les annotateurs, le fait de regarder le comportement des uns et des autres permet de mesurer le décalage. Ainsi, nous avons introduit la notion de consensualité pour observer et quantifier à quel point un annotateur est dissemblable par rapport à un groupe d'annotateurs. La particularité de cette mesure est le fait de s'appuyer sur la capacité de distinguer les annotateurs. Il convient de noter que, à l'inverse des mesures d'accord inter-annotateurs habituelles, une correction par la chance n'est pas nécessaire. La correction par la chance tente, en effet, de comparer les résultats d'une campagne à ceux d'une autre, et d'établir une sorte de norme. Or, au travers de la consensualité, nous ne comparons que des accords

au sein de cette campagne, pas dans l'absolu.

Le calcul de la consensualité n'a toutefois pas vocation à remplacer les mesures d'accord inter-annotateurs : il s'agit d'une proposition complémentaire par rapport à celles-ci. Si les mesures classiques souhaitent rendre compte de l'accord de l'ensemble d'un groupe, la consensualité, elle, s'intéresse à chaque annotateur par rapport au groupe et aide l'identification des annotations en décalage avec le reste du groupe. Dans un futur proche, il est nécessaire de s'interroger sur la meilleure manière de combiner les mesures d'accord et la consensualité, afin d'améliorer le processus d'annotation et l'étude des biais.

Nous avons pour l'instant étudié la pertinence de la consensualité dans un cadre simple : les premiers résultats que nous avons eus sont prometteurs, et déjà utilisables. Nous avons déjà pu observer, lors de nos campagnes de test, un lien entre les annotateurs les moins consensuels et les moins performants, qu'il faudra chercher à généraliser. Les seconds résultats obtenus sur une autre campagne et un autre type d'annotation semblent confirmer les premières observations, et l'ensemble est prometteur. Les moyens d'observation sont, quant à eux, déjà généralisables. Toutefois, il reste un travail à fournir pour confirmer durablement ces observations, et cela sur d'autres jeux de données et différents types d'ancrage et de caractérisation.

L'accord et la consensualité des annotateurs peuvent varier de manière significative d'un item à un autre. Il est alors nécessaire de mettre en place des méthodes permettant de suivre ces variations et d'identifier les items problématiques, procédé que nous avons commencé à mettre en œuvre au cours de la campagne « Erreurs ». La méthode mérite d'être davantage étoffée, ainsi que la question de leur traitement durant la campagne voire durant les traitements ultérieurs. Par exemple, un système d'entraînement ou d'évaluation pourrait moduler la confiance des décisions prises à partir de ces items.

Recommandations

Nous proposons dans cette section de revenir sur les choix et les décisions qui peuvent aider les responsables de campagnes à une meilleure gestion de ces dernières. Au regard de l'état de l'art produit et des problèmes et manquements mis en lumière, ainsi que de nos expériences, nous proposons des recommandations. Ces recommandations restent, bien évidemment, sujettes à précaution, à adaptation et à discussion.

Choisir

Le premier choix auquel les responsables de campagne sont confrontés dépend du phénomène étudié : quoi et comment annoter ? Ces décisions dépendent du positionnement par rapport à un modèle théorique. Bien que Leech (1997) rappelle qu'un schéma d'annotation doit être neutre et ne se rapporter à aucun modèle, être neutre est quasiment impossible, même pour des phénomènes linguistiques qui paraissent les plus consensuels. Faute de pouvoir rester théoriquement neutre, nous pourrions au moins préconiser un affichage clair du modèle théorique sous-jacent.

Concernant la modélisation du phénomène (ancrage et caractérisation), il est difficile de donner des recommandations générales, tant cette étape dépend de la tâche d'annotation. Nous attirons l'attention sur les risques inhérents à la formulation du modèle dans des termes impropres, éventuellement inspirés par l'outil ou le format.

Sur l'aspect de la représentativité du corpus, il s'agit d'une question bien connue des personnes en charge de constituer le corpus de textes à annoter, et à laquelle ces personnes sont sensibilisées. Cela a été largement discuté dans la littérature ; nous pouvons nous reporter à l'étude de Habert (2000).

Les recommandations liées à l'outil concernent surtout la préférence d'utiliser tel ou tel outil. Nous le rappelons une nouvelle fois, mais le logiciel parfait n'existe pas, ou pour très peu de campagnes. Le choix doit s'effectuer sur des critères comme la question de l'expressivité de l'outil, la fidélité au modèle théorique, les fonctionnalités techniques ou encore la facilité d'utilisation. Des grilles de comparaison entre les outils peuvent orienter le choix vers un outil (voir par exemple 5.14). Même si le choix final revient au responsable de campagne, les annotateurs peuvent, et doivent, donner leur avis sur le logiciel.

Le recrutement des annotateurs est un point critique. Dans le cas d'une campagne d'annotation traditionnelle, nous conseillons de recruter des annotateurs experts. Le type d'expertise dépendra cependant de l'objectif de la campagne. S'il s'agit d'une campagne orientée application pour un domaine précis, des experts du domaine seraient plus appropriés pour annoter. À l'inverse, si le but de la campagne est de produire un corpus de référence pour un phénomène linguistique, il faudrait plutôt engager des experts linguistes connaissant bien ce phénomène et ses spécificités.

Pour le nombre d'annotateurs, Neuendorf (2009) recommande au moins deux an-

Comparaison des 3 outils

FIGURE 5.14 – Exemple de grille de comparaison, extraite de HO-DAC et POUDAT (2021).

notateurs pour chaque item.Cependant, si la tâche est complexe ou inédite, si nous ne savons pas encore si l'établissement d'une référence est possible, le mieux serait d'avoir un groupe large d'annotateurs, entre quatre et cinq, voire plus, comme le suggèrent BAYERL et PAUL (2011). Cette réflexion s'appuie essentiellement sur l'hypothèse communément admise que plus il y a d'annotateurs en accord sur une annotation, plus cette annotation est considérée comme valide. Certes, l'augmentation de l'effectif augmente les sources possibles de désaccord, mais, en même temps, seul un accord observé au sein de points de vu distincts sera réellement satisfaisant.

Dans une situation de myriadisation, par principe les annotateurs ne sont pas experts. Il faut compenser leur non-expertise par une augmentation du nombre d'effectif de manière significative. SNOW et al. (2008) estiment que quatre annotateurs non-experts sont nécessaires pour atteindre la qualité de travail d'un expert. Il sera aussi d'autant plus important de repérer dans les cohortes des annotateurs qui ont des comportements fortement divergents, afin de prendre connaissance des raisons de ce comportement.

Un autre point délicat des recommandations concerne l'évaluation des annotations manuelles. Comme nous l'avons vu, il n'existe pas forcément des mesures pour toutes les sortes d'annotation. Pour celles ayant des mesures adaptées, il convient de bien être conscient de l'éventail des mesures, et de choisir la plus pertinente. Par exemple, dans le cadre d'une pure annotation catégorielle, même si le κ de Cohen a fait ses preuves, nous avons pu observer que α de Krippendorff semble être la mesure la plus versatile. Pour les tâches d'*unitizing*, il est d'une façon générale peu importun d'utiliser de mesures qui n'ont pas été nativement prévues pour cela et nous utiliserons préférentiellement des mesures dédiées telles que ${}_u\alpha$ et γ.

Pour les types d'ancrage et de caractérisation pour lesquelles il y a encore une absence de mesures d'accord inter-annotateur, deux solutions sont envisageables :

— utilisation de métriques initialement prévues pour comparer un ensemble de référence avec un ensemble candidat ;

— décomposition de la tâche pour revenir à des tâches ayant des mesures adaptées.

La subdivision de la tâche n'est pas exempte de problèmes, notamment si les phases sont évaluées conjointement, une bonne moyenne d'accord peut cacher des disparités importantes entre les sous-tâches et ayant un très fort impact. Bien sûr, il conviendrait, à terme, de réfléchir à des mesures adaptées pour ces types d'annotations.

Pour l'établissement d'une référence, nous ré-affirmons que la révision des annotations collégiale est la meilleure option pour constituer une référence, comme cela a déjà été souligné par Mathet et Widlöcher (2016), même si elle est coûteuse. Si le vote à l'unanimité et à la majorité sont relativement faciles à mettre en œuvre quand les unités à annoter sont déjà pré-définies, cela est beaucoup plus complexe dans le cadre de l'*unitizing*, comme cela a été vu dans la section 1.6.2.

Être attentif

Pour la rédaction du guide d'annotation, nous listons de nouveau les recommandations fournies par Fort et al. (2009) concernant la rédaction du guide :

— définir précisément les termes, les catégories et justifier les choix effectués ;

— ajouter des exemples ;

— intégrer les potentielles ambiguïtés ;

— préciser l'objectif de la campagne ;

— laisser une part d'interprétation pour les annotateurs.

Nous nuançons toutefois deux points. Le premier concerne les exemples : donner trop d'illustrations est peut-être contre-productif. Cela peut amener l'annotateur à penser que si un cas n'est pas répertorié dans les exemples, alors il n'est pas de cette catégorie. Il faut donc donner des exemples en quantité raisonnable. Les deux recommandations stipulant de définir précisément les termes tout en laissant une part d'interprétation aux annotateurs peuvent paraître contradictoires, et le point d'équilibre entre manque et excès est parfois difficilement atteignable.

La question de la longueur du guide d'annotation se pose aussi. Doit-il être court, pour être plus aisé à consulter, ou au contraire, être autant détaillé que possible (tout en restant dans la limite du raisonnable) ? S'il est bien structuré et précis, ainsi que facilement consultable (par exemple des liens et des renvois internes au document), la longueur ne sera pas forcément un défaut. De plus, il est important que les définitions des termes et des catégories soient clairement identifiables, pour que les annotateurs puissent les repérer rapidement ; il en va de même pour les exemples.

Nous rappelons que la première version du guide d'annotation sera forcément imparfaite et doit être modifiée au fil des annotations et des discussions avec les annotateurs, ainsi qu'à la lumière des mesures d'accord inter-annotateurs.

Le format des annotations dépend certes du logiciel, mais relève aussi d'une décision pouvant avoir une conséquence sur la diffusion. Plus les annotations seront interopérables, plus le corpus sera « utile ». Si les annotations sont difficiles à extraire ou à utiliser, il se peut que le corpus ne soit finalement peu utilisé. Il convient donc de choisir un format de données assez répandu, ou du moins un format facilement manipulable.

Documenter

Un point mis en avant par les travaux de cette thèse est qu'il est essentiel de documenter tout le processus d'une campagne d'annotation, tous les choix et leurs justifications. Nous insistons plus particulièrement sur le fait de conserver toutes les annotations, d'une part dans un but de reproductibilité des expériences, d'autre part parce que l'établissement d'une référence n'est pas toujours la meilleure solution. Ainsi, nous pouvons fournir une référence stabilisée, tout en diffusant aussi l'ensemble des annotations produites au moment de cette référence.

Le guide d'annotation constitue un élément de documentation central. Il permet d'une part d'expliciter les choix réalisés quant au schéma d'annotation et au positionnement relatif à une école, d'autre part d'être le reflet de l'évaluation de la campagne et des dialogues entre les annotateurs et les responsables. Certaines informations doivent figurer dans le guide d'annotation si elles concernent les annotateurs, d'autres, qui les concerneront moins ou qu'éventuellement nous souhaitons lui masquer, devront néanmoins figurer dans un document nécessaire à la reproduction des expériences (par exemple, il aurait été contre-productif de prévenir les annotateurs de l'existence de scénarios différents dans le cadre de nos expériences).

Il convient de noter la source et la licence pour chaque texte du corpus. Une description de chaque texte du corpus constitue une plus-value pour la réutilisation des corpus (autant la collection de textes bruts que celle avec les annotations). Pour les corpus oraux, cette description peut être l'occasion de renseigner les informations liées aux locuteurs et à la situation d'énonciation.

Il n'y a pas de référence absolue

Une seule vérité ?

Il peut exister plusieurs modèles, plusieurs façons de percevoir un phénomène, et donc plusieurs vérités d'annotation. Un même item ne sera pas annoté de la même manière selon le modèle adopté ou les spécification de l'outil d'annotation, et ces annotations seront quand même correctes d'un point de vue de la « vérité ». Pour un même modèle et avec un même outil, il peut y avoir une importante variabilité. Cela sera évidemment le cas si le but de la tâche varie, par exemple pour déterminer l'âge d'une personne : nous pouvons vouloir donner soit l'âge biologique (à un médecin, par exemple), soit l'âge perçu (à un policier, pour dresser un portrait robot). Toutefois, même à but constant, il y a une variabilité irréductible : par exemple, pour l'annotation de passages chargés d'opinion, pour la description d'un même évènement, un annotateur peut y voir une description objective, alors qu'un autre peut y déceler la trace d'une idéologie.

Pour une illustration plus littéraire, nous pouvons citer une tâche de classification de textes par genre de l'imaginaire (science-fiction, fantasy, fantastique). Selon les catégories disponibles (limitées aux trois genres principaux ou aux sous-genres disponibles) et la

sensibilité de l'annotateur, une œuvre telle que *Dune* (Hebert, 1965) peut être classifiée comme uniquement de la *Science-Fiction*, mais aussi comme *Science-Fantasy*. Il en va de même pour la définition du fantastique prise en compte pour la tâche : est-ce la définition stricte du terme comme définie par (Todorov, 1970), ou une définition élargie ces dernières années et se confondant avec de la *low fantasy* (Besson, 2020 ; de Palmas Jauze, 2014) ? Ce problème est un cas particulier de recouvrement entre les catégories, car même s'il y a des critères formels de distinction entre ces genres, il peut arriver que l'auteur en joue pour construire une œuvre ambiguë ou se revendique lui-même d'un autre genre. Une partie peut être clarifiée dans le guide d'annotation, mais certaines ambiguïtés sont inhérentes aux annotateurs, aux termes utilisés ou encore au contexte.

Établir une référence suppose donc de définir ce que l'on souhaite atteindre selon la tâche d'annotation et le corpus qui en sera le résultat. La conception de la référence — ou de la vérité — va donc dépendre de l'objectif et des caractéristiques de la campagne (Habert, 2000 ; Leech, 2005).

Une référence, ou des références ?

Généralement, une campagne d'annotation est réalisée afin d'obtenir une référence unique, pour développer les études et les outils sur un phénomène linguistique. À cet effet, les responsables de campagne diffusent uniquement un corpus de référence, sans proposer les autres annotations qui ont conduits à son établissement.

Aroyo et Welty (2015) soulignent toutefois que considérer une seule vérité pour un item n'est pas toujours correct, du moins pour certains exemples. Les auteurs citent par exemple le cas la tâche d'annotation de relations UMLS [12], pour lesquelles plusieurs interprétations sont acceptables. Dans ce cas, comment gérer de tels items lors de l'établissement d'une référence ? Ou, à défaut de pouvoir établir une seule référence, ne faudrait-il pas mieux partager l'ensemble des annotations, cette diversité de points de vue constituant alors la référence ?

Ce besoin d'exhaustivité ne se manifeste pas forcément avec autant d'intensité selon les réutilisations du corpus. Pour une première approche du corpus ou du phénomène, la référence obtenue permet d'obtenir des résultats tout à fait raisonnables, que l'ensemble des annotations permet de consolider. Si l'utilisation du corpus s'intègre dans une suite

12. Voir www.nlm.nih.gov/research/umls/.

logique d'application, par exemple pour de l'apprentissage, la référence est un point de comparaison pour différents systèmes et l'ensemble des annotations permet de déceler d'éventuels biais ou manques du corpus de référence. Enfin, les utilisateurs du corpus peuvent aussi vouloir construire leur propre référence, si par exemple les chercheurs veulent seulement travailler sur les cas problématiques.

Une bonne pratique pour la communauté est de ne pas se restreindre à la diffusion du seul corpus de référence, mais de l'accompagner de l'ensemble des annotations et discussions qui ont permis de l'obtenir. Certains responsables de campagne diffusent déjà l'ensemble des annotations. Cela est notamment le cas du corpus Annodis (Péry-Woodley et al., 2011). Nous pensons que cette pratique mériterait d'être plus répandue. Ceci est important afin que les futurs utilisateurs du corpus soient éclairés sur la globalité du processus d'annotation et des potentiels biais. En effet, disposer de toutes les données permet à la fois de se confronter à la méthode utilisée, de reproduire les expérimentations et d'être conscient des éléments litigieux et non fiables, la fiabilité dépendant du contexte.

Se donner les moyens d'observer

Les expérimentations réalisées dans le cadre de cette thèse avaient pour objectif principal de poser les bases d'une méthodologie pour observer ce qui perturbe potentiellement l'annotation manuelle ou le lien entre l'annotation et la réalité qu'elle devrait donner à voir. L'apport fondamental de ce mémoire demeure donc une invitation à l'attention des responsables de campagne pour se donner les moyens de regarder finement l'effet de tel ou tel paramètre sur l'annotation. Les responsables doivent en effet se poser, dès la conception et à toutes les étapes d'une campagne, la question de la présence ou de l'impact de biais potentiels. Sans que la campagne soit nécessairement dédiée à l'étude de biais, la répartition en différentes cohortes d'annotateurs peut permettre de déceler des biais auxquels nous ne nous serions pas attendus ; cela suppose toutefois que nous disposons d'éléments connus de différenciation entre les annotateurs, les items, les parties du corpus, etc.

Ce qui a été effectué lors des campagnes de cette thèse peut parfaitement être reproduit pour d'autres campagnes. Avec l'ensemble des annotations, une ou plusieurs analyses statistiques peuvent être réalisées et permettre de mettre au jour *a posteriori* des biais

ou des soucis variés qui auraient pu en affecter la stabilité et la fiabilité. À ce sujet, nous invitons la communauté à développer des outils de validation statistiques génériques qui pourraient être appliqués à l'ensemble des campagnes d'annotation manuelle.

La consensualité constitue typiquement un de ces outils de validation et est une autre pierre à l'édifice. Elle permet de mettre en exergue le comportement individuel et de repérer des dissonances fortes au sein d'un groupe d'annotateurs. À présent, il faut poursuivre les expérimentations pour essayer de voir dans quelle mesure nous pouvons généraliser ces observations.

Bibliographie

Abbott, R., Walker, M., Anand, P., Tree, J. E. F., Bowmani, R. & King, J., (2011), How can you say such things ?!?: Recognizing disagreement in informal political argument, In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, Portland, Oregon, États-Unis.

Abeillé, A. (Éd.), (2003), *Treebanks*, Springer Netherlands, https://doi.org/10.1007/978-94-010-0201-1

Abeillé, A., Clément, L. & Liégeois, L., (2019), Un corpus annoté pour le français : le French Treebank, *Revue TAL*, *602*, 19-43, https://halshs.archives-ouvertes.fr/halshs-02560207

Abeillé, A., Clément, L. & Toussenel, F., (2003), Building a treebank for French. In *Treebanks* (p. 165-187), Springer.

Afantenos, S., Asher, N., Benamara, F., Bras, M., Fabre, C., Ho-dac, M., Draoulec, A. L., Muller, P., Péry-Woodley, M.-P., Prévot, L., Rebeyrolles, J., Tanguy, L., Vergez-Couret, M. & Vieu, L., (2012), An empirical resource for discovering cognitive principles of discourse organisation : the ANNODIS corpus, In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turquie, European Language Resources Association (ELRA), http://www.lrec-conf.org/proceedings/lrec2012/pdf/836_Paper.pdf

Ågren, M., (2008), *À la recherche de la morphologie silencieuse : sur le développement du pluriel en francais L2 écrit* (thèse de doct.), Lund University.

Amalvy, A., Labatut, V. & Dufour, R., (2022), Remplacement de mentions pour l'adaptation d'un corpus de reconnaissance d'entités nommées à un domaine cible (Mention replacement for adapting a named entity recognition dataset to a target domain), In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, Avignon, France, ATALA, https://aclanthology.org/2022.jeptalnrecital-taln.19

Amidei, J., Piwek, P. & Willis, A., (2018), Rethinking the Agreement in Human Evaluation Tasks, In *Proceedings of the 27th International Conference on Compu-*

tational Linguistics, Santa Fe, Nouveau-Mexique, États-Unis, http://oro.open.ac.uk/56443/

Antoine, J.-Y., Schang, E., Muzerelle, J., Lefeuvre, A., Pelletier, A., Désoyer, A., Landragin, F., Tellier, I., Villaneau, J., Eshkol, I. & Maurel, D., (2014), *Corpus ANCOR-Corpus : Présentation générale* (rapp. tech.), Université François Rabelais et Université d'Orléans, https://tln.lifat.univ-tours.fr/medias/fichier/ancor-centre_1562920591452-pdf?ID_FICHE=322825&INLINE=FALSE

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

Aroyo, L. & Welty, C., (2015), Truth Is a Lie : Crowd Truth and the Seven Myths of Human Annotation, *AI Magazine*, *361*, 15-24, https://doi.org/10.1609/aimag.v36i1.2564

Artstein, R. & Poesio, M., (2008), Inter-Coder Agreement for Computational Linguistics, *Comput. Linguist.*, *344*, 555-596, https://doi.org/10.1162/coli.07-034-R2

Asher, N., Nasr, A. & Perrotin, R., (2017), *Manuel d'annotation en actes de dialogue pour le corpus Datcha* (rapp. tech.).

Austin, J. L., (1975), *How to do things with words*, Oxford University Press.

Azé, J., Heitz, T., Roche, M., Mela, A., Peinl, P. & Amar Djalil, M., (2006), Présentation de DEFT 06 (DÉfi Fouille de Textes), In *Atelier DEFT'06 - SDN'06 (Semaine du Document Numérique)*, Fribourg, Suisse, https://hal.inria.fr/lirmm-00113164

Bagga, A. & Baldwin, B., (1998), Algorithms for Scoring Coreference Chains, In *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, Grenade, Espagne.

Baldridge, J., Asher, N. & Hunter, J., (2007), Annotation for and robust parsing of discourse structure on unrestricted texts, *Zeitschrift für Sprachwissenschaft*, *262*, 213-239.

Baledent, A., (2022), Impact des modalités induites par les outils d'annotation manuelle : exemple de la détection des erreurs de français (Impact of modalities induced by manual annotation tools : example of French error detection), In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2 : 24e Rencontres Etudiants Chercheurs en Informatique pour le TAL (RECITAL)*, Avignon, France, ATALA, https://aclanthology.org/2022.jeptalnrecital-recital.7

Baledent, A., Mathet, Y., Widlöcher, A., Couronne, C. & Manguin, J.-L., (2022), Validity, Agreement, Consensuality and Annotated Data Quality, In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France, European Language Resources Association, https://aclanthology.org/2022.lrec-1.315

Bayerl, P. S. & Paul, K. I., (2011), What Determines Inter-Coder Agreement in Manual Annotations ? A Meta-Analytic Investigation, *Computational Linguistics*, *374*, 699-725, https://doi.org/10.1162/COLI_a_00074

Bazillon, T., (2011), *Transcription et traitement manuel de la parole spontanée pour sa reconnaissance automatique* (Thèse 2011LEMA3003), Université du Maine, https://theses.hal.science/tel-00598427

Beck, K., (2011), The Agile Manifesto, http://agilemanifesto.org/

Bejček, E. & Straňák, P., (2010), Annotation of multiword expressions in the Prague dependency treebank, *Language Resources and Evaluation*, *441*, 7-21.

Bender, E., (2019), The #BenderRule : On naming the languages we study and why it matters, https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/

Bender, E. M. & Friedman, B., (2018), Data Statements for Natural Language Processing : Toward Mitigating System Bias and Enabling Better Science, *Transactions of the Association for Computational Linguistics*, *6*, 587-604, https://doi.org/10.1162/tacl_a_00041

Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S., (2021), On the Dangers of Stochastic Parrots : Can Language Models Be Too Big ?, In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event, Canada, Association for Computing Machinery, https://doi.org/10.1145/3442188.3445922

Bennett, E. M., Alpert, R. & Goldstein, A., (1954), Communications through limited-response questioning, *Public Opinion Quarterly*, *183*, 303-308.

Benzitoun, C., Fort, K. & Sagot, B., (2012), TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe, In *JEP-TALN 2012 - Journées d'Études sur la Parole et conférence annuelle du Traitement Automatique des Langues Naturelles*, Grenoble, France, https://hal.archives-ouvertes.fr/hal-00709187

Besson, A., (2020), Les sous-genres de la fantasy, un vaste monde à découvrir, https://fantasy.bnf.fr/comprendre/les-sous-genres-de-la-fantasy-un-vaste-monde-decouvrir/

Bhardwaj, V., Passonneau, R., Salleb-Aouissi, A. & Ide, N., (2010), Anveshan : A Framework for Analysis of Multiple Annotators' Labeling Behavior, In *Proceedings of the Fourth Linguistic Annotation Workshop*, Uppsala, Suède, Association for Computational Linguistics, https://aclanthology.org/W10-1806

Biber, D., (1993), Representativeness in corpus design, *Literary and linguistic computing*, *84*, 243-257.

Böhmová, A., Hajič, J., Hajičová, E. & Hladká, B., (2003), The Prague Dependency Treebank, In A. Abeillé (Éd.), *Treebanks : Building and Using Parsed Corpora* (p. 103-127), Dordrecht, Pays-Bas, Springer Netherlands, https://doi.org/10.1007/978-94-010-0201-1_7

Bonneau-Maynard, H., Rosset, S., Ayache, C., Kuhn, A. & Mostefa, D., (2005), Semantic annotation of the French media dialog corpus, In *INTERSPEECH*.

Boré, C. & Elalouf, M.-L., (2017), Deux étapes dans la construction de corpus scolaires : problèmes récurrents et perspectives nouvelles, *Corpus*, *16*, 31-63, https://doi.org/10.4000/corpus.2731

Bradley, M. M. & Lang, P. J., (1999), *Affective norms for English words (ANEW) : Instruction manual and affective ratings* (rapp. tech.), The Center for Research in Psychophysiology, University of Florida.

Braffort, A., Chételat-Pelé, E. & Segouat, J., (2011), Corpus de langue des signes : situer les biais des méthodes d'annotation et d'analyse, *Corpus*, *10*, 25-40, https://journals.openedition.org/corpus/1992#authors

Brandsen, A., Verberne, S., Wansleeben, M. & Lambers, K., (2020), Creating a Dataset for Named Entity Recognition in the Archaeology Domain, In *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, European Language Resources Association, https://aclanthology.org/2020.lrec-1.562

Bregeon, D., Antoine, J.-Y., Villaneau, J. & Lefeuvre-Halftermeyer, A., (2019), Redonner du sens à l'accord interannotateurs : vers une interprétation des mesures d'accord en termes de reproductibilité de l'annotation, *Traitement Automatique des Langues*, *602*, 23, https://hal.archives-ouvertes.fr/hal-02375240

Buchholz, S. & Marsi, E., (2006), CoNLL-X Shared Task on Multilingual Dependency Parsing, In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, New-York, États-Unis, Association for Computational Linguistics, https://aclanthology.org/W06-2920

Buhmann, J., Caspers, J., van Heuven, V. J., Hoekstra, H., Martens, J.-P. & Swerts, M., (2002), Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus, In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Îles Canaries, Espagne, European Language Resources Association (ELRA), http://www.lrec-conf.org/proceedings/lrec2002/pdf/96.pdf

Callison-Burch, C. & Dredze, M., (2010), Creating speech and language data with amazon's mechanical turk, In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*, Los Angeles, Californie, États-Unis.

Candito, M., Constant, M., Ramisch, C., Savary, A., Guillaume, B., Parmentier, Y. & Cordeiro, S. R., (2020), A French corpus annotated for multiword expressions and named entities, *Journal of Language Modelling*, *82*, 415-479, https://doi.org/10.15398/jlm.v8i2.265

Candito, M., Constant, M., Ramisch, C., Savary, A., Parmentier, y., Pasquer, C. & Antoine, J.-y., (2017), Annotation d'expressions polylexicales verbales en français (J.-Y. A. Iris Eshkol, Éd.), In J.-Y. A. Iris Eshkol (Éd.), *24e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Orléans, France, https://hal.archives-ouvertes.fr/hal-01537880

Candito, M., Perrier, G., Guillaume, B., Ribeyre, C., Fort, K., Seddah, D. & de la Clergerie, É., (2014), Deep Syntax Annotation of the Sequoia French Treebank, In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Islande, European Language Resources Association (ELRA), http://www.lrec-conf.org/proceedings/lrec2014/pdf/494_Paper.pdf

Candito, M. & Seddah, D., (2012), Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical, In *TALN 2012 - 19e conférence sur le Traitement Automatique des Langues Naturelles*, Grenoble, France, https://hal.inria.fr/hal-00698938

Carletta, J., (1996), Assessing Agreement on Classification Tasks : The Kappa Statistic, *Computational Linguistics*, *222*, 249-254, https://aclanthology.org/J96-2004

Carletta, J., Isard, A., Isard, S., Kowtko, J. C., Doherty-Sneddon, G. & Anderson, A. H., (1997), The Reliability of a Dialogue Structure Coding Scheme, *Computational Linguistics*, *231*, 13-31, https://aclanthology.org/J97-1002

Carlson, L., Marcu, D. & Okurowski, M. E., (2002), RST Discourse Treebank, Linguistic Data Consortium.

Caubrière, A., Rosset, S., Estève, Y., Laurent, A. & Morin, E., (2020), Where are we in Named Entity Recognition from Speech ?, In *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, European Language Resources Association, https://aclanthology.org/2020.lrec-1.556

Chamberlain, J., Fort, K., Kruschwitz, U., Lafourcade, M. & Poesio, M., (2013), Using games to create language resources : Successes and limitations of the approach. In *The People's Web Meets NLP* (p. 3-44), Springer.

Chamberlain, J., Kruschwitz, U. & Poesio, M., (2009), Constructing an Anaphorically Annotated Corpus with Non-Experts : Assessing the Quality of Collaborative Annotations, In *Proceedings of the 2009 Workshop on The People's Web Meets NLP : Collaboratively Constructed Semantic Resources (People's Web)*, Suntec, Singapour, Association for Computational Linguistics, https://aclanthology.org/W09-3309

Chinchor, N. & Sundheim, B., (1993), MUC-5 Evaluation Metrics, In *Fifth Message Understanding Conference (MUC-5) : Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*, https://aclanthology.org/M93-1007

Chiril, P., Moriceau, V., Benamara, F., Mari, A., Origgi, G. & Coulomb-Gully, M., (2020), An Annotated Corpus for Sexism Detection in French Tweets, In *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, European Language Resources Association, https://aclanthology.org/2020.lrec-1.175

Clifford, C. W., Watson, T. L. & White, D., (2018), Two sources of bias explain errors in facial age estimation, *Royal Society open science*, *510*, 180841.

Cohen, J., (1960), A coefficient of agreement for nominal scales, *Educational and psychological measurement*, *201*, 37-46.

Cohen, J., (1968), Weighted kappa : nominal scale agreement provision for scaled disagreement or partial credit., *Psychological bulletin*, *704*, 213.

Couillault, A. & Fort, K., (2013), Charte Éthique et Big Data : parce que mon corpus le vaut bien ! [4 pages], In *Linguistique, Langues et Parole : Statuts, Usages et Mésusages*, Strasbourg, France, 4 pages, https://hal.archives-ouvertes.fr/hal-00820352

Dandapat, S., Biswas, P., Choudhury, M. & Bali, K., (2009), Complex linguistic annotation–no easy way out ! A case from Bangla and Hindi POS labeling tasks, In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, Suntec, Singapour.

Davies, M. & Fleiss, J. L., (1982), Measuring agreement for multinomial data, *Biometrics*, 1047-1051.

Dawid, A. P. & Skene, A. M., (1979), Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm, *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, *281*, 20-28, https://doi.org/10.2307/2346806

Day, D., McHenry, C., Kozierok, R. & Riek, L., (2004), Callisto : A Configurable Annotation Workbench, In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbonne, Portugal, European Language Resources Association (ELRA), http://www.lrec-conf.org/proceedings/lrec2004/pdf/612.pdf

Delaborde, M., (2020), *Analyse en corpus de chaînes de coréférence : la coréférence non-stricte à l'épreuve de la linguistique outillée* (Thèse 2020PA030073), Université de la Sorbonne nouvelle - Paris III, https://tel.archives-ouvertes.fr/tel-03425446

de Mazancourt, H., Couillault, A., Adda, G. & Recourcé, G., (2015), Faire du TAL sur des données personnelles : un oxymore ?, In *Actes de la 1e Ethique et TRaitemeNt Automatique des Langues*, Caen, France, Association pour le Traitement Automatique des Langues, http://www.atala.org/taln_archives/ETERNAL/ETERNAL-2015/eternal-2015-long-003

de Palmas Jauze, D., (2014), *Les dragons de la Fantasy : Les du passé et renouveau*, Paris, France, Éditions du Panthéon.

Di Eugenio, B. & Glass, M., (2004), Squibs and Discussions : The Kappa Statistic : A Second Look, *Computational Linguistics*, *301*, 95-101, https://doi.org/10.1162/089120104773633402

Doran, C., Aberdeen, J., Damianos, L. & Hirschman, L., (2001), Comparing Several Aspects of Human-Computer and Human-Human Dialogues, In *Proceedings of the*

Second SIGdial Workshop on Discourse and Dialogue, Aalborg, Danemark, https://aclanthology.org/W01-1607

Ehrmann, M., (2008), *Les Entitées Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation* (Thèse), Paris Diderot University, https://hal.archives-ouvertes.fr/tel-01639190

Ehrmann, M. & Rosset, S., (2018), Entités nommées [[Online ; accessed 20. Jul. 2022]], https://bigdataspeech.github.io/EN/resume/2018/06/21/summary.html

Eshkol, I., Baude, O., Kanaan, L., Maurel, D. & Dugua, C., (2014), " Procédure d'anonymisation et traitement automatique : l'expérience d'ESLO ", In *Journée d'études ATALA, Ethique et TAL*, Paris, France, https://halshs.archives-ouvertes.fr/halshs-01165957

Eshkol-Taravella, I., (2015), *La définition des annotations linguistiques selon les corpus : de l'écrit journalistique à l'oral* (Habilitation à diriger des recherches), Université d'Orléans, https://hal.archives-ouvertes.fr/tel-01250650

Finn, R. H., (1970), A note on estimating the reliability of categorical data, *Educational and psychological measurement*, *301*, 71-76.

Fisas, B., Ronzano, F. & Saggion, H., (2016), A Multi-layered Annotated Corpus of Scientific Papers, *ELRA (European Language Resources Association)*.

Fischer, J. & Whitney, D., (2014), Serial dependence in visual perception, *Nature Neuroscience*, *175*, 738-743, https://doi.org/10.1038/nn.3689

Fiumara, J., Cieri, C., Wright, J. & Liberman, M., (2020), LanguageARC : Developing Language Resources Through Citizen Linguistics, In *Proceedings of the LREC 2020 Workshop on "Citizen Linguistics in Language Resource Development"*, Marseille, France, European Language Resources Association, https://aclanthology.org/2020.cllrd-1.1

Fleiss, J. L., (1971), Measuring nominal scale agreement among many raters., *Psychological bulletin*, *765*, 378.

Fort, K., (2016), *Collaborative Annotation for Reliable Natural Language Processing : Technical and Sociological Aspects*, Hoboken, New Jersey, États-Unis, Wiley, https://www.wiley.com/en-ai/Collaborative+Annotation+for+Reliable+Natural+Language+Processing%3A+Technical+and+Sociological+Aspects-p-9781119307655

Fort, K., (2017), Experts ou (foule de) non-experts ? la question de l'expertise des annotateurs vue de la myriadisation (crowdsourcing), *CORELA - COgnition, REprésentation, LAngage*, *HS-21*, https://doi.org/10.4000/corela.4835

Fort, K., (2022), Cours de Master 2.

Fort, K., (2012), *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus* (Thèse), Université Paris-Nord - Paris XIII, https://tel.archives-ouvertes.fr/tel-00797760

Fort, K. & Claveau, V., (2012a), Annotating Football Matches : Influence of the Source Medium on Manual Annotation, In *LREC - Eight International Conference on Language Resources and Evaluation*, Istanbul, Turquie, https://hal.archives-ouvertes.fr/hal-00709170

Fort, K., Ehrmann, M. & Nazarenko, A., (2009), Vers une méthodologie d'annotation des entités nommées en corpus ?, In *Traitement Automatique des Langues Naturelles 2009*, Senlis, France, https://hal.archives-ouvertes.fr/hal-00402321

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, Turquie, https://hal.archives-ouvertes.fr/hal-00709174

Fort, K., François, C. & Ghribi, M., (2010), Evaluer des annotations manuelles dispersées : les coefficients sont-ils suffisants pour estimer l'accord inter-annotateurs ?, In *Traitement Automatique des Langues Naturelles (TALN)*, Montréal, Canada, https://hal.archives-ouvertes.fr/hal-00484265

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, Pays-Bas, https://hal.inria.fr/hal-00969157

Fort, K., Nazarenko, A. & Rosset, S., (2012), Modeling the Complexity of Manual Annotation Tasks : a Grid of Analysis, In *International Conference on Computational Linguistics*, Mumbaï, Inde, https://hal.archives-ouvertes.fr/hal-00769631

Fort, K. & Claveau, V., (2012b), Annotation manuelle de matchs de foot : Oh la la ! l'accord inter-annotateurs ! et c'est le but ! (Manual Annotation of Football Matches : Inter-annotator Agreement ! Gooooal !) [in French], In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN*, Grenoble, France, ATALA/AFCP, https://aclanthology.org/F12-2031

Fort, K. & Névéol, A., (2018), Présence et représentation des femmes dans le traitement automatique des langues en France, In *Penser la Recherche en Informa-*

tique comme pouvant être Située, Multidisciplinaire Et Genrée (PRISME-G), Paris, France, https://hal.archives-ouvertes.fr/hal-01683774

Fort, K. & Sagot, B., (2010), Influence of Pre-Annotation on POS-Tagged Corpus Development, In *Proceedings of the Fourth Linguistic Annotation Workshop*, Uppsala, Suède, Association for Computational Linguistics, https://aclanthology.org/W10-1807

Francis, W. N., Kucera, H., Kučera, H. & Mackie, A. W., (1982), *Frequency analysis of English usage : Lexicon and grammar*, Houghton Mifflin.

Gabay, S., Clérice, T. & Reul, C., (2020), *OCR17 : Ground Truth and Models for 17th c. French Prints (and hopefully more)* [working paper or preprint], working paper or preprint, https://hal.archives-ouvertes.fr/hal-02577236

Galibert, O., Rosset, S., Grouin, C., Zweigenbaum, P. & Quintard, L., (2011), Structured and Extended Named Entity Evaluation in Automatic Speech Transcriptions, In *Proceedings of 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thaïlande, Asian Federation of Natural Language Processing, https://aclanthology.org/I11-1058

Galliano, S., Gravier, G. & Chaubard, L., (2009), The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts, In *Tenth Annual Conference of the International Speech Communication Association*, Brighton, Royaume-Uni.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H. & Crawford, K., (2021), Datasheets for Datasets, *Commun. ACM*, *6412*, 86-92, https://doi.org/10.1145/3458723

Geiger, D., Seedorf, S., Schulze, T., Nickerson, R. C. & Schader, M., (2011), Managing the crowd : towards a taxonomy of crowdsourcing processes, In *AMCIS 2011 Proceedings*, Detroit, michigan, États-Unis.

Gravier, G., Adda, G., Paulson, N., Carré, M., Giraudel, A. & Galibert, O., (2012), The ETAPE corpus for the evaluation of speech-based TV content processing in the French language, In *LREC-Eighth international conference on Language Resources and Evaluation*, Istanbul, Turquie.

Green, A. M., (1997), Kappa statistics for multiple raters using categorical classifications, In *Proceedings of the 22nd annual SAS User Group International conference*, San Diego, Californie.

Grishman, R. & Sundheim, B., (1996), Message Understanding Conference- 6 : A Brief History, In *COLING 1996 Volume 1 : The 16th International Conference on Computational Linguistics*, Copenhague, Danemark, https://aclanthology.org/C96-1079

Gross, M., (1982), Une classification des phrases « figées » du français, *Revue québécoise de linguistique*, *112*, 151-185, https://doi.org/https://doi.org/10.7202/602492ar

Grouin, C., (2013), *Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique* (Thèse), Université Pierre et Marie Curie - Paris VI, https://tel.archives-ouvertes.fr/tel-00848672

Grouin, C., (2018), Simplification de schémas d'annotation : un aller sans retour ? (Annotation scheme simplification : a one way trip with no return ?), In *Actes de la Conférence TALN. Volume 1 - Articles longs, articles courts de TALN*, Rennes, France, ATALA, https://aclanthology.org/2018.jeptalnrecital-court.32

Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O. & Quintard, L., (2011), Proposal for an Extension of Traditional Named Entitites : from Guidelines to Evaluation, an Overview, In *5th Linguistics Annotation Workshop (The LAW V)*, Portland, États-Unis, https://hal.archives-ouvertes.fr/hal-00604369

Guéron, J., (1979), Relations de coréférence dan la phrase et dans le discours, *Langue Française*, *44*, 42-79.

Guillaume, B., Fort, K. & Lefèbvre, N., (2016), Crowdsourcing Complex Language Resources : Playing to Annotate Dependency Syntax, In *International Conference on Computational Linguistics (COLING)*, Osaka, Japon, https://hal.inria.fr/hal-01378980

Gut, U. & Bayerl, P. S., (2004), Measuring the Reliability of Manual Annotations of Speech Corpora, In *Proceedings of the Speech Prosody*, Nara, Japon.

Habert, B., (2000a), *Corpus. Méthodologie et applications linguistiques* (T. 31), Presses Universitaires de Perpignan.

Habert, B., (2000b), Des corpus représentatifs : de quoi, pour quoi, comment, *Cahiers de l'Université de Perpignan*, *31*, 11-58.

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*, *101*, 25-45, https://journals.lib.unb.ca/index.php/CJAL/article/view/19733

Hamon, T., Fraisse, A., Paroubek, P., Zweigenbaum, P. & Grouin, C., (2015), Analyse des émotions, sentiments et opinions exprimés dans les tweets : présentation et résultats de l'édition 2015 du défi fouille de texte (DEFT), In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2015)*, Caen, France, https://hal.archives-ouvertes.fr/hal-01617180

Hebert, F., (1965), *Dune*, Chilton Books.

Hiebel, N., Ferret, O., Fort, K. & Névéol, A., (2022), CLISTER : A Corpus for Semantic Textual Similarity in French Clinical Narratives, In *Proceedings of the Language Resources and Evaluation Conference*, Marseille, France, European Language Resources Association, https://aclanthology.org/2022.lrec-1.459

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*, Marseille, France, European Language Resources Association, https://aclanthology.org/2020.lrec-1.533

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*, Paris, France, https://hal.archives-ouvertes.fr/hal-01378351

Ho-Dac, L.-M. & Poudat, C., (2021), Annoter, partager et comparer avec Glozz, Inception et TXM-URS, In *Plate-forme collaborative et participative de transcription, relecture et annotation de corpus (CORLI)*, Paris, France, https://hal.archives-ouvertes.fr/hal-03560281

Hovy, D. & Spruit, S. L., (2016), The Social Impact of Natural Language Processing, In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, Berlin, Allemagne, Association for Computational Linguistics, https://doi.org/10.18653/v1/P16-2096

Hripcsak, G. & Rothschild, A. S., (2005), Agreement, the F-Measure, and Reliability in Information Retrieval, *Journal of the American Medical Informatics Association*, *123*, 296-298, https://doi.org/10.1197/jamia.M1733

Ide, N. & Suderman, K., (2007), GrAF : A Graph-based Format for Linguistic Annotations, In *Proceedings of the Linguistic Annotation Workshop*, Prague, République tchèque, Association for Computational Linguistics, https://aclanthology.org/W07-1501

Inel, O., Khamkham, K., Cristea, T., Dumitrache, A., Rutjes, A., van der Ploeg, J., Romaszko, L., Aroyo, L. & Sips, R.-J., (2014), Crowdtruth : Machine-human

computation framework for harnessing disagreement in gathering annotated data, In *International semantic web conference*, Riva del Garda, Italie, Springer.

Jiang, M., Hu, Y., Worthey, G., Dubnicek, R. C., Capitanu, B., Kudeki, D. & Downie, J. S., (2021), The Gutenberg-HathiTrust Parallel Corpus : A Real-World Dataset for Noise Investigation in Uncorrected OCR Texts, *iSchools*, https://www.ideals.illinois.edu/handle/2142/109695

Kang, D., Ammar, W., Dalvi, B., van Zuylen, M., Kohlmeier, S., Hovy, E. & Schwartz, R., (2018), A Dataset of Peer Reviews (PeerRead) : Collection, Insights and NLP Applications, In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, Nouvelle Orléans, Louisiane, Association for Computational Linguistics, https://doi.org/10.18653/v1/N18-1149

Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E. & Gurevych, I., (2018), The INCEpTION Platform : Machine-Assisted and Knowledge-Oriented Interactive Annotation [Event Title : The 27th International Conference on Computational Linguistics (COLING 2018)], In *Proceedings of the 27th International Conference on Computational Linguistics : System Demonstrations*, Santa Fe, Nouveau-Mexique, États-Unis, Association for Computational Linguistics, Event Title : The 27th International Conference on Computational Linguistics (COLING 2018), http://tubiblio.ulb.tu-darmstadt.de/106270/

Krippendorff, K., (1995), On the Reliability of Unitizing Continuous Data, *Sociological Methodology*, *25*, 47-76.

Krippendorff, K., (2013), *Content analysis : An Introduction to its Methodology* (3e éd.), Sage publications.

Krippendorff, K., Mathet, Y., Bouvry, S. & Widlcher, A., (2016), On the reliability of unitizing textual continua : Further developments, *Quality & Quantity*, *506*, 2347-2364, https://doi.org/10.1007/s11135-015-0266-1

Kucera, H. & Francis, N. W., (1967), *Computational Analysis of Present-Day American English*, Providence, Rhode Island, États-Unis, Brown University Press.

Labadié, A., Enjalbert, P. & Ferrari, S., (2012), Transitions thématiques : Annotation d'un corpus journalistique et premières analyses, In *Joint Conference JEP-TALN-RECITAL 2012*, Grenoble, France, https://hal.archives-ouvertes.fr/hal-01071693

Labadié, A., Enjalbert, P., Mathet, Y. & Widlöcher, A., (2010), Discourse structure annotation : Creating reference corpora (Budin, Gerhard, Romary, Laurent, Declerck, Thierry, Wittenburg & Peter, Éd.), In Budin, Gerhard, Romary, Laurent, Declerck, Thierry, Wittenburg & Peter (Éd.), *Workshop on Language Resource and Language Technology Standards - state of the art, emerging needs, and future developments*, La Valette, Malte, https://hal.archives-ouvertes.fr/hal-01016656

Lafourcade, M. & Joubert, A., (2008), JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes, In *JADT'08 : Journées internationales d'Analyse statistiques des Données Textuelles*, France, https://hal-lirmm.ccsd.cnrs.fr/lirmm-00358848

Laignelet, M., (2009), *Analyse discursive pour le reperage automatique de segments obsolescents dans des documents encyclopédiques.* (Thèse), Université Toulouse le Mirail - Toulouse II, https://tel.archives-ouvertes.fr/tel-00461579

Landis, J. R. & Koch, G. G., (1977), The Measurement of Observer Agreement for Categorical Data, *Biometrics*, *331*, 159-174.

Landragin, F., (2021), Le corpus Democrat et son exploitation. Présentation, *Langages*, *224*, 11-24, https://doi.org/10.3917/lang.224.0011

Landragin, F., Potier, J. & Bothua, M., (2017), Annotation manuelle d'expressions référentielles : expérimentations pour simplifier les prises de décisions et optimiser le processus, In *9èmes Journées Internationales de la Linguistique de corpus (JLC 2017)*, Grenoble, France, https://halshs.archives-ouvertes.fr/halshs-01513810

Leech, G., (1993), Corpus annotation schemes, *Literary and linguistic computing*, *84*, 275-281.

Leech, G., (1997), Grammatical tagging, In R. G. Garside, G. Leech & A. M. Mcenery (Éd.), *Corpus Annotation : Linguistic Information from Computer Text Corpora* (p. 19-33), Bologne, Italie, Routledge.

Leech, G., (2005), *Developing Linguistic Corpora : a Guide to Good Practice*, Oxford, Royaume-Uni, Oxbow Books.

Lefeuvre, A., Antoine, J.-Y. & Schang, E., (2014), Le corpus ANCOR_Centre et son outil de requêtage : application à l'etude de l'accord en genre et nombre dans les coréférences et anaphores en français parlé, *SHS Web of Conferences*, *8*, 2691-2706, https://doi.org/10.1051/shsconf/20140801359

Lefeuvre, A., Antoine, J.-Y. & Allegre, W., (2015), Éthique conéquentialiste et traitement automatique des langues : une typologie de facteurs de risques adaptée aux technologies langagières, In *Atelier Ethique et TRaitemeNt Automatique des Langues (ETeRNAL'2015), conférence TALN'2015*, Caen, France, https://hal.archives-ouvertes.fr/hal-01170630

Lefeuvre, A., Antoine, J.-Y., Savary, A., Schang, E., Abouda, L., Maurel, D. & Eshkol, I., (2014), Annotation de la temporalité en corpus : contribution à l'amélioration de la norme TimeML (ATALA, Éd.), In ATALA (Éd.), *TALN'2014*, Marseille, France, https://hal.archives-ouvertes.fr/hal-01075207

Liberman, A., Fischer, J. & Whitney, D., (2014), Serial Dependence in the Perception of Faces, *Current biology : CB*, *2421*, 2569, https://doi.org/10.1016/j.cub.2014.09.025

Lion-Bouton, A., Grobol, L., Antoine, J.-Y., Billot, S. & Lefeuvre-Halftermeyer, A., (2020), Comment arpenter sans mètre : les scores de résolution de chaînes de coréférences sont-ils des métriques ? (Do the standard scores of evaluation of coreference resolution constitute metrics ?), In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). 2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)*, Nancy, France, ATALA et AFCP, https://aclanthology.org/2020.jeptalnrecital-eternal.2

Luo, X., (2005), On Coreference Resolution Performance Metrics, In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, Canada, Association for Computational Linguistics, https://www.aclweb.org/anthology/H05-1004

Makhoul, J., Kubala, F., Schwartz, R. & Weischedel, R., (1999), Performance Measures For Information Extraction, In *Proceedings of DARPA Broadcast News Workshop*, Herndon, Virginie, États-Unis.

Malaterra, G., (2016, cop. 2016), Histoire du grand comte Roger et de son frère Robert Guiscard . Vol. I Livres I & II, *Caen, France*, Centre Michel de Boüard-CRAHAM, Centre de recherches archéologiques et historiques anciennes et mediévales Presses universitaires de Caen.

Manning, C. D., (2011), Part-of-Speech Tagging from 97% to 100% : Is It Time for Some Linguistics ?, In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part I*, Tokyo, Japan, Springer-Verlag.

Marcus, M. P., Santorini, B. & Marcinkiewicz, M. A., (1993), Building a Large Annotated Corpus of English : The Penn Treebank, *Computational Linguistics*, *192*, 313-330, https://aclanthology.org/J93-2004

Martínez Alonso, H., Seddah, D. & Sagot, B., (2016), From Noisy Questions to Minecraft Texts : Annotation Challenges in Extreme Syntax Scenario, In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, Osaka, Japon, The COLING 2016 Organizing Committee, https://aclanthology.org/W16-3905

Maslow, A. H., (1966), The psychology of science a reconnaissance.

Mathet, Y., (2017), The Agreement Measure γ_{cat} a Complement to γ Focused on Categorization of a Continuum, *Computational Linguistics*, *433*, 661-681, https://doi.org/10.1162/COLI_a_00296

Mathet, Y. & Widlöcher, A., (2016), Évaluation des annotations : ses principes et ses pièges, *Revue TAL*, *572*, 73-98, https://hal.archives-ouvertes.fr/hal-01712282

Mathet, Y., Widlöcher, A., Fort, K., François, C., Galibert, O., Grouin, C., Kahn, J., Rosset, S. & Zweigenbaum, P., (2012), Manual Corpus Annotation : Giving Meaning to the Evaluation Metrics, In *International Conference on Computational Linguistics*, Mumbaï, Inde, https://hal.archives-ouvertes.fr/hal-00769639

Mathet, Y., Widlöcher, A. & Métivier, J.-P., (2015), The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment, *Computational Linguistics*, *413*, 437-479, https://doi.org/10.1162/COLI_a_00227

McClosky, D., Charniak, E. & Johnson, M., (2006), Reranking and Self-Training for Parser Adaptation, In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australie, Association for Computational Linguistics, https://doi.org/10.3115/1220175.1220218

McEnery, T. & Hardie, A., (2011), *Corpus linguistics : Method, theory and practice*, Cambridge University Press.

Metsis, V., Androutsopoulos, I. & Paliouras, G., (2006), Spam filtering with naive bayes-which naive bayes ?, In *CEAS*, Mountain View, Californie, États-unis.

Millour, A., (2020), *Myriadisation de ressources linguistiques pour le traitement automatique de langues non standardisées* (Thèse), Sorbonne Universite, https://hal.archives-ouvertes.fr/tel-03083213

Millour, A., Dupont, Y., Jouglar, A. & Fort, K., (2022), FENEC : un corpus équilibré pour l'évaluation des entités nommées en français (FENEC : a balanced sample corpus for French named entity recognition), In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, Avignon, France, ATALA, https://aclanthology.org/2022.jeptalnrecital-taln.8

Miltsakaki, E., Prasad, R., Joshi, A. & Webber, B., (2004), The Penn Discourse Treebank, In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbonne, Portugal, European Language Resources Association (ELRA), http://www.lrec-conf.org/proceedings/lrec2004/pdf/618.pdf

Moosavi, N. S. & Strube, M., (2016), Which Coreference Evaluation Metric Do You Trust ? A Proposal for a Link-based Entity Aware Metric, In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Berlin, Allemagne, Association for Computational Linguistics, https://doi.org/10.18653/v1/P16-1060

Mpouli Njanga Seh, S. P., (2016), *Automatic annotation of similes in literary texts* (Thèse 2016PA066298), Université Pierre et Marie Curie - Paris VI, https://tel.archives-ouvertes.fr/tel-01467081

Müller, C., (2006), Representing and Accessing Multi-Level Annotations in MMAX2, In *Proceedings of the 5th Workshop on NLP and XML (NLPXML-2006) : Multi-Dimensional Markup in Natural Language Processing*, Trente, Italie, https://aclanthology.org/W06-2712

Muzerelle, J., Lefeuvre, A., Schang, E., Antoine, J.-Y., Pelletier, A., Maurel, D., Eshkol, I. & Villaneau, J., (2014), ANCOR_Centre, a Large Free Spoken French Coreference Corpus : description of the Resource and Reliability Measures (ELRA, Éd.), In ELRA (Éd.), *LREC'2014, 9th Language Resources and Evaluation Conference.* Reyjavik, Islande, https://hal.archives-ouvertes.fr/hal-01075679

Nazarenko, A., Habert, B. & Salem, A., (1997), *Les linguistiques de corpus*, Armand Colin, https://hal.archives-ouvertes.fr/hal-00619268

Nédellec, C., Bessieres, P., Bossy, R., Kotoujansky, A. & Manine, A. P., (2006), Annotation guidelines for machine learning-based named entity recognition in mi-

crobiology, In *Proceeding of Data and Text Mining for Integrative Biology Workshop 17. European Conference on Machine Learning 10. European Conference on Principles and Practice of Knowledge Discovery in Databases*, Berlin, Allemagne, Springer-Verlag.

Nedoluzhko, A., Novák, M., Popel, M., Žabokrtský, Z., Zeldes, A. & Zeman, D., (2022), CorefUD 1.0 : Coreference Meets Universal Dependencies, In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France, European Language Resources Association, https://aclanthology.org/2022.lrec-1.520

Neuendorf, K. A., (2002), Defining content analysis, *Content analysis guidebook. Thousand Oaks, CA : Sage.*

Neuendorf, K. A., (2009), Reliability for content analysis, In A. B. Jordan, D. Kunkel, J. Manganello & M. Fisbhein (Éd.), *Media Messages and Public Health* (p. 67-87), Routledge.

Neves, M. & Ševa, J., (2019), An extensive review of tools for manual annotation of documents, *Briefings in Bioinformatics*, *221*, https ://academic.oup.com/bib/article-pdf/22/1/146/35934686/bbz130.pdf, 146-163, https : / / doi . org / 10 . 1093 / bib / bbz130

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R. & Zeman, D., (2016), Universal Dependencies v1 : A Multilingual Treebank Collection, In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovénie, European Language Resources Association (ELRA), https://aclanthology.org/L16-1262

Nouvel, D., Ehrmann, M. & Rosset, S., (2015), *Les entités nommées pour le traitement automatique des langues*, ISTE Editions, https://hal-inalco.archives-ouvertes.fr/hal-01359438

Ogrodniczuk, M., Katarzyna, G., Mateusz, K., Savary, A. & Magdalena, Z., (2014), *Coreference. Annotation, Resolution and Evaluation in Polish*, Walter de Gruyter, https://hal.archives-ouvertes.fr/hal-01174653

Ogrodniczuk, M. & Ng, V. (Éd.), (2016), *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, San Diego, Californie, États-Unis, Association for Computational Linguistics, https://doi.org/10.18653/v1/W16-07

Ogrodniczuk, M. & Ng, V. (Éd.), (2017), *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, Valence, Espagne, Association for Computational Linguistics, https://doi.org/10.18653/v1/W17-15

Orăsan, C., (2003), PALinkA : A highly customisable tool for discourse annotation, In *Proceedings of the Fourth SIGdial Workshop of Discourse and Dialogue*, Sapporo, Japon, https://aclanthology.org/W03-2120

Ortiz Suárez, P. J., Sagot, B. & Romary, L., (2019), Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures (P. Bański, A. Barbaresi, H. Biber, E. Breiteneder, S. Clematide, M. Kupietz, H. Lüngen & C. Iliadi, Éd.), In P. Bański, A. Barbaresi, H. Biber, E. Breiteneder, S. Clematide, M. Kupietz, H. Lüngen & C. Iliadi (Éd.), *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, Royaume-Uni, Leibniz-Institut für Deutsche Sprache, https://doi.org/10.14618/IDS-PUB-9021

Paroubek, P., Grouin, C., Bellot, P., Claveau, V., Eshkol-Taravella, I., Fraisse, A., Jackiewicz, A., Karoui, J., Monceaux, L. & Torres-Moreno, J.-M., (2018), DEFT2018 : Recherche d'information et analyse de sentiments dans des tweets concernant les transports en Île de France, In *DEFT 2018 - 14ème atelier Défi Fouille de Texte*, Rennes, France, https://hal.archives-ouvertes.fr/hal-01839407

Pasquer, C., (2017), Expressions polylexicales verbales : étude de la variabilité en corpus, In *TALN-RECITAL 2017*, Orléans, France, https://hal.archives-ouvertes.fr/hal-01637355

Passonneau, R. J., Bhardwaj, V., Salleb-Aouissi, A. & Ide, N., (2012), Multiplicity and word sense : evaluating and learning from multiply labeled word sense annotations, *Language Resources and Evaluation*, *462*, 219-252.

Paun, S., Artstein, R. & Poesio, M., (2022), Statistical Methods for Annotation Analysis, *Synthesis Lectures on Human Language Technologies*, *151*, https ://-doi.org/10.2200/S01131ED1V01Y202109HLT054, 1-217, https://doi.org/10.2200/S01131ED1V01Y202109HLT054

Pegors, T. T., Mattar, M. G., Bryan, P. B. & Epstein, R. A., (2015), Simultaneous perceptual and response biases on sequential face attractiveness judgments, *Journal of Experimental Psychology : General*, *1443*, 664-673, https://doi.org/10.1037/xge0000069

Perrotin, R., Nasr, A. & Auguste, J., (2018), Annotation en Actes de Dialogue pour les Conversations d'Assistance en Ligne, In *25e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Rennes, France, https://hal.archives-ouvertes.fr/hal-01943345

Péry-Woodley, M.-P., Afantenos, S., Ho-Dac, L.-M. & Asher, N., (2011), La ressource ANNODIS, un corpus enrichi d'annotations discursives, *Revue TAL*, *523*, 71-101, https://halshs.archives-ouvertes.fr/halshs-00935201

Pitrelli, J. F., Beckman, M. E. & Hirschberg, J., (1994), Evaluation of prosodic transcription labeling reliability in the tobi framework, In *Proc. 3rd International Conference on Spoken Language Processing (ICSLP 1994)*, Yokohama, Japon.

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.*, *31*, https://doi.org/10.1145/2448116.2448119

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R. & Xue, N., (2011), CoNLL-2011 Shared Task : Modeling Unrestricted Coreference in OntoNotes, In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning : Shared Task*, Portland, Oregon, États-Unis, Association for Computational Linguistics, https://aclanthology.org/W11-1901

Pustejovsky, J. & Stubbs, A., (2012), *Natural Language Annotation for Machine Learning*, Sebastopol, Californie, aux États-Unis, O'Reilly Media, Inc., https://www.oreilly.com/library/view/natural-language-annotation/9781449332693

Recasens, M. & Hovy, E., (2011), BLANC : Implementing the Rand index for coreference evaluation, *Natural Language Engineering*, *1704*, 485-510, https://doi.org/10.1017/S135132491000029X

RO, G. & Ledegen, G., (2012), Orthographe : ce qui est jugé difficile, *Glottopol*, *19*, 17-36, https://hal.archives-ouvertes.fr/hal-01114713

Rosset, S., Grouin, C., Fort, K., Galibert, O., Kahn, J. & Zweigenbaum, P., (2012), Structured named entities in two distinct press corpora : Contemporary broadcast news and old newspapers, In *6th Linguistics Annotation Workshop (The LAW VI)*, Jeju, Corée du Sud.

Roubaud, M.-N., (2014), *De la description de la langue à son enseignement* (Habilitation à diriger des recherches), UNIVERSITÉ STENDHAL – GRENOBLE 3, https://halshs.archives-ouvertes.fr/tel-01102494

Ruokolainen, T., Kauppinen, P., Silfverberg, M. & Lindn, K., (2020), A Finnish news corpus for named entity recognition, *Language Resources and Evaluation*, *541*, 247-272, https://doi.org/10.1007/s10579-019-09471-7

Sagot, B., Fort, K., Adda, G., Mariani, J. & Lang, B., (2011), Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé, In *TALN'2011 - Traitement Automatique des Langues Naturelles*, Montpellier, France, https://hal.inria.fr/inria-00617067

Sampson, G., (1995), *English for the Computer*, Oxford, Royaume-Uni, Oxford University Press, https://global.oup.com/academic/product/english-for-the-computer-9780198240235

Sand, G., (1881), *Pauline*, Revue des Deux Mondes.

Sankaran, B., Bali, K., Choudhury, M., Bhattacharya, T., Bhattacharyya, P., Jha, G. N., Rajendran, S., Saravanan, K., Sobha, L. & Subbarao, K., (2008), A Common Parts-of-Speech Tagset Framework for Indian Languages, In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Maroc, European Language Resources Association (ELRA), http://www.lrec-conf.org/proceedings/lrec2008/pdf/337_paper.pdf

Scott, W. A., (1955), Reliability of content analysis : The case of nominal scale coding, *Public opinion quarterly*, 321-325.

Searle, J., (1972), *Les actes de langage : essai de philosophie du langage*, Hermann.

Searle, J., (1982), *Sens et expression : Études de théorie des actes de langage*, Les Éditions de Minuit.

Seddah, D., Essaidi, F., Fethi, A., Futeral, M., Muller, B., Ortiz Suárez, P. J., Sagot, B. & Srivastava, A., (2020), Building a User-Generated Content North-African Arabizi Treebank : Tackling Hell, In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, Association for Computational Linguistics, https://doi.org/10.18653/v1/2020.acl-main.107

Segura-Bedmar, I., Martínez, P. & Herrero-Zazo, M., (2013), SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013), In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2 : Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, Georgie, États-Unis, Association for Computational Linguistics, https://aclanthology.org/S13-2056

Sharoff, S., (2006), Creating general-purpose corpora using automated search engine queries, In M. Baroni & S. Bernardini (Éd.), *WaCky ! Working papers on the Web as Corpus* (p. 63-98), Bologne, Italie, GET.

Sinclair, J., (2005), Corpus and Text — Basic Principles, In M. Wynne (Éd.), *Developing Linguistic Corpora : a Guide to Good Practice* (p. 1-16), Owbow Books.

Snow, R., O'Connor, B., Jurafsky, D. & Ng, A., (2008), Cheap and Fast – But is it Good ? Evaluating Non-Expert Annotations for Natural Language Tasks, In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaï, Association for Computational Linguistics, https://aclanthology.org/D08-1027

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S. & Tsujii, J., (2012), brat : a Web-based Tool for NLP-Assisted Text Annotation, In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, Association for Computational Linguistics, https://aclanthology.org/E12-2021

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C. & Meteer, M., (2000), Dialogue act modeling for automatic tagging and recognition of conversational speech, *Computational Linguistics*, *263*, 339-374, https://aclanthology.org/J00-3003

Stubbs, A., (2012), Developing specifications for light annotation tasks in the biomedical domain, In *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining*, Istanbul, Turquie.

Syrdal, A. K. & McGory, J., (2000), Inter-transcriber reliability of ToBI prosodic labeling, In *Sixth International Conference on Spoken Language Processing*, Pékin, Chine.

Taylor, M., Ann and Marcus & Santorini, B., (2003), The Penn Treebank : An Overview, In A. Abeillé (Éd.), *Treebanks : Building and Using Parsed Corpora* (p. 5-22), Dordrecht, Pays-Bas, Springer Netherlands, https://doi.org/10.1007/978-94-010-0201-1_1

Teufel, S. Et al., (1999), *Argumentative zoning : Information extraction from scientific text* (Thèse), University of Edinburgh, Édimbourg, Écosse.

Tjong Kim Sang, E. F. & De Meulder, F., (2003), Introduction to the CoNLL-2003 Shared Task : Language-Independent Named Entity Recognition, In *Proceedings*

of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, https://aclanthology.org/W03-0419

Todorov, T., (1970), *Introduction à la littérature fantastique*, Paris, France, Seuil.

Valette, M., (2016), Analyse statistique des données textuelles et traitement automatique des langues. Une étude comparée (D. Mayaffre, C. Poudat, L. Vanni, V. Magri & P. Follette, Éd.), In D. Mayaffre, C. Poudat, L. Vanni, V. Magri & P. Follette (Éd.), *International Conference on Statistical Analysis of Textual Data (JADT2016)*, Nice, France, https://hal-inalco.archives-ouvertes.fr/hal-01335084

Vestlund, J., Langeborg, L., Srqvist, P. & Eriksson, M., (2009), Experts on age estimation, *Scandinavian Journal of Psychology*, *504*, 301-307, https://doi.org/10.1111/j.1467-9450.2009.00726.x

Vilain, M., Burger, J. D., Aberdeen, J., Connolly, D. & Hirschman, L., (1995), A Model-Theoretic Coreference Scoring Scheme, *ACL Anthology*, https://www.aclweb.org/anthology/M95-1005

Voelkle, M. C., Ebner, N. C., Lindenberger, U. & Riediger, M., (2012), Let me guess how old you are : effects of age, gender, and facial expression on perceptions of age, *Psychology and Aging*, *272*, 21895379, 265-277, https://doi.org/10.1037/a0025065

von Ahn, L., (2006), Games with a purpose, *Computer*, *396*, 92-94, https://doi.org/10.1109/MC.2006.196

Voorhees, E. M. & Harman, D., (1998), The Text REtrieval Conferences (TRECs), In *TIPSTER TEXT PROGRAM PHASE III : Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998*, Baltimore, Maryland, États-Unis, Association for Computational Linguistics, https://doi.org/10.3115/1119089.1119127

Voormann, H. & Gut, U., (2008), Agile corpus creation, *Corpus Linguistics and Linguistic Theory*, *42*, 235-251, https://doi.org/doi:10.1515/CLLT.2008.010

Watson, T. L., Otsuka, Y. & Clifford, C. W. G., (2016), Who are you expecting ? Biases in face perception reveal prior expectations for sex and age, *Journal of Vision*, *163*, 5, https://doi.org/10.1167/16.3.5

Wayne, C. L., (2000), Multilingual Topic Detection and Tracking : Successful Research Enabled by Corpora and Evaluation, In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athènes, Grèce,

European Language Resources Association (ELRA), http://www.lrec-conf.org/proceedings/lrec2000/pdf/168.pdf

Widlöcher, A. & Mathet, Y., (2009), La plate-forme Glozz : environnement d'annotation et d'exploration de corpus, In *Actes de la 16e Conférence Traitement Automatique des Langues Naturelles (TALN'09), session posters*, Senlis, France, https://hal.archives-ouvertes.fr/hal-01011969

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B., (2016), The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data*, *3160018*, 1-9, https://doi.org/10.1038/sdata.2016.18

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*, Montréal, Canada, ATALA, https://aclanthology.org/2010.jeptalnrecital-long.13

Woliński, M., Głowińska, K. & Świdziński, M., (2011), A preliminary version of Składnica — a treebank of Polish, In *Proceedings of the 5th Language and Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznań, Pologne.

Yimam, S. M., Gurevych, I., Eckart de Castilho, R. & Biemann, C., (2013), WebAnno : A Flexible, Web-based and Visually Supported System for Distributed Annotations, In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, Sofia, Bulgarie, Association for Computational Linguistics, https://aclanthology.org/P13-4001

Yvon, F. & Segal, N., (2012), *Des corpus d'erreurs pour TRACE* (rapp. tech.), LIMSI.

Zhou, D., Basu, S., Mao, Y. & Platt, J., (2012), Learning from the Wisdom of Crowds by Minimax Entropy (F. Pereira, C. Burges, L. Bottou & K. Weinberger, Éd.), In F. Pereira, C. Burges, L. Bottou & K. Weinberger (Éd.), *Advances in Neural Information Processing Systems*, Curran Associates, Inc., https://proceedings.neurips.cc/paper/2012/file/46489c17893dfdcf028883202cefd6d1-Paper.pdf

Annexes

Les annexes qui suivent sont composées de :

- Campagne d'annotations Portraits :
  - texte d'appel à participation aux étudiants de la licence Humanités ;
  - corpus « Portraits » ;
- Campagne d'annotations Fautes :
  - texte d'appel à participation ;
  - guide d'explication/annotation fourni aux annotateurs ;
  - corpus « Erreurs » ;
  - document explicatif des réponses et des règles abordées durant l'expérience ;

Annexe A

Campagne d'annotations « Portraits »

A.1 Texte d'appel à participation transmis aux étudiants de la licence Humanité

Le message ci-dessus a été transmis via la fonction « Forum » du Moodle associé au cours, par un des responsables du diplôme.

> Bonjour à toutes et à tous,
>
> La semaine prochaine, la conférence de culture numérique de Anaëlle Baledent et Yann Mathet sera consacrée à l'annotation de corpus et la mesure de l'accord inter-annotateurs pour la constitution de données de référence.
>
> Afin d'illustrer concrètement certains points importants de cette présentation, un exercice d'annotation vous est soumis aujourd'hui, dont les résultats seront exploités pendant la conférence.
>
> Cet exercice, dont vous ne comprendrez probablement pas l'intérêt avant la semaine prochaine (cela est parfaitement normal), consistera, pour chacun d'entre vous, à répondre à une série de 100 questions. Pour chaque question, la photographie d'un individu humain vous sera présentée et vous devrez en déterminer l'âge aussi précisément que possible, âge que vous indiquerez simplement, dans la zone de saisie prévue pour cela, en saisissant un simple nombre entier correspondant à votre estimation.
>
> Pour ce travail, nous vous remercions de respecter scrupuleusement les instructions suivantes :
>
> 1. Bien entendu, votre performance à cet exercice ne sera pas prise en compte dans votre évaluation universitaire. Nous vous prions toutefois de jouer le jeu aussi sérieusement que possible.
> 2. L'accès à l'exercice est disponible tout en bas de la page consacrée aux conférences de culture numérique (https://ecampus.unicaen.fr/course/view.php?id=18563). Chacun d'entre vous peut accéder à un exercice nommé « Scénario X » (où X est un nombre compris entre 1 et 7).

3. À partir du moment où vous démarrez l'exercice, vous disposez d'une heure pour répondre à l'ensemble des questions. Cela est largement suffisant.

4. Pour répondre à chaque question, vous ne devez vous appuyer que sur les informations visuelles que vous donne la photographie elle-même (sans chercher notamment à identifier la personne concernée et/ou à retrouver la photographie sur le web en quête d'indications complémentaires).

5. Nous vous demandons par ailleurs de traiter les questions les unes après les autres, dans l'ordre dans lequel elles vous sont présentées, sans revenir en arrière.

6. Une fois l'ensemble de questions traité, pensez bien à « Terminer le test » en utilisant le bouton prévu pour cela.

L'exercice est disponible à partir d'aujourd'hui (jeudi 26/03). Nous vous demandons de bien vouloir l'effectuer avant vendredi 27/03 à 12h.

Bien cordialement, Anaëlle Baledent, Yann Mathet et Antoine Widlöcher.

A.2 Corpus Portraits

Code	Énoncé	Âge réel
I001	1280px-Crown_Princess_Victoria_2016_10.jpg	0.24
I002	Naruhito19610204.jpg	0.95
I003	Princess_Alexia_Juliana_Marcela_Laurentien_%282006%29.jpg	1.28
I004	800px-Estelle_of_Sweden.jpg	1.29
I005	Prince_Oscar%2C_Duke_of_Sk%C3%A5ne_in_2018_%28cropped%29.jpg	2.26
I006	Princess_Leonore_May_2016_%28cropped%29.jpg	2.26
I007	Catharina-Amalia_Beatrix_Carmen_Victoria_%282006%29.jpg	2.84
I008	1024px-Princess_Estelle_of_Sweden.jpg	4.26
I009	Princess_Ariane.jpg	6.06
I010	800px-Prince_Hisahito_of_Akishino.jpg	7.15

Code	Énoncé	Âge réel
I011	Nationaldagen_EM1B2126_%2848018060506%29.jpg	7.28
I012	800px-Prinses-ariane.jpg	7.65
I013	336px-Catharina-Amalia_Beatrix_Carmen_Victoria_%282013%29.jpg	9.4
I014	800px-Prinses_Alexia%2C_Wassenaar%2C_najaar_2014.jpg	9.44
I015	800px-Malia_Obama.jpg	10.55
I016	800px-Wassenaar%2C_najaar_2014%2C_de_Prinses_van_Oranje.jpg	10.99
I017	Catharina-Amalia_Beatrix_Carmen_Victoria_%282015%29.jpg	11.64
I018	Koningsdag_2019_Amersfoort_15.jpg	12.05
I019	800px-Dafne_Keen_Press_Conference_Logan_Berlinale_2017_02.jpg	12.12
I020	800px-Sverre_Magnus_de_Norv%C3%A8ge.png	12.45
I021	800px-Prinses_Elisabeth_van_Belgi%C3%AB.jpg	12.5
I022	800px-Tom_Holland_Billy_Elliott_2010_1b.jpg	13.83
I023	Princess_Alexia_Juliana_Marcela_Laurentien_%282019%29.jpg	13.83
I024	800px-Ingrid_Alexandra_de_Norv%C3%A8ge.png	14.32
I025	800px-Millie_Bobby_Brown_2018.jpg	14.42
I026	800px-Asa_Butterfield_in_2011_%28cropped%29.jpg	14.64
I027	Catharina-Amalia_Beatrix_Carmen_Victoria_%282019%29.jpg	15.39
I028	Andrea_bowen_headshot.jpg	15.92
I029	600px-Malala_Yousafzai_2015.jpg	18.22
I030	800px-Prins_Willem-Alexander_als_LTZ3.jpg	18.68
I031	Amandla_Stenberg_2018.jpg	19.81
I032	800px-Aml-Ameen-200%C2%A7-10a.jpg	21.19
I033	Princess_Mako_and_Princess_Kako_at_the_Tokyo_Imperial_Palace_%28cropped%29.jpg	23.2
I034	800px-Chris_Colfer_2013.jpg	23.21

Code	Énoncé	Âge réel
I035	800px-Princess_Noriko_cropped_2_Princess_Noriko_2013.JPG	24.45
I036	800px-Demi_Lovato_in_2017_cropped_02.JPG	25.07
I037	Tena_desae.jpg	25.23
I038	800px-Samantha_Bailly_-_Utopiales_2014_-_P960148.jpg	25.95
I039	618px-Prince_William_at_seedhill_mills.jpg	27.45
I040	800px-Audrey_-Alwett-2.jpg	28.0
I041	800px-Rencontre_Abonn%C3%A9s_2016_Montbazon_-_001.jpg	28.1
I042	800px-Patricia_Lyfoung_20070511_Fnac.jpg	29.39
I043	800px-2018.06.24._Clementine_Beauvais_Fot_Mariusz_Kubik.JPG	29.46
I044	800px-Marie_Lu.JPG	29.97
I045	800px-20190831RS0119_%2848684497503%29.jpg	32.15
I046	800px-Marjorie_Liu%2C_2012_%28cropped%29.jpg	33.08
I047	Emmanuel_Moire_NRJ_Music_Awards_2013.jpg	33.62
I048	800px-Naomi_Novik_July08.jpg	35.21
I049	800px-Mae_Carol_Jemison_%28cropped_2%29.jpg	35.7
I050	800px-Prince_Carl_Philip_of_Sweden_8255.jpg	36.07
I051	800px-Doona_Bae_promoting_The_Tunnel.png	36.75
I052	Rogue_One-_A_Star_Wars_Story_Japan_Premiere_Red_Carpet-_Diego_Luna_%2834959299874%29.jpg	36.94
I053	1280px-Timothee_de_Fombelle_20100329_Salon_du_livre_de_Paris_2.jpg	36.95
I054	800px-Crown_Prince_Haakon_of_Norway_2012-03-26_001.jpg	38.68
I055	800px-Victoria%2C_Crown_Princess_of_Sweden_in_2018.jpg	41.25

Code	Énoncé	Âge réel
I056	Diane_Kruger_C%C3%A9sar_2018_%28cropped%29.jpg	41.51
I057	N._K._Jemisin_%28cropped%29.jpg	42.72
I058	800px-Nnedi_Okorafor_%2837108184821%29.jpg	43.36
I059	330px-Pierre_Bottero_20080315_Salon_du_livre_1.jpg	44.08
I060	King_Mswati_III_2014.jpg	46.29
I061	Indvielse_af_Ninjago_World_-_Prins_Joachim.jpg	46.76
I062	800px-Katherine_Johnson_at_NASA%2C_in_1966_-_Original.jpg	47.35
I063	800px-L%E2%80%99infante_H%C3%A9l%C3%A8ne_d%E2%80%99Espagne%2C_duchesse_de_Lugo_en_2011.jpg	47.43
I064	%28Felipe_de_Borb%C3%B3n%29_Inauguraci%C3%B3n_de_FITUR_2018_%2839840659951%29_%28cropped%29.jpg	49.98
I065	Defense.gov_News_Photo_030612-D-2987S-002_%28cropped%29.jpg	50.87
I066	1280px-2019-05-30_Felipe_VI_of_Spain-6125.jpg	51.33
I067	800px-Italiaanse_schrijver_Umberto_Eco_%2C_kop%2C_Bestanddeelnr_932-9758.jpg	52.38
I068	1280px-Hirsch_morris.jpg	52.51
I069	1024px-Mae_Jemison_crop_2009_CHAO.jpg	53.07
I070	800px-Henri_of_Luxembourg_%282009%29.jpg	54.04
I071	Empress_Masako_at_TICAD7.jpg	55.72
I072	Crown_Prince_Naruhito_%282018%29.jpg	58.07
I073	800px-Jean_dAillon_20090315_Salon_du_livre_1.jpg	60.91
I074	800px-Bryan_Cranston_at_the_2018_Berlin_Film_Festival_%282%29.jpg	61.94
I075	800px-Katherine_Johnson_1983.jpg	64.35
I076	800px-King_Rama_X_official_%28crop%29.png	64.43
I077	Beatriz_dos_Pa%C3%ADses_Baixos.jpg	65.14

Code	Énoncé	Âge réel
I078	800px-Hamad_bin_Isa_Al_Khalifa_April_2016.jpg	66.19
I079	1280px-Koningin_Juliana%2C_Bestanddeelnr_254-9845.jpg	67.0
I080	800px-Carl_XVI_of_Sweden.jpg	68.59
I081	800px-Portrait_photoshoot_at_Worldcon_75%2C_Helsinki%2C_before_the_Hugo_Awards_%E2%80%93_George_R._R._Martin_%28cropped%29.jpg	68.89
I082	HM_Margrethe_II_2010_%28cropped%29.jpg	70.66
I083	800px-Umberto_Eco_04.jpg	73.34
I084	Drottning_Margrethe_av_Danmark_crop.jpg	74.11
I085	Prince_Masahito_cropped_2_Prince_Masahito_Prince_Albert_II_Princess_Hanako_and_Yukiya_Amano_20100713.jpg	74.54
I086	UrsulaLeGuin.01.jpg	74.73
I087	800px-Donald_Sutherland_%28cropped%29.JPG	77.9
I088	Empress_Michiko_cropped_20140424.jpg	79.51
I089	800px-Sheikh_Sabah_IV.jpg	80.13
I090	800px-Simone_Veil%2C_gymnase_Japy_2008_02_27_n5.jpg	80.63
I091	800px-Leiji_Matsumoto_-_Lucca_Comics_%26_Games_2018_02.jpg	80.77
I092	1280px-Robert_Silverberg_-_Samedi_-_Utopiales_2015_-_E96A1660.jpg	80.79
I093	800px-Harald_V_en_2018.jpg	81.23
I094	Emperor_Akihito_%282016%29.jpg	82.09
I095	Rosa_Parks_1997.jpg	83.61
I096	Maurice_Druon_2003_Orenburg_crop.jpg	84.69
I097	800px-Queen_Elizabeth_II_of_New_Zealand_%28cropped%29.jpg	84.7
I098	Katherine_Johnson_in_2008.jpg	90.0
I099	Elizabeth_II_2018_birthday_%28cropped%29.jpg	92.0
I100	800px-Katherine_Johnson_medal_%28cropped%29.jpeg	97.25

Annexe B

Campagne d'annotations « Fautes »

B.1 Texte d'appel à participation

Les cohortes étaient contactées par mail par ce texte :

> Dans le cadre de ma thèse, je souhaiterais recueillir des participations à un questionnaire disponible au lien suivant : https://baleden191.users.greyc.fr/campagnes/fautes.php . Généralement, l'expérience prend environ 30 minutes.
>
> Ce questionnaire contient une série de 100 questions ; pour chaque question, un énoncé (une phrase) vous sera présenté et vous devez cocher une seule des deux réponses disponibles : *Faute* ou *Sans faute*, selon si vous jugez que l'énoncé contient ou non une faute. Une phrase ne peut contenir qu'une seule faute au maximum.
>
> Jouez le jeu au maximum sans aller chercher la réponse sur internet. Pensez à bien envoyer vos réponses lorsque toutes les questions seront complétées. Une fois les réponses envoyées, vous pourrez obtenir votre score ainsi qu'un fichier expliquant certaines règles abordées durant l'expérience.
>
> Comme dit précédemment, l'expérience prend environ 30 minutes. Aucune information personnelle, mis à part votre niveau d'étude, ne vous sera demandée et les réponses seront anonymisées.
>
> Je vous remercie par avance de votre participation !

B.2 Guide d'annotation

Le guide d'annotation était commun à toutes les cohortes et à tous les scénarios. Il était affiché sur la première page du questionnaire :

> Ce questionnaire contient une série de 100 questions ; pour chaque question, un énoncé (une phrase) vous sera présenté et vous devrez cocher une seule des

deux réponses disponibles : *Faute* ou *Sans faute*, selon que vous jugez que l'énoncé contient ou non une faute. Une phrase peut contenir au plus une faute.

Dans cette expérience, nous employons le terme *faute* pour désigner une graphie ou une formulation qui diffère de ce qui est prescrit par l'Académie Française et les manuels de français. Cela comprend :

— les fautes d'orthographe ou lexicales : oubli ou rajout de doubles consonnes, de lettres finales ou d'accents, une confusion entre deux sons proches ou deux graphies, etc.
  — *proffesseur* au lieu de *professeur*
  — *dicernement* au lieu de *discernement*
  — *toujour* au lieu de *toujours*
— les fautes de grammaire : accord verbal ou au sein du groupe nominal, conjugaison (mauvais accord ou concordance des temps), homophonie entre deux mots appartenant à des catégories distinctes, etc.
  — *La jeune fille est tombé* au lieu de *La jeune fille est tombée*
  — *J'ai assister au cours* au lieu de *J'ai assisté au cours*
  — *S'est à toi de voir* au lieu de *C'est à toi de voir*,
— les fautes syntaxiques : inversion de l'ordre des mots, mauvaise préposition employée, etc.
  — *Tu te rappelles du film d'hier ?* au lieu de *Tu te rappelles le film d'hier ?*,
  — *Je t'ai pas entendu arriver* au lieu de *Je ne t'ai pas entendu arriver*,
  — *Le chat que je parle est roux* au lieu de *Le chat dont je parle est roux*,

Jouez le jeu au maximum sans aller chercher la réponse sur internet. Si les énoncés apparaissent tous sur la même page, vous pouvez modifier vos réponses tant que vous ne validez pas vos réponses. Ainsi, pensez à bien envoyer vos réponses lorsque toutes les questions seront complétées. Une fois les réponses envoyées, vous pourrez obtenir votre score ainsi qu'un fichier expliquant certaines règles abordées durant l'expérience.

B.3 Corpus Erreur

Code	Énoncé
SF001	Les devoirs que l'étudiant a rendus sont archivés pendant plusieurs années.
SF002	Il a rencontré ses camarades et leur a donné le document.
SF003	Les étudiants sont curieux ; il faut leur donner matière à réflexion.

Code	Énoncé
SF004	Le dysfonctionnement observé est probablement lié à une incompréhension des instructions.
SF005A	Cette semaine, je ne pourrai malheureusement pas assister à la réunion habituelle du mardi.
SF006	Nous pourrons discuter de ce point quand la suite aura été traitée.
SF007	Je vous renvoie à la discussion de la semaine dernière pour éclairer ce point.
SF008B	Les poésies qu'il a entendu chanter en Grèce lui ont donné le sens de la prosodie.
SF009C	Les poètes qu'il a entendus chanter en Grèce lui ont donné le sens de la prosodie.
SF010	L'adresse de l'éditeur doit impérativement figurer dans chaque notice bibliographique.
SF011D	Vous devez indiquer les phrases comportant des erreurs.
SF012E	Le comportement de cet individu laisse penser qu'il ignore le règlement intérieur.
SF013	Sans préjuger de la décision qui sera prise, nous pouvons espérer avoir éclairé le débat.
SF014	Ils ont formulé de nombreuses propositions, desquelles on pourrait déduire que le sujet les intéresse.
SF015F	Cette interprétation du texte mériterait d'être davantage diffusée.
SF016	C'est l'été qu'Isabelle est le mieux préparée car elle a plus de temps que l'hiver.
SF017G	Je vous saurais gré de bien vouloir me retourner le formulaire dans les plus brefs délais.
SF018H	Nous sommes convenus de nous voir prochainement.
SF019I	Tu es sûr que c'est bien de Pierre que tu parles ?
SF020J	C'est un film que je me rappelle très précisément.
SF021	Je commencerai le tri après qu'il aura fini de réunir tous les livres.
SF022	L'épidémie dégrade les conditions d'enseignement, voire décourage complètement certains étudiants.

Code	Énoncé
SF023	Le soi-disant policier n'était en fait qu'un escroc.
SF024	Ce manteau coûte quatre cent cinquante euros.
SF025K	Quelles que soient les conditions climatiques, ces pneus tiennent bien la route.
SF026	C'est l'une des raisons pour lesquelles cette décision me semble absurde.
SF027	Je suis désolée, une panne de voiture m'a empêchée de me rendre à mon rendez-vous.
SF028	Sophie, il l'a envoyée sur les roses !
SF029	Il m'a suggéré de venir un peu en avance.
SF030	Il y a quelque 2000 personnes qui se sont rendues à la manifestation
SF031	Le maire a été pris à partie par les manifestants.
SF032	Il a tiré parti de l'expérience de ses collègues.
SF033	L'étudiante attend que le chat errant mange.
SF034	Il entend ces phrases, que l'on prononce et oublie si vite.
SF035	Elle gardait une place dans ses pensées quoiqu'il advînt.
SF036	Les peines qu'elle avait connues s'envolèrent.
SF037	Les cours ont repris depuis bientôt un mois.
SF038	Quand j'ai eu terminé mon travail, j'ai mangé.
SF039	La factrice distribue le courrier à huit heures.
SF040	Le beau temps a permis de faire une balade à vélo.
SF041	Le jeune homme regarda les éclairs déchirer le ciel nocturne.
SF042	Leur moment était celui de ce battement de cœur raté.
SF043L	Le festival est censé se dérouler au printemps
SF044	L'adolescente porte une robe à carreaux rouge et blanc.
SF045M	Les deux jumeaux se ressemblent à s'y méprendre.
SF046	Si j'avais assez de temps, j'aimerais lire toutes les œuvres d'Alexandre Dumas.
SF047	Combien de cœurs a-t-elle brisés ?
SF048	Au vu de leurs notes, elles ont l'air sérieux comme candidates à une bourse de thèse.

Code	Énoncé
SF049	Pour pallier le manque d'enseignants, l'Éducation supérieure fait souvent appel à des vacataires.
SF050	Elle se fait fort de ne pas craquer devant ses harceleurs.
AF001	Les options que l'étudiant a retenu ne lui permettent pas de se réorienter dans la filière espérée.
AF002	Espéré depuis de longues années, cette réforme est une déception aux yeux de la communauté.
AF003	Précédamment, il avait déjà exprimé cette opinion.
AF004	Le recours à des procédés réthoriques ne dispense pas du respect de la logique.
AF005	La réduction de la place du latin dans l'enseignement secondaire cause du tord à la pratique universitaire des Humanités.
AF006	Cette proposition n'est ni raisonable, ni rationnelle.
AF007A	Cette semaine, je ne pourrais malheureusement pas assister à la réunion habituelle du mardi.
AF008B	Les poésies qu'il a entendues chanter en Grèce lui ont donné le sens de la prosodie.
AF009C	Les poètes qu'il a entendu‘ chanter en Grèce lui ont donné le sens de la prosodie.
AF010D	Vous devez indiquer les phrases comportants des erreurs.
AF011E	Le comportement de cet individu laisse pensé qu'il ignore le règlement intérieur.
AF012	Ces règles, des quelles on peut déduire toutes les autres, seront dites « axiomes ».
AF013	Les données que je me suis servi pour ce travail ont été publiées en 2020.
AF014F	Cette interprétation du texte mériterai d'être davantage diffusée.
AF015	Ce livre est un espèce de mélange entre la fantasy et la science-fiction.
AF016	Parmis ces figures de style, lesquelles vous semblent être des métonymies ?
AF017	Cette situation est moins pire que l'autre.

Code	Énoncé
AF018G	Je vous serais gré de bien vouloir me retourner le formulaire dans les plus brefs délais.
AF019H	Nous avons convenu d'un rendez-vous prochain.
AF020	Il y a d'avantage de personnes touchées par l'épidémie parmi les personnes âgées.
AF021I	Tu es sûr que c'est bien de Pierre dont tu parles ?
AF022J	C'est un film dont je me rappelle très précisément.
AF023	Te souviens-tu de la question qu'a posé l'inspecteur ?
AF024	C'est comme même pas courant qu'il neige en mai.
AF025	Quand à vous, je vous donnerai les consignes prochainement.
AF026	Nous constatons une flambée de violance dans les établissements scolaires.
AF027	Cette voiture coûte chère.
AF028K	Quelque soient les conditions climatiques, ces pneus tiennent bien la route.
AF029	C'est entre autre pour cette raison qu'elle s'est inscrite en retard.
AF030	La grève des transports a empêchée Lucie d'arriver à l'heure à son cours.
AF031	Il faut finaliser le travail commencer en classe sur les interprétations.
AF032	Aujourd'hui dans cette vidéo, je vous partage mon expérience pour réussir vos cheese-cakes.
AF033	La brise agitait légèremment le feuillage des arbres.
AF034	Les enfants regardent dans le puit.
AF035	L'acceuil se situe au rez-de-chaussée du bâtiment.
AF036L	Le festival est sensé se dérouler au printemps.
AF037	Le combattant dépose enfin les armes, triomfant.
AF038	As-tu appellé ?
AF039	Malgrés le beau temps, la course n'a pas pu avoir lieu.
AF040	J'ai beau laver le tapis, la tâche ne veut pas partir.
AF041	La jeune fille aide à ranger les affaires tombé.
AF042	Seul le bruit de ses pas rythmaient son trajet et ses pensées.

Code	Énoncé
AF043	Il a du revoir son organisation pour finir à temps son travail.
AF044	Des pâtes, j'en ai faites plein durant mes études.
AF045M	Les deux jumeaux se ressemblent à si méprendre.
AF046	Vôtre chat est obèse.
AF047	La plupart de ces phrases est fausse.
AF048	Je ne veux plus te voir, va-t-en !
AF049	Il joue avec moi aux jeux vidéo après qu'il ait dormi.
AF050	Les cahiers oranges appartiennent à mon frère.

B.4 Document explicatif des réponses et des règles

Campagne « Fautes » : réponses et explications

Anaëlle Baledent, Yann Mathet, Antoine Widlöcher

Afin de ne pas perturber l'expérience, nous vous prions de ne pas diffuser ce document, a fortiori à d'autres participant(e)s n'ayant pas encore participé.

Dans ce document, vous trouverez les réponses attendues à l'expérience ainsi que des explications aux règles de français associées. À des fins expérimentales, le questionnaire ne vous a pas été forcément présenté dans l'ordre originel. Il se peut donc que les réponses ci-dessous n'apparaissent pas dans le même ordre que celui correspondant à votre expérience.

Énoncé(s)	**Correction**
Précédamment, il avait déjà exprimé cette opinion.	**Faute**

Règle

Orthographe de « précédemment ».

* * *

Énoncé(s)	**Correction**
Le recours à des procédés réthoriques ne dispense pas du respect de la logique.	**Faute**

Règle

Orthographe de « rhétorique ».

* * *

Énoncé(s)	**Correction**
La réduction de la place du latin dans l'enseignement secondaire cause du tord à la pratique universitaire des Humanités.	**Faute**

Règle

Orthographe de « tort » (dommage, mal).

* * *

Énoncé(s)	**Correction**
Cette proposition n'est ni raisonable, ni rationnelle.	**Faute**

Règle

Orthographe de « raisonnable ».

* * *

Énoncé(s)	**Correction**
Le dysfonctionnement observé est probablement lié à une incompréhension des instructions.	Sans Faute

Règle

« dysfonctionnement » s'écrit bien avec un y.

* * *

Énoncé(s)	**Correction**
L'adresse de l'éditeur doit impérativement figurer dans chaque notice bibliographique.	Sans Faute

Règle

Orthographe de « adresse ». Attention à ne pas confondre avec le « address » de la langue anglaise.

* * *

Énoncé(s)	**Correction**
Parmis ces figures de style, lesquelles vous semblent être des métonymies ?	Faute

Règle

Orthographe de la préposition « parmi ».

* * *

Énoncé(s)	**Correction**
L'épidémie dégrade les conditions d'enseignement, voire décourage complètement certains étudiants.	Sans Faute

Règle

Dans ce contexte-là, « voire » est équivalant à « même ».

* * *

Énoncé(s)	**Correction**
C'est comme même pas courant qu'il neige en mai.	Faute

Règle

Il y a ici confusion entre les quasi-homophones « comme même » et « quand même » (qui signifie « malgré tout »).

* * *

Énoncé(s)	Correction
Quand à vous, je vous donnerai les consignes prochainement.	Faute

Règle

« quant à X » signifie « concernant X », à ne pas confondre avec son homophone « quand » qui est une conjonction signifiant « à ce moment ».

* * *

Énoncé(s)	Correction
Le soi-disant policier n'était en fait qu'un escroc.	Sans Faute

Règle

« soi-disant » signifie « se disant lui-même », le « soi » correspond donc à la personne qui se prétend telle, à ne pas confondre avec son homophone « soit ». Formellement, on ne devrait d'ailleurs pas dire « Cette soi-disant voiture » mais plutôt « Cette prétendue voiture », une voiture n'étant pas douée de parole.

* * *

Énoncé(s)	Correction
Nous constatons une flambée de violance dans les établissements scolaires.	Faute

Règle

Simple erreur d'orthographe pour violence.

* * *

Énoncé(s)	Correction
Le maire a été pris à partie par les manifestants.	Sans Faute

Règle

Dans cet emploi, il s'agit de « partie » et non de « parti ».

* * *

Énoncé(s)	Correction
Il a tiré parti de l'expérience de ses collègues.	Sans Faute

Règle

Dans cet emploi, il s'agit de « parti » et non de « partie ».

* * *

Énoncé(s)	**Correction**
Le beau temps a permis de faire une balade à vélo.	**Sans Faute**

Règle

Dans cet emploi, il s'agit de « balade » (promenade) et non de « ballade » (chanson).

* * *

Énoncé(s)	**Correction**
La brise agitait légèremment le feuillage des arbres.	**Faute**

Règle

L'adverbe se finissant par le son [emã], il ne prend qu'un seul « m ».

* * *

Énoncé(s)	**Correction**
Leur moment était celui de ce battement de cœur raté.	**Sans Faute**

Règle

Les mots sont bien orthographiés.

* * *

Énoncé(s)	**Correction**
Les enfants regardent dans le puit.	**Faute**

Règle

Le mot « puits » est tiré du latin « puteus ». Nous conservons encore cette étymologie dans son orthographe actuel, avec le « -ts ».

* * *

Énoncé(s)	**Correction**
L'acceuil se situe au rez-de-chaussée du bâtiment.	**Faute**

Règle

Pour conserver la prononciation, on écrit -ueil lorsque le son [œj] est précédé par les sons [k] ou [g].

* * *

Énoncé(s)	Correction
Le combattant dépose enfin les armes, triomfant.	Faute

Règle

Le mot est tiré du latin « triumphus », on conserve donc l'étymologie du mot en l'orthographiant avec un -ph- : « triomphant ».

* * *

Énoncé(s)	Correction
As-tu appellé ?	Faute

Règle

Lorsque nous entendons le son [ə], on ne met qu'un seul « l » au verbe « appeler ».

* * *

Énoncé(s)	Correction
Malgrés le beau temps, la course n'a pas pu avoir lieu.	Faute

Règle

Le mot « malgré » est invariable et s'écrit toujours ainsi.

* * *

Énoncé(s)	Correction
J'ai beau laver le tapis, la tâche ne veut pas partir.	Faute

Règle

Dans cette phrase, nous pouvons remplacer « tâche » par « salissure » ; donc la bonne orthographe est « tache », sans accent circonflexe.

* * *

Énoncé(s)	Correction
Il y a d'avantage de personnes touchées par l'épidémie parmi les personnes âgées.	Faute

Règle

« d'avantage » peut être utilisé dans « il y a beaucoup d'avantages à être une grande lectrice », alors qu'il s'agit ici d'un tout autre mot signifiant une plus grande quantité.

* * *

5

Énoncé(s)	Correction
Quelque soient les conditions climatiques, ces pneus tiennent bien la route.	Faute
Quelles que soient les conditions climatiques, ces pneus tiennent bien la route.	Sans Faute

Règle

« Quelque » est utilisé devant un nom, un adjectif ou un adverbe (« quelque souriant qu'il semble, il est malheureux »), mais pas devant un verbe comme ici. Il faut utiliser « quel » et l'accorder.

Énoncé(s)	Correction
Espéré depuis de longues années, cette réforme est une déception aux yeux de la communauté.	Faute

Règle

Accord du participe « espéré » avec « cette réforme ».

* * *

Énoncé(s)	Correction
Il a rencontré ses camarades et leur a donné le document.	Sans Faute
Les étudiants sont curieux ; il faut leur donner matière à réflexion.	Sans Faute

Règle

« leur », pronom personnel, est invariable.

* * *

Énoncé(s)	Correction
Cette semaine, je ne pourrais malheureusement pas assister à la réunion habituelle du mardi.	Faute
Cette semaine, je ne pourrai malheureusement pas assister à la réunion habituelle du mardi.	Sans Faute

Règle

Rien ne justifie ici l'usage du conditionnel. Il convient donc d'écrire « pourrai » correspondant au futur à l'indicatif.

* * *

Énoncé(s)	Correction
Nous pourrons discuter de ce point quand la suite aura été traitée.	Sans Faute

Règle

« Traiter » est ici à la voix passive (« être traité ») et est conjugué au futur antérieur. Il convient donc d'accorder le participe passé avec le sujet du verbe (« la suite »).

* * *

Énoncé(s)	Correction
Je vous renvoie à la discussion de la semaine dernière pour éclairer ce point.	Sans Faute

Règle

« renvoyer » est un verbe du premier groupe. À la première personne du singulier, on a donc : « je renvoie ».

* * *

Énoncé(s)	Correction
Vous devez indiquer les phrases comportants des erreurs.	Faute
Vous devez indiquer les phrases comportant des erreurs.	Sans Faute

Règle

« comportant » est ici participe présent.

* * *

Énoncé(s)	Correction
Le comportement de cet individu laisse pensé qu'il ignore le règlement intérieur.	Faute
Le comportement de cet individu laisse penser qu'il ignore le règlement intérieur.	Sans Faute
Le jeune homme regarda les éclairs déchirer le ciel nocturne.	Sans Faute

Règle

Quand deux verbes se suivent (hors construction avec auxiliaires), le second est à l'infinitif. « penser » doit donc être à l'infinitif.

* * *

Énoncé(s)	Correction
Sans préjuger de la décision qui sera prise, nous pouvons espérer avoir éclairé le débat.	Sans Faute

Règle

Cette phrase suppose une bonne compréhension de la place de chaque verbe dans la seconde partie, avec un enchaînement IND INF INF PPE :

- si deux verbes se suivent, le second est à l'infinitif (« pouvons espérer »)
- le verbe « éclairer » est ici l'infinitif passé (« avoir éclairé »).

* * *

Énoncé(s)	**Correction**
Cette interprétation du texte mériterai d'être davantage diffusée.	**Faute**
Cette interprétation du texte mériterait d'être davantage diffusée.	**Sans Faute**

Règle

À la troisième personne du singulier, le conditionnel de « mériter » s'écrit « mériterait ».

* * *

Énoncé(s)	**Correction**
Je commencerai le tri après qu'il aura fini de réunir tous les livres.	**Sans Faute**

Règle

« après que », tout comme « depuis que », et contrairement à « avant que », s'utilise avec l'indicatif et non le subjonctif.

* * *

Énoncé(s)	**Correction**
Ce manteau coûte quatre cent cinquante euros.	**Sans Faute**

Règle

Une règle veut que « cent » est au pluriel dans « quatre cents », mais pas dans « quatre cent cinquante ».

* * *

Énoncé(s)	**Correction**
Cette voiture coûte chère.	**Faute**

Règle

On accorderait dans la phrase « Cette voiture est chère », mais pas avec le verbe « coûter ».

* * *

Énoncé(s)	**Correction**
C'est entre autre pour cette raison qu'elle s'est inscrite en retard.	**Faute**

Règle

« autre » est forcément au pluriel dans « entre autres » car pour utiliser « entre », il faut au moins deux choses.

* * *

Énoncé(s)	Correction
Il y a quelque 2000 personnes qui se sont rendues à la manifestation.	Sans Faute

Règle

« Quelque » a ici le sens de « environ », et ne s'accorde pas.

* * *

Énoncé(s)	Correction
Il faut finaliser le travail commencer en classe sur les interprétations.	Faute

Règle

La faute provient de l'homophonie avec les verbes du premier groupe, et on ne commettrait pas la faute par exemple dans « le travail fini hier » (et non « le travail finir hier »)

* * *

Énoncé(s)	Correction
L'étudiante attend que le chat errant mange.	Sans Faute
Il entend ces phrases, que l'on prononce et oublie si vite.	Sans Faute

Règle

Les verbes « attendre » et « entendre » à la troisième personne du singulier au présent de l'indicatif s'écrivent bien -end, sans -s.

* * *

Énoncé(s)	Correction
Elle gardait une place dans ses pensées quoi qu'il advînt.	Sans Faute

Règle

« quoi que » est toujours suivi d'un verbe au subjonctif. L'emploi de l'imparfait du subjonctif pour advînt est aussi correct, pour la concordance des temps.

* * *

Énoncé(s)	Correction
Les cours ont repris depuis bientôt un mois.	Sans Faute

Règle

Le participe passé de « reprendre » se termine par un -s.

* * *

Énoncé(s)	Correction
Quand j'ai eu terminé mon travail, j'ai mangé.	Sans Faute

Règle

Il s'agit du passé surcomposé (nous retrouvons deux fois l'auxiliaire avoir) du verbe « terminer ».

Énoncé(s)	Correction
La jeune fille aide à ranger les affaires tombé.	Faute

Règle

« Tombé » est employé ici comme un adjectif et il s'accorde avec le nom qu'il qualifie. On écrira donc : « les affaires tombées ».

* * *

Énoncé(s)	Correction
Seul le bruit de ses pas rythmaient son trajet et ses pensées	Faute

Règle

Le verbe s'accorde avec « le bruit », il convient d'écrire « rythmait ».

* * *

Énoncé(s)	Correction
Il a du revoir son organisation pour finir à temps son travail	Faute

Règle

Le mot « du » est, dans cette phrase, le participe passé de « devoir ». On doit donc ajouter un accent circonflexe et écrire « Il a dû ».

* * *

Énoncé(s)	Correction
Les deux jumeaux se ressemblent à si méprendre	Faute
Les deux jumeaux se ressemblent à s'y méprendre	Sans Faute

Règle

Lorsque nous pouvons remplacer « si/s'y » par « se » (pronom réfléchi), alors il faut utiliser « s'y » suivi d'un verbe. En l'occurrence, la construction du verbe est « se méprendre », on devra donc écrire « s'y méprendre ».

* * *

Énoncé(s)	Correction
Vôtre chat blanc est obèse	Faute

Règle

Si « votre/vôtre » est suivi d'un groupe nominal, alors il s'agit de l'adjectif possessif « votre ». Il faut écrire « Votre chat est obèse ». En revanche, on écrit : « le vôtre est obèse ».

* * *

Énoncé(s)	Correction
Si j'avais assez de temps, j'aimerais lire toutes les œuvres d'Alexandre Dumas.	Sans Faute

Règle

Dans cette phrase, le « si » dénote un souhait, et implique donc l'utilisation du conditionnel ; il faut donc écrire « j'aimerais ».

* * *

Énoncé(s)	Correction
La plupart de ces phrases est fausse.	Faute

Règle

Dans le cas où « la plupart » est suivi d'un complément, on doit accorder le verbe avec le complément. Ainsi, nous devons écrire « La plupart de ces phrases sont fausses ».

* * *

Énoncé(s)	Correction
Je ne veux plus te voir, va-t-en !	Faute

Règle

À l'impératif, les pronoms sont élidés s'ils sont suivis de « en » ou « y ». L'élision s'écrit avec une apostrophe et dans cette phrase, nous devons écrire « va-t'en », puisqu'il s'agit de l'élision de « toi ».

* * *

Énoncé(s)	Correction
Il joue avec moi aux jeux vidéo après qu'il ait dormi.	Faute

Règle

« Après que » est toujours suivi d'un verbe à l'indicatif (l'action a déjà eu lieu). Il faut donc écrire « après qu'il a dormi ».

* * *

Énoncé(s)	Correction
Au vu de leurs notes, elles ont l'air sérieux comme candidates à une bourse de thèse.	Sans Faute

Règle

Lorsque nous utilisons l'expression « avoir l'air » et que « air » est précédé d'un déterminant ou suivi d'un complément, on accorde l'adjectif avec air. Dans cette phrase, « avoir l'air sérieux » est suivi d'un complément (introduit par « comme »), on accorde donc sérieux avec air.

* * *

Énoncé(s)	**Correction**
Les cahiers oranges appartiennent à mon frère.	**Faute**

Règle

Les adjectifs de couleur sont invariables lorsqu'ils sont à l'origine un nom ; en l'occurrence, « orange » désigne à l'origine un fruit.

* * *

Énoncé(s)	**Correction**
Elle se fait fort de ne pas craquer devant ses harceleurs.	**Sans Faute**

Règle

Devant un infinitif, on n'accorde pas l'expression « se faire fort de » : dans ce cas-là, l'expression est figée et est employée au sens « se dire capable de ».

* * *

Énoncé(s)	**Correction**
Ces règles, des quelles on peut déduire toutes les autres, seront dites « axiomes ».	**Faute**
Ils ont formulé de nombreuses propositions, desquelles on pourrait déduire que le sujet les intéresse.	**Sans Faute**

Règle

Dans le premier exemple, il conviendrait d'écrire « desquelles », pronom relatif composé de « de » et « lesquelles ».

* * *

Énoncé(s)	**Correction**
Ce livre est un espèce de mélange entre la fantasy et la science-fiction.	**Faute**

Règle

« Espèce » est un nom féminin.

* * *

Énoncé(s)	Correction
La factrice distribue le courrier à huit heures.	Sans Faute

Règle

« Heures » s'accorde en nombre avec « huit ».

* * *

Énoncé(s)	Correction
L'adolescente porte une robe bleu marine.	Sans Faute

Règle

Lorsque deux adjectifs de couleur sont juxtaposés ou reliés par un trait d'union, aucun ne s'accorde en genre et/ou en nombre.

Énoncé(s)	Correction
Les données que je me suis servi pour ce travail ont été publiées en 2020.	Faute

Règle

Le pronom relatif « dont » doit être utilisé à la place de « que » car l'antécédent « les données » est complément indirect du verbe servir.

* * *

Énoncé(s)	Correction
Cette situation est moins pire que l'autre.	Faute

Règle

On peut paraphraser « pire » par « plus mauvaise », donc « moins pire » correspondrait à « moins plus mauvaise ». Il faut donc seulement dire « moins mauvaise ».

* * *

Énoncé(s)	Correction
C'est l'été qu'Isabelle est le mieux préparée car elle a plus de temps que l'hiver.	Sans Faute

Règle

On n'accorde pas « le mieux » car il s'agit ici d'une comparaison de différents états d'une même personne. On ne compare pas Isabelle à d'autres personnes (auquel cas on accorderait comme dans « Isabelle est la mieux préparée de toutes les candidates »).

* * *

Énoncé(s)	Correction
Je vous serais gré de bien vouloir me retourner le formulaire dans les plus brefs délais.	Faute
Je vous saurais gré de bien vouloir me retourner le formulaire dans les plus brefs délais.	Sans Faute

Règle

Il s'agit de « savoir gré » et non de « être gré »

* * *

Énoncé(s)	Correction
Nous sommes convenus de nous voir prochainement.	Sans Faute
Nous avons convenu d'un rendez-vous prochain.	Faute

Règle

L'auxiliaire avoir ne peut être employé avec ce sens de convenir, mais dans le sens classique tel que dans : « Son cadeau lui a convenu ». On ne convient pas de quelque chose, mais on est convenu de quelque chose.

* * *

Énoncé(s)	Correction
Tu es sûr que c'est bien de Pierre dont tu parles ?	Faute
Tu es sûr que c'est bien de Pierre que tu parles ?	Sans Faute

Règle

Il y a déjà « de » dans « de Pierre », si bien que l'emploi de « dont » crée une redondance. Il faut donc utiliser « que ». En revanche on pourrait dire « Tu es sûr que c'est Pierre dont tu parles ? ».

* * *

Énoncé(s)	Correction
C'est un film dont je me rappelle très précisément.	Faute
C'est un film que je me rappelle très précisément.	Sans Faute

Règle

On « se rappelle quelque chose » (verbe transitif direct), et non « de quelque chose ». L'erreur provient d'une confusion avec le verbe « se souvenir » : « se souvenir de quelque chose ».

* * *

Énoncé(s)	Correction
Aujourd'hui dans cette vidéo je vous partage mon expérience pour réussir vos cheese-cakes.	Faute

Règle

On partage quelque chose « avec » quelqu'un. La bonne structure est donc « Je partage avec vous mon expérience ». En revanche on partage un gâteau en deux.

* * *

Énoncé(s)	Correction
Le festival est sensé se dérouler au printemps.	Faute
Le festival est censé se dérouler au printemps.	Sans Faute

Règle

Dans cette phrase, nous pouvons remplacer le verbe passif « être censé » par « supposer » : « Le festival est supposé se dérouler au printemps ». C'est donc le verbe « censer » qui doit être utilisé ; par ailleurs, il est toujours suivi d'un infinitif, comme c'est le cas dans cet exemple. L'adjectif « sensé » existe bien, mais signifie « avoir du sens », comme dans « Cette proposition est tout à fait sensée ».

* * *

Énoncé(s)	Correction
Pour pallier le manque d'enseignants, l'Éducation supérieure fait souvent appel à des vacataires.	Sans Faute

Règle

Il s'agit d'un verbe transitif direct donnant donc lieu à la structure « pallier quelque chose », et non « pallier à quelque chose ».

* * *

Énoncé(s)	Correction
C'est l'une des raisons pour lesquelles cette décision me semble absurde.	Sans Faute

Règle

Il faut accorder avec « des raisons », car même si l'on se focalise sur l'une d'entre elles, on en évoque plusieurs. Autrement, s'il s'agissait de la seule raison, il faudrait simplement dire : « c'est la raison pour laquelle cette décision me semble absurde ».

Énoncé(s)	Correction
La grève des transports a empêchée Lucie d'arriver à l'heure à son cours.	Faute

Règle

Avec l'auxiliaire avoir, il faut accorder avec le complément d'objet direct (C.O.D.) s'il est placé avant, ce qui n'est pas le cas de « Lucie » ici.

* * *

Énoncé(s)	Correction
Je suis désolée, une panne de voiture m'a empêchée de me rendre à mon rendez-vous.	Sans Faute
Sophie, il l'a envoyée sur les roses !	Sans Faute
Les options que l'étudiant a retenu ne lui permettent pas de se réorienter dans la filière espérée.	Faute
Les devoirs que l'étudiant a rendus sont archivés pendant plusieurs années.	Sans Faute
Les peines qu'elle avait connues s'envolèrent.	Sans Faute
Te souviens-tu de la question qu'a posé l'inspecteur ?	Faute
Combien de cœurs a-t-elle brisés ?	Sans Faute

Règle

Lorsque le C.O.D. est placé avant l'auxiliaire avoir, le participe passé s'accorde.

* * *

Énoncé(s)	Correction
Il m'a suggéré de venir un peu en avance.	Sans Faute

Règle

La règle de l'accord avec un C.O.D. placé avant l'auxiliaire avoir ne concerne cependant pas un Complément d'Objet Indirect (C.O.I.), comme c'est le cas ici, même si celui-ci était féminin.

* * *

Énoncé(s)	Correction
Des pâtes, j'en ai faites plein durant mes études.	Faute

Règle

Bien que le complément d'objet direct (C.O.D.) soit placé avant le verbe « avoir », il y a aussi la présence de « en » qui signale un article partitif et annule l'accord du participe (« J'ai fait plein de pâtes »). On doit donc plutôt écrire « Des pâtes, j'en ai fait plein ».

* * *

Énoncé(s)	Correction
Les poésies qu'il a entendues chanter en Grèce lui ont donné le sens de la prosodie.	Faute
Les poésies qu'il a entendu chanter en Grèce lui ont donné le sens de la prosodie.	Sans Faute
Les poètes qu'il a entendus chanter en Grèce lui ont donné le sens de la prosodie.	Sans Faute
Les poètes qu'il a entendu chanter en Grèce lui ont donné le sens de la prosodie.	Faute

Règle

Quand le participe est suivi d'un infinitif, il s'accorde avec le COD placé en amont si ce COD réalise l'action décrite par l'infinitif.
Dans le premier cas, ce ne sont pas les « poésies » qui chantent, donc le participe ne s'accorde pas avec elles.
Dans le second cas, ce sont bien les « poètes » qui chantent, donc le participe s'accorde en conséquence (« entendus »).

De la complexité de l'annotation manuelle : méthodologie, biais et recommandations

Résumé

Les corpus de référence annotés constituent des éléments primordiaux de nombreuses tâches du Traitement Automatique des Langues. Leur construction fait l'objet d'une attention particulière, notamment lors de campagnes d'annotation manuelle. Ces dernières impliquent de multiples aspects, déjà étudiés dans la littérature mais souvent de manière séparée. Nous présentons une synthèse des problèmes rencontrés lors des différentes étapes d'une campagne, attirant l'attention des gestionnaires sur des points de vigilance, afin qu'ils fassent preuve de prudence durant leur campagne.

Cette thèse donne une première définition des biais d'annotation, qui sont des phénomènes perturbateurs et variés pouvant avoir une incidence sur les annotations. Nous proposons une méthode et des moyens d'observation pour détecter et analyser la présence de biais d'annotation. Deux campagnes d'annotation, menées spécialement dans le but d'étudier des biais particuliers, servent d'illustration et nous ont permis de constater l'influence tangible de certains paramètres sur l'annotation. Dans cette optique, nous avons aussi introduit la notion de consensualité, qui permet en particulier de situer un annotateur par rapport à un groupe. Nous montrons un premier lien entre les annotateurs les moins consensuels et les moins performants.

Mots clefs Annotation manuelle ; Biais d'annotation ; Corpus annoté ; Traitement Automatique des Langues ; Méthodologie

On the complexity of manual annotation : methodology, bias and recommendations

Abstract

Annotated reference corpora are essential elements of many tasks in Natural Language Processing. Their construction is the object of particular attention, especially during manual annotation campaigns. The latter involve multiple aspects, already studied in the literature but often separately. We present a synthesis of the problems encountered during the different stages of a campaign, drawing the attention of managers to points of vigilance, so that they can be careful during their campaign.

This thesis gives a first definition of annotation biases, which are disturbing and varied phenomena that can have an impact on annotations. We propose a method and means of observation to detect and analyze the presence of annotation bias. Two annotation campaigns, conducted specifically to study particular biases, serve as illustrations and have allowed us to observe the tangible influence of certain parameters on the annotation. In this perspective, we have also introduced the notion of consensuality, which allows us to situate an annotator in relation to a group. We show a first link between the least consensual annotators and the least efficient ones.

Keywords: Manual annotation; Annotation bias; Annotated corpora; Natural language processing; Methodology