Redonner du sens à l'accord interannotateur : vers une interprétation des mesures d'accord en termes de reproductibilité de l'annotation

Dany Bregeon[*,***] **— Jean-Yves Antoine**[*] **— Jeanne Villaneau**[**] **— Anaïs Halftermeyer**[***]

[*] *LIFAT (EA 6300), ICVL, Université de Tours*
jean-yves.antoine@univ-tours.fr
[**] *IRISA (UMR 6074) D6-Expression,Université de Bretagne Sud*
jeanne.villaneau@univ-ubs.fr
[***] *LIFO, ICVL, Université d'Orléans*
dany.bregeon@etu.univ-orleans.fr, anais.halftermeyer@univ-orleans.fr

RÉSUMÉ. *Les mesures d'accord interannotateur sont utilisées en routine par le TAL pour évaluer la fiabilité des annotations de référence. Pourtant, les seuils de confiance liés à cette estimation relèvent d'opinions subjectives et n'ont fait l'objet d'aucune expérience de validation dédiée. Dans cet article, nous présentons des résultats expérimentaux sur données réelles ou simulées qui visent à proposer une interprétation des mesures d'accord en termes de stabilité de la référence produite, sous la forme d'un taux moyen de variation de la référence entre différents groupes d'annotateurs.*

ABSTRACT. *Inter-coders agreement measures are used to assess the reliability of annotated corpora in NLP. Now, the interpretation of these agreement measures in terms of reliability level relies on pure subjective opinions that are not supported by any experimental validation. In this paper, we present several experiments on real or simulated data that aim at providing a clear interpretation of agreement measures in terms of the level of reproductibility of the reference annotation with any other set of coders.*

MOTS-CLÉS : *accord interannotateur, reproductibilité, niveau de fiabilité.*

KEYWORDS: *inter-coders agreement, reproductibility, reliability level.*

1. Introduction

Utilisant de manière intensive des techniques d'apprentissage automatique entraînées sur des corpus, le traitement automatique des langues (TAL) a un besoin de plus en plus insatiable de ressources langagières massives. Face à ces besoins toujours croissants, le TAL a désormais recours fréquemment à ces corpus annotés automatiquement. La masse de données alors produite constitue une contrepartie intéressante du biais introduit, du moins si l'on fait le pari que les systèmes n'apprendront pas ce biais. Le recours à des ressources de qualité, annotées ou révisées manuellement, reste toutefois toujours pertinent, soit que l'on ne dispose pas de solution d'annotation automatique efficace, soit que la qualité des données d'apprentissage est un plus pour l'application visée.

La qualité des données annotées va au-delà des erreurs observées par rapport au guide d'annotation. Dans le cas de tâches complexes et/ou soumises à une forte subjectivité (pensons par exemple à la détection d'émotion), cette qualité répond au contraire avant tout à une exigence de fiabilité, c'est-à-dire de reproductibilité de l'annotation. Un corpus ne pourra en effet être considéré comme fiable et représentatif de la tâche considérée que si les annotations obtenues sont reproductibles par d'autres annotateurs que ceux choisis à l'initial (à l'idéal).

Dans le cas de corpus annotés par plusieurs personnes, on estime cette reproductibilité en observant l'accord qui existe entre chaque annotation individuelle. Le principe sous-jacent ici est que plus les annotateurs s'accordent entre eux, plus il y a de chances qu'ils se seraient également accordés avec n'importe quels autres annotateurs. Les chances que l'annotation soit reproductible sont donc d'autant plus élevées que l'accord entre les annotateurs l'est également.

Plusieurs métriques d'évaluation de l'accord interannotateur ont été proposées dans la littérature et sont utilisées en routine en TAL. Les plus répandues dans notre communauté sont ainsi le κ de Cohen (1960) et ses différents avatars, ou, plus récemment, le α de Krippendorff (2008) .

Pour un état de l'art récent, complet et très fouillé de la question de l'estimation de l'accord interannotateur, on pourra consulter (Mathet, 2017a). Dans cette introduction nous nous contenterons de constater que ces métriques diffèrent uniquement par leur façon de corriger l'accord brut observé sur le corpus par une estimation de la part de chance due au hasard dans cet accord. Les subtilités statistiques sous-jacentes à cette estimation de l'accord par hasard ne sont pas sans intérêt. Il a ainsi été démontré qu'elles avaient un impact sur la qualité d'estimateur de la reproductibilité, en particulier dans le cas d'annotations ordinales (valeurs discrètes ordonnées) (Antoine *et al.*, 2014). Dans cet article, nous souhaitons aborder toutefois une question bien trop négligée de notre point de vue, alors qu'elle revêt une grande importance pratique : il s'agit du manque d'intelligibilité des valeurs d'accord retournées par ces métriques.

Quelle que soit la métrique considérée, celle-ci retourne une valeur d'accord corrigée par la chance d'une valeur maximale de 1 (accord parfait). Du fait de la correction

statistique, il est difficile de relier cette valeur d'accord avec une idée claire de la qualité de l'annotation. Au cours du temps, de nombreux auteurs (cf. figure 1, page 51) ont proposé des seuils de qualité acceptable pour chaque métrique, sans que ces seuils ne reposent sur une argumentation démontrée. Cet article présente précisément des résultats expérimentaux qui tendent à faire sortir l'évaluation de l'accord interannotateur du seul argument d'autorité en matière de seuils d'acceptabilité. Nous proposons pour cela de revenir au critère de reproductibilité qui a fondé les recherches sur le sujet. Notre étude vise, en effet, à relier la mesure de l'accord interannotateur à l'estimation du taux de modifications de l'annotation que l'on obtiendrait avec un autre ensemble d'annotateurs.

Dans un premier temps, nous allons faire une brève présentation de la problématique de l'évaluation de l'accord interannotateur en nous attachant avant tout à décrire tous les facteurs qui peuvent influer sur l'estimation de cet accord, et les difficultés qu'il y a à interpréter les mesures obtenues. Nous proposerons ensuite de réinterpréter ces mesures en reliant la mesure d'accord interannotateur avec la stabilité de l'annotation obtenue avec n'importe quel ensemble d'annotateurs d'une taille donnée. Ainsi, nous lions directement accord interannotateur et reproductibilité de l'annotation.

Cette réinterprétation passe par la mise en place d'une batterie d'expérimentations portant sur des données annotées réelles ou simulées. La quatrième section de cet article présente en détail le cadre méthodologique que nous avons adopté pour cette étude, en insistant en particulier sur la technique de génération d'annotations simulées à partir de données réelles que nous avons utilisée. Nous présenterons enfin les résultats expérimentaux que nous avons obtenus en nous focalisant dans un premier temps sur le κ de Cohen. Ces résultats suggèrent qu'il est possible d'établir une corrélation entre valeurs d'accord interannotateur et taux de reproductibilité de l'annotation. La conclusion nous permettra enfin de détailler l'ensemble des études qui restent à conduire pour arriver à une interprétation directe réellement opérationnelle des métriques d'accord interannotateur.

2. Estimation de l'accord interannotateur : état de l'art et limitations

2.1. *Processus d'annotation*

L'annotation recouvre des processus d'enrichissement de corpus variés, parmi lesquels on distingue deux grandes classes d'activités, qui peuvent, suivant la tâche concernée, correspondre à deux étapes successives d'annotation (Mathet, 2017b) :

– la segmentation, appelée *unitizing* chez Krippendorff (2013), consiste à localiser des unités dignes d'intérêt dans le flux langagier. Elle peut être continue, à savoir qu'elle couvre l'ensemble d'un texte à annoter, ou discontinue. Dans ce second cas, seules quelques portions du texte à annoter seront localisées. C'est par exemple le cas de l'annotation en entités nommées ou en mentions référentielles, où seuls quelques mots ou expressions polylexicales seront localisés dans le continuum du texte ;

– la catégorisation revient, quant à elle, à associer une description linguistique à des unités déjà caractérisées. Elle peut se limiter à associer une catégorie à chaque unité. C'est par exemple le cas d'une annotation en entités nommées où l'on associerait un type d'entité (personne, organisation, lieu...) à chaque unité. La catégorisation peut également être bien plus fine et associer tout un ensemble de traits qualificatifs aux entités étudiées.

La catégorisation peut concerner le document dans sa globalité, auquel cas elle ne succède pas à une étape préalable de segmentation. Dans d'autres situations, la détermination des unités d'intérêt est immédiate, voire automatisable, et ne pose pas de problème de fiabilité. Toutefois, le processus d'annotation englobe en général les deux étapes de segmentation et de catégorisation. La nature très différente de ces deux activités (localiser et qualifier) fait que les guides d'annotation recommandent souvent de les réaliser de manière séparée. Il semble, dès lors, avisé d'évaluer séparément la fiabilité de ces deux opérations. C'est, par exemple, la démarche que nous avons adoptée pour le corpus ANCOR annoté en coréférence (Muzerelle *et al.*, 2014).

Il reste bien sûr envisageable de conduire une évaluation unique, intégrant segmentation et catégorisation, de la fiabilité des annotations. Se pose alors la délicate question de l'alignement des segmentations dans le calcul de l'accord, ainsi que celle de l'estimation d'un accord par chance sur cette délimitation des segments. La famille de métriques γ définie par Yann Mathet (2017, 2017b) constitue la proposition la plus aboutie en la matière. Ce γ ne résout toutefois pas la question de l'intelligibilité des mesures d'accord interannotateur que nous allons étudier dans cet article.

C'est pourquoi nous avons décidé d'étudier cette question, qui n'a jamais été abordée frontalement à notre connaissance, en nous focalisant sur la question de la fiabilité de l'étape de catégorisation, ainsi que sur la métrique la plus répandue pour l'estimer : le κ de Cohen (1960). Ceci, sans ignorer l'existence de propositions alternatives de métriques, qui feront l'objet d'études ultérieures de notre part.

2.2. *La famille de métriques* κ

Dès qu'une annotation de référence est obtenue à partir de plusieurs annotations concurrentes, l'estimation de sa fiabilité repose sur le calcul de l'accord entre les annotateurs. L'accord brut entre les annotateurs ne peut toutefois tenir lieu de bon estimateur de la qualité de la référence, car il n'intègre pas la part d'aléatoire (accord par chance) qui entre dans la mesure finale observée. Cet accord au hasard a pourtant un impact évident sur l'accord brut observé, puisqu'il est *a priori* plus facile d'obtenir un bon accord lorsque la catégorisation ne concerne que deux classes d'annotation que lorsqu'elle en implique dix. On peut songer au cas limite où il n'y aurait qu'une seule classe d'annotation et où l'accord serait, de fait, parfait dès le départ.

Les mesures de fiabilité de l'annotation en catégorisation corrigent l'accord brut observé par une estimation statistique de la part d'accord qui est due à la chance. Les métriques diffèrent par la manière selon laquelle elles estiment cet accord au hasard.

Considérons une tâche d'annotation nominale qui consiste à affecter à chaque entité une catégorie parmi un ensemble de valeurs totalement indépendantes (par exemple, un type d'entité nommée). Le κ de Cohen est estimé par la formule générale [1], où A_o est l'accord brut estimé entre les annotateurs, et A_e est l'estimation de l'accord qui est dû à la chance :

$$\kappa = \frac{A_o - A_e}{1 - A_e} \qquad [1]$$

Pour estimer A_e, Cohen postule que l'accord par hasard dépend uniquement du comportement individuel de chaque utilisateur, qu'il résume par la distribution statistique des catégories d'annotation utilisées par chacun d'entre eux. Selon cette hypothèse, l'accord entre deux annotateurs sera d'autant plus élevé que leur fréquence d'utilisation de chaque catégorie est proche. Ainsi, pour une tâche de catégorisation avec N items annotés, réalisée par deux annotateurs, A_e est estimé par la formule suivante, où n_c^i correspond au nombre de fois où la catégorie c a été utilisée par l'annotateur i :

$$A_e = \frac{1}{N^2} . \sum_c n_c^1 n_c^2 \qquad [2]$$

Cette estimation correspond à une annotation en catégories nominales par deux utilisateurs. Cohen (1968) a proposé une généralisation de la métrique à une catégorisation ordinale ou multivaluée. Cette situation survient, par exemple, pour l'annotation en émotion, où des tours de parole sont catégorisés suivant une échelle ordinale de valences (– 2 = très négatif, – 1 = négatif, 0 = neutre, 1 = positif, 2 = très positif). Dans ce cas, le κ est dit pondéré, puisque la métrique tient compte du fait qu'un désaccord entre deux catégories proches est moins grave que celui entre deux catégories éloignées. Cette généralisation consiste donc à passer d'une distance binaire entre catégories à une distance euclidienne.

Enfin, Davies et Fleiss (1982) ont, de leur côté, défini une généralisation du κ binaire à un ensemble quelconque d'annotateurs. Dans leur proposition, la valeur de A_e correspond à la moyenne, sur l'ensemble des paires des P annotateurs, des valeurs de A_e définies par la formule [2] pour deux annotateurs.

$$A_e = \sum_c \frac{2}{P(P-1)} \sum_{m\,=\,1}^{P\,-\,1} \sum_{n\,=\,m\,+\,1}^{P} \frac{n_c^m n_c^n}{N^2} \qquad [3]$$

Comme le rappellent Artstein et Poesio (2008) dans un état de l'art très complet sur l'accord interannotateur, il n'existe toutefois pas à ce jour de proposition concernant la métrique κ qui soit à la fois pondérée et adaptée à un nombre variable d'utilisateurs. La métrique α définie par Krippendorff (2004) permet au contraire une généralisation qui englobe à la fois un nombre quelconque d'annotateurs et une distance euclidienne entre classes d'annotation. Dans le cadre de cette étude, qui relève avant

tout de la preuve de concept, nous avons considéré la version binaire multi-utilisateur de κ telle que proposée par Davies et Fleiss (1982) (cf. formule [3]). Nos travaux futurs concerneront toutefois aussi bien le α que le κ.

2.3. *Biais d'estimation de l'accord interannotateur*

La question de la pertinence des valeurs fournies par les différentes métriques d'accord interannotateur a fait l'objet de nombreuses études expérimentales ou théoriques. Celles-ci se sont avant tout intéressées aux facteurs d'impact qui peuvent biaiser l'estimation de l'accord interannotateur, indépendamment de tout questionnement sur l'interprétation directe d'une valeur d'accord donnée. On peut citer ici quelques-uns des biais potentiels les mieux identifiés dans la littérature :

1) Biais annotateur et nombre d'annotateurs

Le biais annotateur est une question centrale en termes de fiabilité des données, et il sera au centre des expérimentations présentées dans cet article. Il est en effet directement relié à la notion de reproductibilité, puisqu'il décrit l'influence du comportement idiosyncratique d'un annotateur donné sur la construction de la référence. Il peut être estimé par une mesure (bias index) qui quantifie l'amplitude des variations, entre chaque annotateur, de la distribution de la fréquence d'utilisation de chaque catégorie d'annotation (Sim et Wright, 2005). Le κ cherche précisément à intégrer ce biais dans l'estimation de l'accord par chance. La bibliographie est toujours partagée sur la pertinence de cette prise en compte. S'appuyant sur des considérations purement théoriques, Feinstein et Cicchetti (1990) et Di Eugenio et Glass (2004) affirment que ce biais peut avoir un impact sur les valeurs d'accord obtenues. Ils notent en particulier que l'estimation de l'accord par chance A_e sera biaisé si la distribution des annotations varie fortement d'un expert à l'autre. Artstein et Poesio (2008) réfutent au contraire cet argument, en estimant que cette influence concernera A_o et A_e de concert, ce qui en limite l'impact. En dépit de ces controverses, les études expérimentales menées sur le sujet permettent d'arriver à un consensus sur un point : plus le nombre d'annotateurs mobilisés pour construire la référence est élevé, plus le biais annotateur est limité.

2) Prévalence d'une catégorie donnée

La prévalence traduit l'existence d'une surreprésentation d'une catégorie donnée dans les choix d'annotation des annotateurs, que cette prédominance soit due à un biais d'annotation ou résulte de la nature même des données. Dans une telle situation, la probabilité d'arriver à un accord est bien entendu plus importante : la correction due à l'accord par chance A_e est alors susceptible d'être plus importante, et de réduire d'autant la valeur du κ final (Brennan et Silman, 1992 ; Di Eugenio et Glass, 2004). Sim et Wright (2005) ont proposé là encore de définir une mesure (*prevalence index*) pour quantifier l'importance de la prévalence dans les données annotées. Ce *prevalence index* est directement intégré dans le calcul du PABAK, une adaptation du κ de Cohen cherchant à limiter ce biais (Byrt *et al.*, 1993).

3) Nombre de catégories

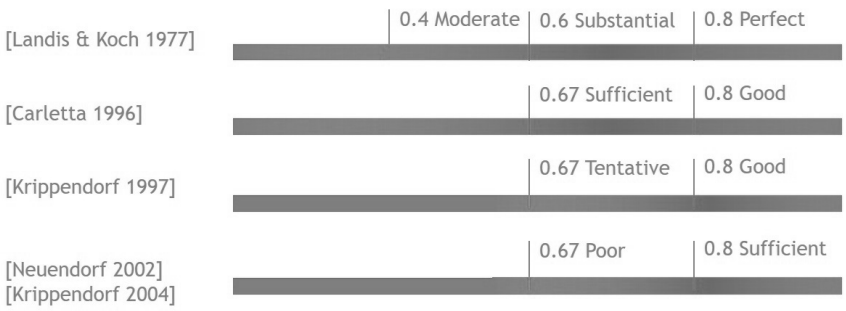

Figure 1. *Échelles subjectives de fiabilité de l'annotation en fonction de l'accord*

Nous avons vu plus haut que le nombre de catégories d'annotation pouvait avoir une influence directe sur la valeur d'accord interannotateur brut Ao. La correction par l'accord au hasard A_e doit limiter cet impact. Les études expérimentales menées sur le sujet ont toutefois montré que les valeurs moyennes de κ observées baissent lorsque le nombre de catégories d'annotation augmente (Brenner et Kliebsch, 1996).

On le voit, les métriques d'accord interannotateur telles que le κ sont potentiellement affectées par de nombreux facteurs d'influence propres aux caractéristiques de l'annotation (nombre de catégories, nombre d'annotateurs, etc.). Ces biais, largement étudiés dans la littérature, rendent difficile l'interprétation d'une valeur d'accord interannotateur donnée. Une même valeur de κ obtenue sur deux annotations différant totalement d'un point de vue méthodologique permet-elle la même conclusion quant à la fiabilité des données annotées, et, dès lors, à partir de quel seuil de κ peut-on estimer qu'une annotation de référence est de qualité suffisante ?

2.4. *Quelle interprétation objective des mesures d'accord interannotateur ?*

L'étude des différents biais qui peuvent entacher la mesure de l'accord interannotateur explique la difficulté qu'il y a à définir une échelle d'interprétation objective de ces mesures. Pourtant, nous avons besoin d'une telle lecture objective, afin de pouvoir associer directement valeur d'accord et fiabilité de l'annotation. Aucune recherche n'a pourtant cherché à creuser cette question à notre connaissance. Ainsi, alors que les métriques d'accord interannotateur mobilisent des calculs statistiques subtils pour nous renseigner sur la qualité de nos données, l'interprétation finale de leur valeur n'a, jusqu'ici, répondu qu'à des échelles subjectives relevant de l'intime opinion de leurs auteurs. On ne compte ainsi plus le nombre de propositions de seuils de bonne fiabilité liés à ces métriques, comme le montre la figure 1.

Tous les auteurs semblent considérer qu'un accord supérieur à 0,8 est un gage relativement satisfaisant de fiabilité des données. Ce seuil porte toutefois une signification très variable d'un auteur à l'autre : là où Landis et Koch (1977) parlent de qualité par-

faite, celle-ci n'est que suffisante par Neuendorf (2002) et Krippendorff (2004). Pour des valeurs inférieures d'accord, les divergences de vue sont encore plus sensibles. À la suite de l'article fondateur de Carletta (1996), le seuil de 0,67 est le plus souvent retenu comme gage de fiabilité acceptable par notre communauté. Elle est pourtant jugée faible par Neuendorf (2002) et Krippendorf (2004), tandis que Landis et Koch (1977) se satisfont même d'un accord interannotateur à 0,4... Cette diversité d'avis résume parfaitement la fragilité méthodologique de ces échelles, qui ne semblent relever que de l'argument d'autorité. Elle est encore évidente lorsque l'on observe qu'un auteur aussi rigoureux que Krippendorff révise lui-même au fil du temps son jugement, sans justifier cette réévaluation. L'objectif des travaux que nous présentons dans cet article est précisément de répondre à cette faiblesse méthodologique, en conduisant des expériences pour élaborer une interprétation objective des valeurs d'accord interannotateur en termes de niveau de reproductibilité.

3. Interpréter les valeurs d'accord en termes de niveau de reproductibilité

L'idée centrale de nos travaux est de revenir à la notion première de fiabilité d'une annotation : une annotation doit être jugée de bonne qualité non pas parce que son accord interannotateur est jugé acceptable suivant une échelle très subjective, mais parce que la référence construite avec tout autre ensemble d'annotateurs reste stable (exigence de reproductibilité). Nous proposons donc d'interpréter la valeur de l'accord interannotateur en termes de taux moyen de variation de la référence que l'on obtiendrait avec d'autres ensembles d'annotateurs.

Notre objectif est donc de conduire une étude expérimentale sur des jeux de données variés, correspondant à des annotations réelles ou simulées à partir de données réelles, pour une correspondance entre valeur d'accord interannotateur et taux de variation de la référence. Nous avons privilégié ici une démarche expérimentale, et non pas une étude statistique théorique, car la question de l'accord interannotateur est hautement multifactorielle (nombreux biais d'estimation bien établis dans la littérature, propositions multiples de métriques d'accord reposant sur des hypothèses fortes sur le comportement des annotateurs qui rendent difficile une théorisation du lien entre accord et stabilité de la référence). Notre article se veut toutefois également une incitation à des tentatives de formalisation plus poussées de ce questionnement méthodologique.

Nous avons observé dans la section précédente que les valeurs d'accord mesurées avec le κ étaient susceptibles de dépendre du nombre d'annotateurs et du nombre de catégories d'annotation. Les correspondances que nous désirons établir à terme dépendront donc de :

– la métrique d'estimation de l'accord (ici : κ) ;

– le nombre d'annotateurs : P ;

– le nombre de catégories : C.

Notre objectif est ainsi d'aller vers une interprétation des mesures d'accord en termes de niveau de stabilité de l'annotation de référence. Ceci pour que les concepteurs de corpus soient à même de juger si le nombre d'annotateurs, de classes ou encore si la complexité de la tâche et la précision du guide d'annotation permettent de fournir des données fiables car reproductibles. Par exemple, un concepteur de corpus qui désirerait s'assurer que son annotation de référence est fiable avec une marge de 5 % d'erreur cherchera la valeur d'accord interannotateur qui lui assure, pour P et C donnés, d'un taux de variation moyen de la référence de 5 % au maximum.

Il convient de noter que l'utilisation d'un vote majoritaire et, entre autres, l'usage d'un tirage aléatoire en cas de ballottage, complexifie encore un peu plus le lien théorique entre valeurs de κ et stabilité de la référence produite. Les travaux qui sont présentés dans cet article sont donc purement expérimentaux et ne fournissent pas encore de tables complètes de fiabilité, mais ils ambitionnent de démontrer la faisabilité et l'intérêt de cette démarche. Nos résultats nous permettront toutefois de proposer des premiers éléments de compréhension du lien entre stabilité de la référence et échelles de fiabilité subjectives proposées dans la littérature. Notons enfin que nos travaux s'inscrivent dans un champ d'application bien précis : celui d'une annotation de type catégorisation d'observables, obtenue par vote majoritaire sur les annotations individuelles et réalisée par plusieurs (au moins trois) experts.

4. Méthodologie expérimentale

Cette partie décrit l'ensemble du cadre expérimental qui a été développé pour estimer le taux moyen de variation d'une annotation en regard de l'accord interannotateur observé sur une annotation de référence. Dans un premier temps, nous décrivons les principes sous-jacents à l'estimation de la stabilité d'une annotation, puis les données, réelles et ensuite simulées, qui ont été utilisées pour nos expériences.

4.1. *Estimation de la stabilité d'une annotation : principes*

Considérons une annotation de type catégorisation réalisée par une population de P annotateurs dont la tâche est d'associer à chacune des N entités d'intérêt (ou *observables*) du corpus une catégorie parmi un ensemble de C catégories. L'annotation est considérée comme parfaitement reproductible si n'importe quel groupe de k personnes, choisies au hasard dans la population, produit toujours la même annotation. Cette annotation idéale correspond à celle que produirait un nombre infini d'annotateurs. Si la reproductibilité n'est pas parfaite, plusieurs annotations différentes peuvent être observées suivant le groupe considéré.

Soit $\mathcal{G} = \{G_i | i \in \{1 : n\}\}$, un ensemble de n groupes comportant chacun k annotateurs, pour des valeurs de k et n données. On note κ_i l'accord interannotateur du groupe G_i. Chacun des groupes G_i produit une référence $R_i = (r_{i_j}, j \in \{1 : N\})$ (par exemple par vote majoritaire), l'annotation la plus fréquente sur l'ensemble

des annotateurs de tous les groupes rassemblés étant considérée comme la *référence absolue*, notée $R = (r_j, j \in \{1 : N\})$. On note τ_i, le taux d'erreurs de la référence produite par le groupe G_i par rapport à la référence absolue R. τ_i se définit par :

$$\tau_i = \frac{1}{N} \sum_{j=1}^{N} d(r_{i_j}, r_j) \quad [4]$$

où $d(a, b) = 0$ si $a = b$ et $d(a, b) = 1$ si $a \neq b$ (distance discrète).

L'accord interannotateur moyen calculé sur l'ensemble $\mathcal{G}$ est la moyenne des accords κ_i calculés sur chacun des groupes G_i :

$$\kappa_{\mathcal{G}} = \frac{1}{n} \sum_{i=1}^{n} \kappa_i \quad [5]$$

alors que le taux d'erreurs moyen calculé sur $\mathcal{G}$ se calcule par

$$\tau_{\mathcal{G}} = \frac{1}{n} \sum_{i=1}^{n} \tau_i \quad [6]$$

Si le nombre d'annotateurs P est suffisamment grand pour que l'on puisse considérer que la référence absolue obtenue est proche d'une référence idéale, et si k est suffisamment petit par rapport à P, alors $\tau_{\mathcal{G}}$ peut être considéré comme une approximation valide du taux de variation attendu de la référence construite avec k annotateurs avec une référence idéale.

Compte tenu des données dont nous disposons, à savoir un corpus annoté par P annotateurs, nous réalisons cette approximation en respectant la procédure suivante :

1) nous calculons la référence absolue R suivant une procédure de vote majoritaire entre les votes des P annotateurs ;

2) nous considérons un nombre k d'annotateurs, avec $k < P$. Pour que la notion de vote majoritaire entre ces annotateurs devienne efficiente, k doit être strictement supérieur à 2[1]. D'où : $2 < k < P$;

3) les sous-ensembles de k annotateurs parmi P sont au nombre de $\complement_P^k$. Suivant leur nombre, ils sont considérés en totalité ou non[2] pour obtenir un ensemble $\mathcal{G}$ de n groupes de k annotateurs avec n assez grand (au moins plusieurs centaines) ;

4) pour chaque sous-ensemble G_i de k annotateurs, nous calculons leur accord interannotateur κ_i et construisons la référence R_i produite par leurs annotations ;

5) nous construisons la référence absolue R correspondant aux P annotateurs par vote majoritaire. Nous calculons également un accord interannotateur théorique $\kappa_{\mathcal{G}}$ comme étant la moyenne des κ_i obtenus sur les différents groupes de k annotateurs ;

1. Lorsque deux classes obtiennent le même nombre de votes sur un observable donné, elles sont départagées par un tirage aléatoire.

2. Pour limiter les temps de calcul, nous nous limitons au tirage aléatoire de 1 000 d'entre eux lorsque leur nombre dépasse plusieurs milliers.

6) la comparaison entre la référence $R = (r_j, j \in \{1 : N\})$ et les références $R_i = (r_{i_j}, j \in \{1 : N\})$ nous donne alors le taux de variation moyen obtenu à partir des annotations produites par les groupes de k annotateurs (cf. formules [4] et [6]) :

$$\tau_{\mathcal{G}} = \frac{1}{n} \sum_{i=1}^{i=n} \sum_{j=1}^{j=N} d(r_j, r_{i_j}) \qquad [7]$$

où d est la distance discrète.

À l'issue de cette procédure, on dispose d'une valeur d'accord interannotateur $\kappa_{\mathcal{G}}$ qui peut être mise en correspondance avec une moyenne des taux de variation $\tau_{\mathcal{G}}$.[3]

4.2. *Corpus et annotations*

Nos expérimentations ont été menées sur les données annotées de quatre corpus différents[4], chacun d'entre eux ayant été collecté pour une tâche spécifique et pour les besoins d'autres projets dans lesquels certains d'entre nous étaient partenaires. Nous sommes en présence d'une simple annotation en classes, tout problème de segmentation préalablement résolu. Ces corpus ont été également choisis car ils impliquaient un grand nombre d'annotateurs. Cette caractéristique sert nos objectifs expérimentaux, mais nous verrons que nos conclusions concernent également les annotations à nombre réduit de codeurs (supérieur ou égal à trois).

4.2.1. *Corpus en émotion*

Une annotation en émotion consiste à ajouter au texte des informations quant à l'émotion que peut ressentir son lecteur. Il n'existe pas de consensus concernant la façon de décrire une émotion dans une tâche d'annotation mais, quelle que soit l'approche choisie, les accords entre les annotateurs sont généralement médiocres, voire faibles, ce qui conduit à prendre les avis d'un grand nombre de codeurs pour définir une annotation de référence (Schuller *et al.*, 2009).

Nos travaux concernaient la détection automatique de l'émotion dans des contes destinés aux enfants. Considérant la difficulté de la tâche, nous avons choisi l'approche

3. Une autre procédure a été expérimentée sur les données réelles : elle consiste à comparer les références produites par deux groupes disjoints de k annotateurs pris parmi les P annotateurs disponibles, ce qui a nécessité d'imposer sur k la contrainte : $2 < k < \frac{P}{2}$. Moyenné sur toutes les paires de groupes disjoints de k annotateurs possibles, le calcul du taux d'erreurs donne également une approximation de la non-reproductibilité. Cette procédure est celle qui avait été mise en œuvre dans Antoine *et al.* (2014). Elle donne des résultats très proches de la procédure utilisée dans les travaux relatés dans cet article.

4. Ces corpus sont disponibles, à fin de reproductibilité, directement auprès des auteurs, à savoir : jean-yves.antoine@univ-tours.fr, jeanne.villaneau@univ-ubs.fr ou encore anais.halftermeyer@univ-orleans.fr.

la plus simple, qui consiste à classifier les émotions suivant une échelle multidimensionnelle ; les deux dimensions essentielles sont la *valence* qui permet de préciser si l'émotion est positive, négative ou neutre et l'*intensité* qui précise le niveau de l'émotion ressentie. Une annotation ordinale en cinq classes {– 2, – 1, 0, 1, 2} permet de réaliser une classification qui combine la *valence* et l'*intensité* réunies. Cette annotation se réduit à trois classes {– 1, 0, 1} si seule la *valence* est considérée.

Le corpus émotion utilisé dans cette étude regroupe 230 phrases issues de deux textes peu connus (Vassallo, 2004 ; Vanderheyden, 1995) et annotées par 25 codeurs. Pour les besoins de nos travaux, deux annotations différentes ont été réalisées : chaque phrase considérée isolément (annotation hors contexte) ou présentée dans l'ordre du récit (annotation en contexte)[5].

4.2.2. *Corpus d'opinion*

Les principes de l'annotation en opinion sont très similaires à ceux pratiqués pour l'opinion : la *polarité* y joue le rôle de la *valence* et le terme d'*activation* peut être utilisé pour désigner l'intensité.

Le corpus est composé de 183 phrases qui expriment des avis déposés sur le site `www.allocine.fr`. Elles ont été annotées par le même groupe d'annotateurs, avec la même échelle de valeurs et dans des conditions strictement semblables à celles du corpus précédent, y compris pour ce qui est de la présentation des phrases hors et en contexte.

4.2.3. *Corpus de coréférence*

Le corpus d'annotation en coréférence utilisé ici est un échantillon produit pour mesurer l'accord interannotateur sur une ressource d'envergure, le corpus ANCOR (Muzerelle *et al.*, 2014). Cet échantillon, qui correspond à un dialogue court à forte interaction a été annoté par 9 annotateurs experts (étudiants de master, doctorants, et enseignants-chercheurs dans la thématique). La tâche de résolution de la coréférence peut être découpée en trois phases :

1) l'identification des mentions référentielles (segmentation) ;

2) l'identification de lien de coréférence entre paires de mentions (segmentation) ;

3) le typage de la relation de coréférence pour chaque paire dont une relation a été identifiée (catégorisation).

Nous utilisons ici uniquement la dernière sous-tâche (3), chaque annotateur a dû choisir pour chaque paire proposée une classe de relation parmi cinq disponibles :

– coréférence directe : les deux mentions présentent la même tête lexicale (*la maison ... cette maison*) ;

5. Pour plus de précisions sur le corpus et son annotation, on pourra se reporter à (Le Tallec *et al.*, 2011).

– coréférence indirecte : les deux mentions présentent deux têtes lexicales différents (*la maison ... cette demeure*) ;

– coréférence pronominale : la seconde mention est un pronom (*la maison ... elle*) ;

– anaphore associative : les deux mentions ne sont pas coréférentes mais il est nécessaire de connaître l'interprétation sémantique de la première pour comprendre la seconde (*cette maison ... la porte*) ;

– anaphore associative pronominale : cette relation présente la même dépendance sémantique entre mentions que pour l'anaphore associative et la seconde mention est un pronom (*cette maison ... ils* s'agissant des habitants de la maison).

4.2.4. *Similarité entre phrases*

Évaluer la similarité entre deux phrases est une tâche classique du TAL, souvent utilisée comme sous-tâche de tâches plus ambitieuses, telles que le résumé automatique. SemEval a proposé des confrontations de systèmes sur ce thème à partir de 2012 et jusqu'en 2017, pour l'essentiel en langue anglaise.

Les données utilisées dans cette étude correspondent à deux petits corpus en langue française créés pour les besoins de nos travaux : le premier a pour thème la conquête spatiale, le second porte sur le thème des épidémies. Pour chacun des corpus, soixante-dix phrases ont été sélectionnées. Dix d'entre elles ont servi de phrases de référence, qui contiennent des informations importantes sur le domaine testé. Chacune de ces phrases a été associée à six autres phrases du corpus, choisies pour que différents niveaux de similarité avec elle soient représentés. Dans chacun des corpus, les soixante paires de phrases ont été annotées par dix annotateurs[6].

L'annotation en similarité est une tâche largement différente de celle en émotion ou opinion, ainsi que de celle qui concerne la coréférence : elle nous semblait donc importante pour augmenter la généricité de notre étude. Pour adopter les données à notre expérimentation, nous avons défini des seuils dans l'échelle des annotations permettant un partage des paires de phrases annotées en trois et en cinq classes[7].

4.3. *Données artificielles*

Notre ambition étant d'étudier une correspondance entre toute valeur d'accord interannotateur et la stabilité de la référence suivant le groupe d'annotateurs choisi, il était indispensable de disposer de données liées à différentes valeurs de cet accord, en l'occurrence κ. La seule façon réaliste d'y parvenir est la génération de données fictives. En même temps, pour garder la possibilité d'étudier l'influence de la tâche, il convient, dans ces données artificiellement générées, de préserver ce qui la caractérise, particulièrement la distribution des annotations initiales.

6. Pour plus de précisions, se reporter à (Vu *et al.*, 2015).

7. Dans le cas de trois classes, les intervalles d'annotations sont $[0\ ;\ 1[$, $[1\ ;\ 2,5[$ et $[2,5\ ;\ 4]$; pour cinq classes, les intervalles sont $[0\ ;\ 0,5]$, $[0,6\ ;\ 1,5]$, $[1,6\ ;\ 2,5]$, $[2,6\ ;\ 3,2]$ et $[3,3\ ;\ 4]$.

4.3.1. *Génération d'annotations fictives*

Le problème consiste donc à générer des annotations fictives qui permettent de modifier les accords interannotateurs tout en respectant la répartition des désaccords dans les données réelles. La méthodologie adoptée est la suivante.

Supposons que l'on ait les données réelles de P annotateurs a_i, $i \in \{1 : P\}$ qui aient donné chacun leur avis sur N observables o_j, $j \in \{1 : N\}$.

Pour chaque observable o_j, soit r_j la référence obtenue par vote majoritaire sur cet observable. On pose $e_{ij} = 0$ si l'observateur a_i a voté pour la référence r_j concernant l'observable o_j et $e_{ij} = 1$ sinon. Suivant cette définition, le nombre e_j des « erreurs » commises par les annotateurs sur l'observable o_j s'obtient par $e_j = \sum_{i=1}^{i=P} e_{ij}$ et f_j, la fréquence de ces « erreurs » se définit par $f_j = \frac{e_j}{P}$.

Par ailleurs, on définit Meo le nombre moyen d'erreurs par annotateur :

$$Meo = \frac{\sum_{i=1}^{i=P} \sum_{j=1}^{j=N} e_{ij}}{P} = \frac{\sum_{j=1}^{j=N} e_j}{P} = \sum_{j=1}^{j=N} f_j$$

Pour générer des annotations fictives, on crée des groupes d'annotateurs plus ou moins « bons », et donc on fait varier le paramètre κ en jouant sur le nombre moyen d'erreurs commises par les annotateurs. Par ailleurs, les paramètres f_j permettent de respecter la répartition des désaccords entre les observables. L'utilisation simultanée de ces deux paramétrages est assurée par le protocole défini ci-dessous.

1) On choisit, en fonction du désaccord que l'on veut obtenir entre les k annotateurs que l'on veut créer, un intervalle $[M - A,\ M + A]$ dans lequel on va faire varier le nombre d'erreurs commises par chacun d'entre eux. Dans la pratique, le centre M de l'intervalle doit être inférieur à Meo, si l'on désire améliorer l'accord obtenu dans les annotations réelles, et supérieur dans le cas contraire. Par ailleurs, le choix de l'amplitude A permet de jouer sur la dispersion du nombre d'erreurs entre les différents annotateurs fictifs créés.

2) Pour chaque annotateur fictif créé, on tire au sort le nombre d'erreurs (nbe) qu'il va commettre dans l'intervalle choisi précédemment $[M - A,\ M + A]$: on crée ainsi un annotateur plus ou moins « bon », à l'intérieur des limites que l'on s'est fixées.

3) On effectue un tirage aléatoire des nbe observables pour lesquels l'annotateur va choisir une annotation autre que la référence. Ce tirage est pondéré par les fréquences $f_j,\ j \in \{1...N\}$ des erreurs sur les N observables présentes dans les annotations réelles. On respecte ainsi la répartition des désaccords dans les données initiales.

On obtient donc ainsi un groupe fictif G de k annotateurs, qui ont chacun un nombre d'erreurs nbe compris entre $M - A$ et $M + A$, et dont les erreurs respectives par rapport à la référence se répartissent préférentiellement sur les observables ayant fait l'objet de désaccords dans les annotations réelles.

5. Résultats expérimentaux

5.1. *Validation de l'idée sur données réelles*

La première expérience que nous avons conduite a consisté à étudier sur des annotations réelles la pertinence du principe de correspondance entre la valeur de κ et le taux de variabilité de la référence. Nous avons pour cela travaillé sur les corpus annotés en émotion et en opinion. Partant de ces ressources annotées par 25 personnes, nous avons créé, pour multiplier les observations, trois sous-corpus obtenus par vote majoritaire en conservant à chaque fois $P = 9$ annotations (respectivement les annotations [1, 9], [9, 17] et [17, 25][8]). Nous avons alors appliqué notre méthode d'estimation du taux de variabilité pour plusieurs valeurs de k annotateurs. Dans cette section, nous présentons à titre illustratif les résultats obtenus pour $k = 3$ à partir des corpus annotés avec trois classes d'émotion ou opinion. Des résultats équivalents sont observés avec d'autres valeurs de k et avec cinq classes d'émotion.

Nos observations (cf. tableau de la figure 2) couvrent une large amplitude de valeurs de κ comprises entre 0,25 et 0,66. On observe que ces variations ont un impact sensible sur l'estimation de la probabilité de reproductibilité de l'annotation, qui varie de son côté de 78 % à 93 %. Ce résultat confirme notre intuition, mais surtout, la figure 2 suggère l'existence d'une forte corrélation entre les valeurs d'accord et la probabilité de variabilité. Il semble donc *a priori* possible d'interpréter les valeurs de κ en termes de reproductibilité de l'annotation, ce qui intéresse directement le concepteur d'un corpus annoté. Il est toutefois nécessaire de disposer de jeux de données en plus grand nombre pour confirmer ces premières observations. C'est l'objectif des expérimentations que nous allons présenter, qui ont été réalisées sur les données simulées.

5.2. *Validation de l'idée sur données simulées*

Nous avons reproduit nos expériences en simulant pour chaque corpus 1 000 nouvelles annotations avec notre procédure de génération de données fictives à partir d'annotations réelles. Cette procédure a été renouvelée pour un nombre d'annotateurs variant entre 2 et 8. La figure 3 présente les résultats ainsi obtenus à partir du corpus annoté avec cinq classes d'opinion. Une fois encore, notons que des résultats équivalents ont été obtenus à partir des autres corpus.

On retrouve ici l'existence d'une forte corrélation négative entre les valeurs d'accord et l'estimation de la probabilité de variabilité (et donc une corrélation positive avec la probabilité de reproductibilité), mais validée cette fois sur un nombre très significatif de jeux de données.

8. Nous utilisons deux fois les annotations 9 et 17 pour une raison purement pratique : obtenir trois groupes d'annotateurs à partir de vingt-cinq. Ce découpage n'induit pas de biais sur l'accord puisque nous ne conservons aucune paire d'annotateurs entre chaque groupe d'annotateurs.

Émotion	HC1	HC2	HC3	EC1	EC2	EC3
Taux variabilité	0,22	0,198	0,206	0,217	0,178	0,164
κ_3	0,248	0,323	0,335	0,273	0,359	0,423
Opinion	HC1	HC2	HC3	EC1	EC2	EC3
Taux variabilité	0,119	0,100	0,109	0,099	0,099	0,087
κ_3	0,557	0,598	0,6	0,571	0,611	0,655

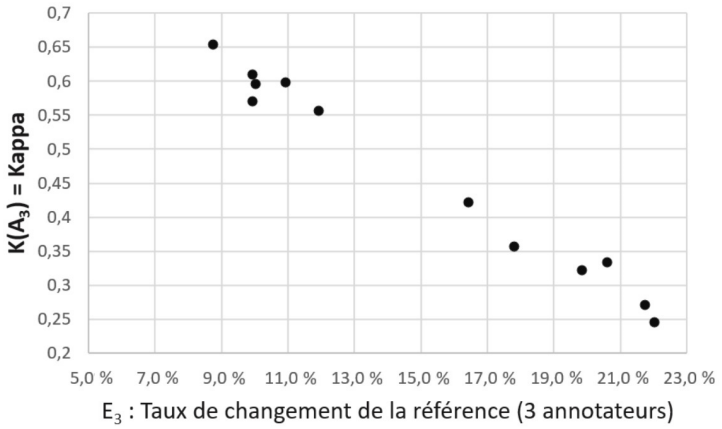

Figure 2. *Résultats sur les corpus en émotion et en opinion à trois classes hors (HC) et en contexte (EC), avec trois annotateurs : valeurs brutes du κ et du taux de modifications de la référence (tableau) et leur représentation graphique*

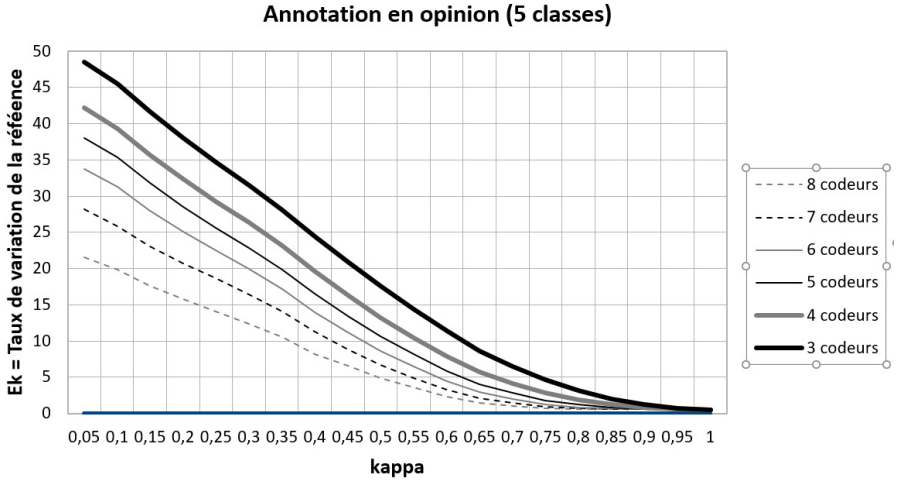

Figure 3. *κ_5 et estimation du taux E_k de variabilité de la référence, pour des données simulées avec k = 2 à 8 annotateurs à partir de l'annotation en cinq classes d'opinion*

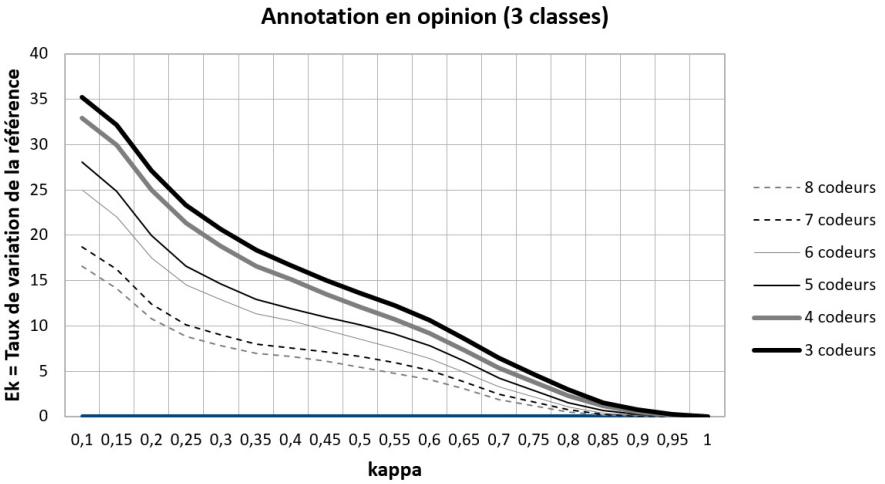

Figure 4. *κ_3 estimation du taux E_k de variabilité de la référence, pour des données simulées avec k = 2 à 8 annotateurs à partir de l'annotation en trois classes d'opinion*

On remarque par ailleurs que cette relation de corrélation entre accord et reproductibilité dépend du nombre d'annotateurs considéré. Pour une valeur de κ donnée, la probabilité de variation de la référence est en effet bien plus élevée si celle-ci a été obtenue avec deux annotateurs qu'avec huit. Si l'on considère par exemple un seuil de fiabilité de données de 0,8 pour le κ, la référence est susceptible de changer dans 7 % des cas avec deux annotateurs, alors que cette probabilité est inférieure à 1 % pour huit annotateurs.

Cette influence du nombre d'annotateurs est facilement interprétable. Il est en effet plus difficile de faire bouger une annotation obtenue par vote majoritaire si celui-ci a été obtenu avec un nombre d'annotateurs plus élevé. Comme nous allons le voir, d'autres facteurs peuvent avoir un impact sur le niveau de reproductibilité.

5.3. *Généricité : influence du nombre de catégories*

Nous avons vu que le nombre de catégories d'annotation pouvait influencer les mesures d'accord observées en corpus (Brenner et Kliebsch, 1996). La comparaison des résultats obtenus sur des jeux de données générés à partir d'annotations à cinq catégories et ceux générés sur trois catégories va nous renseigner sur l'influence du nombre de catégories sur nos observations. À titre d'exemple, la figure 4 donne les résultats obtenus à partir du corpus annoté en opinion, mais cette fois avec seulement trois classes.

On note tout d'abord que l'on retrouve ici la même corrélation négative entre les valeurs d'accord et le taux de variabilité, et ce, pour tous les nombres d'annotateurs considérés. En revanche, les courbes présentées sur la figure 4 se caractérisent par des taux de variation de la référence significativement moindres que sur la figure 3, correspondant à une annotation à cinq catégories.

Ce résultat s'explique simplement pour une annotation par vote majoritaire. Considérons un observable annoté donné dans le corpus. Le nombre de votes que reçoit chaque catégorie est *a priori* plus important dans une annotation à nombre réduit de catégories, puisqu'alors les choix d'annotation se distribuent entre un nombre moindre d'annotations. Dès lors, il est plus difficile de faire bouger une majorité portant sur un plus grand nombre de votes, d'où une variabilité plus faible. On en conclut donc que les tables de correspondance que nous souhaitons établir doivent également considérer le nombre de catégories d'annotation.

Nous cherchons d'ailleurs à mieux caractériser cet impact, et voir si une estimation théorique d'une modification de la référence, par hasard, en fonction du nombre de classes d'annotation ne pourrait pas rapprocher les observations. D'un point de vue pratique, il nous semble toutefois important de donner une estimation brute de la reproductibilité d'une annotation aux concepteurs de corpus.

5.4. *Généricité : influence de la tâche*

Les résultats que nous avons présentés jusqu'ici convergent tous vers le fait qu'il semble possible de proposer une interprétation des valeurs de κ sous forme de stabilité de la référence. D'un point de vue pratique, la question se pose toutefois de savoir s'il est possible de définir une échelle unique d'interprétation des mesures d'accord. Nous avons déjà observé qu'une telle échelle ne peut s'entendre que pour un nombre d'annotateurs et de catégories donné. Mais la question la plus importante reste de savoir si la tâche d'annotation donnée influe, elle aussi, sur l'interprétation. Nos corpus d'expérimentations recouvrent une diversité significative de tâches et d'objets linguistiques (émotions, opinions, coréférences, similarités sémantiques). Pour répondre à cette interrogation, nous avons donc décidé de comparer les résultats spécifiques à chaque tâche à un nombre d'annotateurs et de catégories constant, ceci en considérant toujours des corpus simulés à partir des corpus réels.

La figure 5 donne l'ensemble des observations relevées sur nos corpus pour trois annotateurs, et respectivement pour trois et cinq classes d'annotation. On observe une remarquable convergence des courbes entre tous les corpus. Cela suggère la possibilité de définir des intervalles de confiance (en termes de reproductibilité) en fonction du κ qui soient génériques, c'est-à-dire qui ne dépendent pas de la tâche d'annotation considérée.

Ce premier constat doit toutefois être tempéré. Une étude plus attentive montre en effet que les courbes correspondant aux corpus annotés en opinion (courbes pleines) présentent un comportement légèrement différent. Cette différence de comportement

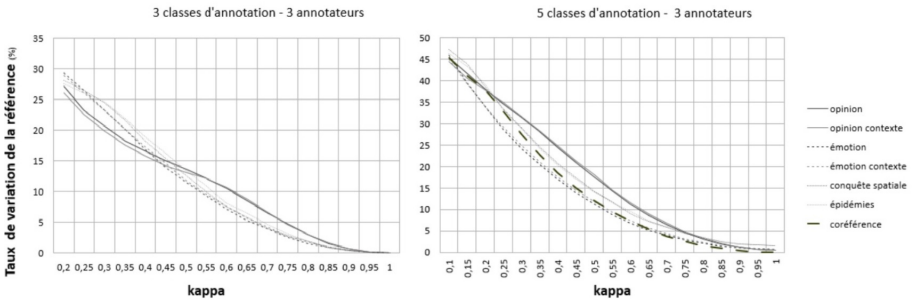

Figure 5. *Comparaison du κ_3 et du taux E_3 de variation de la référence sur tous les corpus, pour des données simulées avec trois annotateurs et trois ou cinq catégories d'annotation*

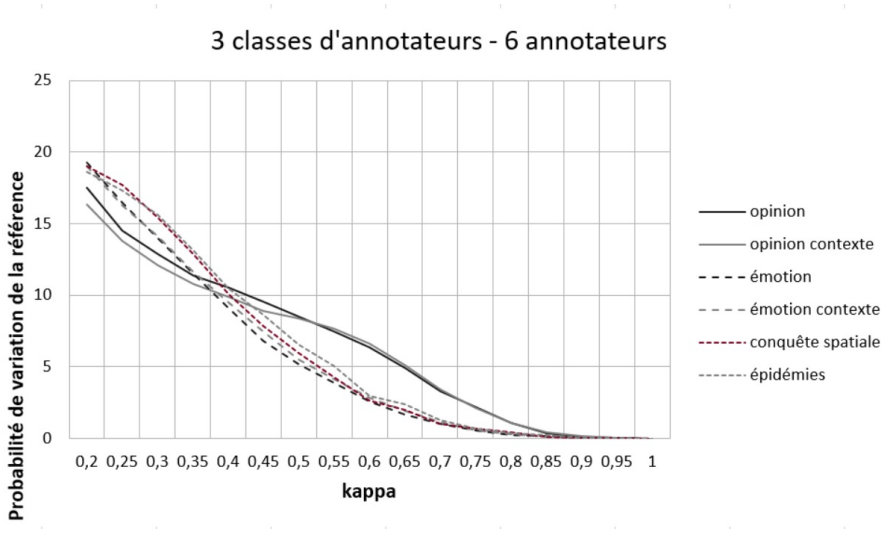

Figure 6. *Comparaison du κ_3 et du taux E_6 de variation de la référence sur tous les corpus, pour des données simulées avec six annotateurs et trois classes d'annotation*

s'accentue lorsque l'on considère un nombre croissant d'annotateurs, comme sur la figure 6. Si les courbes correspondant aux corpus en émotion et en similarité sémantique (épidémie, conquête spatiale) présentent encore une remarquable proximité, la spécificité du corpus en opinion est ici encore plus sensible pour des valeurs de κ intermédiaires : pour les valeurs de κ comprises entre 0,5 et 0,75, la courbe correspondant au corpus en opinion se distingue très sensiblement des autres.

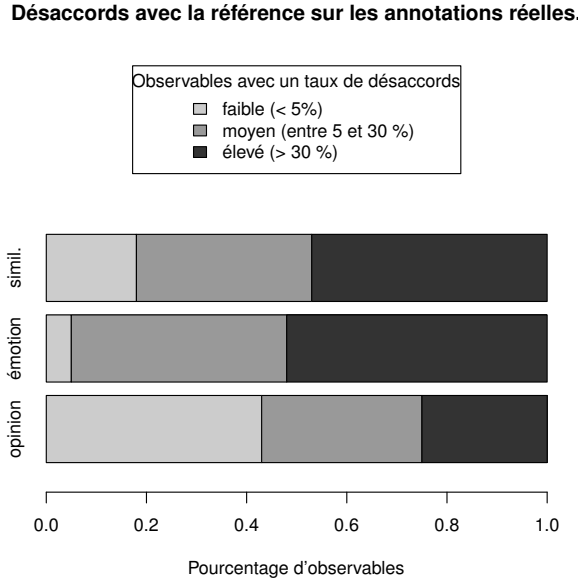

Figure 7. *Histogrammes comparés des désaccords avec la référence sur les corpus émotion, opinion et conquête spatiale*

La nature de l'annotation ne semble pas pouvoir expliquer en elle-même ces divergences de comportement. Comme nous l'avons noté lors de la présentation des corpus (cf. 4.2.2), l'annotation en opinion est en effet proche de l'annotation en émotion dans ses principes, et les deux corpus ont été réalisés dans des conditions identiques, alors que leurs courbes sont différentes. Au contraire, le corpus en émotion donne une courbe proche de celle des annotations en similarité sémantique, alors que ces tâches n'ont rien en commun.

Une analyse des corpus semble suggérer une explication de nature statistique. L'annotation en opinion se caractérise en effet, par rapport aux autres corpus, par une distribution différente des « erreurs » d'annotation par rapport à la référence. Comme on peut le voir dans l'histogramme de la figure 7, le corpus en opinion se caractérise, en effet, par un pourcentage important d'observables (supérieur à 40 %) pour lesquels il y a une quasi–unanimité des annotateurs.

Il reste à élaborer un modèle mathématique qui rendrait compte exactement de l'impact de cette variation de la distribution des divergences d'annotation entre observables. Mais le fait que les autres corpus conduisent à des courbes d'évolution proches, suggère que notre démarche est générique en termes de tâche d'annotation. La section

suivante cherche à éclaircir l'impact de cette distribution non homogène des écarts à la référence.

6. Étude de l'impact de la distribution des divergences d'annotation

Selon nos observations, la distribution des divergences d'annotation entre les observables pourrait être un facteur qui modifie la correspondance entre les valeurs de κ et la stabilité de la référence des différents groupes d'annotateurs. Les résultats présentés dans cette section ont été réalisés sur des annotations entièrement fictives spécialement créées pour confirmer ou infirmer cette observation et pour pouvoir, ultérieurement, explorer d'autres paramètres fréquemment cités, tels que la prévalence.

6.1. *Création des données fictives*

On veut contrôler le nombre de classes (*C*), le nombre d'annotateurs dans chaque groupe créé (*k*), l'accord entre les annotateurs de chaque groupe (le κ), la distribution du taux de désaccords entre les annotations et la référence produite, ainsi que les prévalences entre classes. En l'absence de référence réelle, les taux de variation de la référence seront calculés entre les paires de groupes d'annotateurs créés.

Le protocole adopté est très proche de celui présenté dans la section 4.3.1. Il en diffère cependant sur les points suivants :

– en l'absence de données réelles, une référence initiale des annotations est définie aléatoirement sur les observables[9] ;

– à côté du paramètre M qui gère le nombre d'observables pour lequel un annotateur est en désaccord avec la référence, on introduit un nouveau paramètre : son écart-type σ. Les nombres de désaccords pour chacun des k annotateurs d'un groupe sont tirés aléatoirement suivant une distribution normale de paramètres $(M,\ \sigma)$.

Malgré sa simplicité, ce protocole permet de créer un ensemble de n groupes de k annotateurs ayant un κ relativement stable. Par exemple, on peut constater la faible valeur des écarts-types dans le tableau 1 qui donne la moyenne et l'écart-type des valeurs de κ avec $k = 4$ pour $n = 200$ groupes d'annotateurs créés suivant ce procédé. Par ailleurs, si la référence initiale est un artifice efficace pour constituer des groupes d'annotateurs de κ homogène, elle n'intervient pas dans les calculs qui s'ensuivent.

6.2. *Expérimentations et résultats*

Les expérimentations actuellement réalisées concernent essentiellement la distribution des désaccords sur les observables. Les premiers tests concernant la prévalence

9. Pour contrôler la prévalence, on définit le poids de chaque classe ; bien sûr, la référence initiale est créée en respectant ces poids.

κ :	0,247	0,319	0,398	0,484	0,580	0,684	0,799	0,923
Écart-type κ :	0,029	0,028	0,025	0,022	0,020	0,016	0,015	0,010

Tableau 1. *Moyenne et écart-type des valeurs de κ calculées sur 200 groupes de quatre annotateurs*

ne seront pas présentés : ils semblent indiquer une relation complexe entre prévalence et stabilité de la référence qui demande une étude approfondie.

La figure 8 permet de comparer les résultats obtenus dans le cas où les désaccords entre annotateurs sont répartis uniformément sur l'ensemble des observables, avec celui où 20 % des observables donnent lieu à un accord total, les désaccords étant uniformément répartis sur les 80 % restants. On y observe que la courbe qui correspond à une distribution uniforme est, pour un nombre donné d'annotateurs, en dessous de la courbe correspondante où 20 % des observables font l'unanimité. Il apparaît donc nettement qu'effectivement, la distribution des désaccords entre les observables est un paramètre important de la relation entre la valeur de κ et la stabilité de la référence produite par les annotateurs. De façon prévisible, un report des désaccords sur un plus petit nombre d'observables induit une plus grande instabilité de la référence.

7. Conclusion

Les expérimentations que nous avons présentées dans cet article ont cherché à interpréter les valeurs d'accord interannotateur (ici, le κ de Cohen) sous la forme d'une probabilité de reproductibilité de l'annotation. Cette étude a été menée sur des corpus réels relevant de tâches d'annotation variées, puis a été complétée sur des données d'envergure simulées à partir de ces corpus réels. On peut donc espérer que sa portée est suffisamment large pour nous permettre de tirer certaines conclusions génériques sur la question.

Les résultats que nous avons détaillés montrent qu'il existe une corrélation forte entre la mesure d'accord interannotateur liée à une annotation obtenue par vote majoritaire et la probabilité de reproduire la même annotation de référence avec un autre ensemble d'annotateurs. La nature exacte de cette corrélation reste encore à préciser, une fois mieux modélisée l'influence de facteurs tels que le nombre de classes d'annotation, le nombre de codeurs et surtout la distribution des divergences d'annotation suivant les observables. Il nous semble toutefois que ces travaux sont une piste encourageante vers une interprétation des valeurs d'accord interannotateur utiles au concepteur de corpus annotés. Quelques conclusions prudentes peuvent à ce sujet déjà être tirées de cette étude.

Un κ de 0,8 est une valeur seuil acceptée par tous les auteurs comme gage de bonne fiabilité des données annotées. Nos expériences confirment ces opinions objectives, puisqu'elles se traduisent par une probabilité de variation de l'annotation relativement

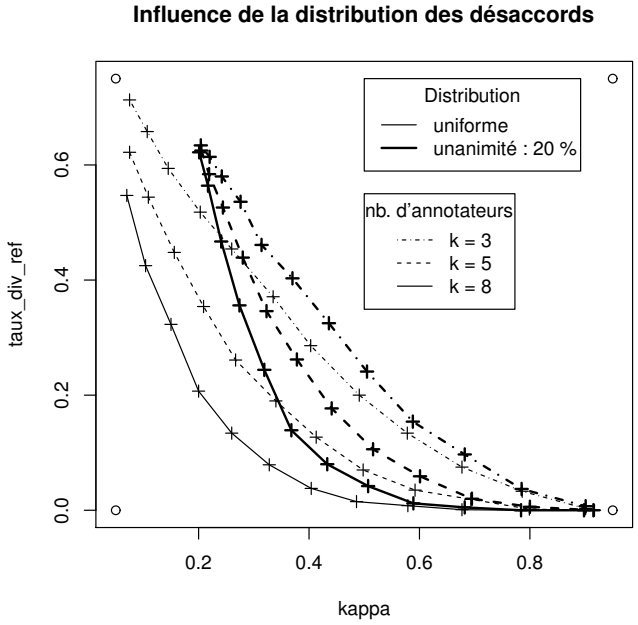

Figure 8. *Résultats comparés : répartition uniforme des désaccords contre 20 % des observables faisant l'unanimité*

faible. Ainsi, toutes nos expériences avec trois annotateurs ou plus montrent que, pour un κ de 0,8, l'annotation de référence a toujours moins de 3 % de chances d'être modifiée avec un autre ensemble d'annotateurs.

À l'opposé, une valeur de κ de 0,67 semble constituer une garantie de fiabilité plus modeste : pour trois annotateurs (figure 5 par exemple) la probabilité de variation de l'annotation de référence est en effet comprise entre 5 % et 10 % cette fois. De tels seuils de fiabilité peuvent déjà nous interroger. Ils posent également question pour ce qui concerne la significativité statistique des campagnes d'évaluation menées en TAL : quel crédit donner aux résultats d'une telle évaluation, si l'annotation de référence qui a servi à l'apprentissage ou au test est susceptible de varier de 10 % avec un autre ensemble d'annotateurs ?

Forts de ces résultats encourageants, nous envisageons d'étendre cette étude en intégrant de nouvelles tâches d'annotation, de nouvelles métriques d'accord telles que le α de Krippendorff. Enfin et surtout, nous aimerions mieux caractériser l'influence de la distribution des annotations de la référence sur les mesures de taux de reproductibilité, afin d'arriver à une interprétation directe et générique des valeurs d'accord.

Nous tenons enfin nos corpus librement à disposition de toute personne qui aimerait reproduire ou compléter cette étude, sur demande auprès des auteurs.

Remerciements

Les auteurs remercient la fédération ICVL (Informatique Centre Val de Loire) pour son soutien à cette recherche, dans le cadre du financement du stage de Dany Brégeon.

8. Bibliographie

Antoine J.-Y., Villaneau J., Lefeuvre A., « Weighted Krippendorff's alpha is a more reliable metrics for multi- coders ordinal annotations : experimental studies on emotion, opinion and coreference annotation. », *EACL 2014*, Gotenborg, Sweden, April, 2014. `http://www.aclweb.org/anthology/E14-1058`.

Artstein R., Poesio M., « Bias decreases in proportion to the number of annotators. », *Proceedings FG-MoL'2005*, Edinburgh, UK, p. 141-150, 2005.

Artstein R., Poesio M., « Inter-coder Agreement for Computational Linguistics », *Computational Linguistics*, vol. 34, n° 4, p. 555-596, December, 2008.

Brennan P., Silman A., « Statistical methods for assessing observer variability in clinical measures. », *BMJ*, vol. 304, p. 1491-1494, 1992.

Brenner H., Kliebsch U., « Dependance of weighted kappa coefficients on the number of categories. », *Epidemiology*, vol. 7, p. 199-202, 1996.

Byrt T., Bishop J., Carlin J., « Bias, prevalence and kappa. », *Journal of Clinical Epidemiology*, vol. 46, p. 423-429, 1993.

Carletta J., « Assessing agreement on classification tasks : the Kappa statistic. », *Computational Linguistics*, vol. 22, n° 2, p. 249-254, 1996.

Cohen J., « A coefficient of agreement for nominal scales. », *Educational and Psychological Measurement*, vol. 20, p. 37-46, 1960.

Cohen J., « Weighted kappa : nominal scale agreement with provision for scaled disagreement of partial credit. », *Psychological Bulletin*, vol. 70, p. 213-220, 1968.

Davies M., Fleiss J., « Measuring agreement for multinomial data. », *Biometrics*, vol. 38, p. 1047-1051, 1982.

Di Eugenio B., Glass M., « The Kappa Statistic : A Second Look », *Computational Linguistics*, vol. 30, n° 1, p. 95-101, 2004.

Feinstein A., Cicchetti D., « High agreement but low Kappa : the problem of two paradoxes. », *Journal of Clinical Epidemiology*, vol. 43, p. 543-549, 1990.

Krippendorff K., « Reliability in content analysis : Some common misconceptions and recommendations. », *Human Communication Research*, vol. 30, n° 3, p. 411–433, 2004.

Krippendorff K., *The content analysis reader.*, Sage, Beverly Hills, CA, chapter Testing the reliability of content analysis data : what is involved and why., 2008.

Krippendorff K., *Content Analysis : An Introduction to Its Methodology*, Sage, Beverly Hills, CA, chapter 11, 2013.

Landis J., Koch G., « The measurement of observer agreement for categorical data. », *Biometrics*, vol. 33, p. 159-174, 1977.

Le Tallec M., Villaneau J., Antoine J.-Y., Duhaut D., « Affective Interaction with a Companion Robot for vulnerable Children : a Linguistically based Model for Emotion Detection », *Proceedings of LTC'2011, Language Technology Conference*, Poznan, Poland, p. 445-450, 2011.

Mathet Y., A contribution to Computational Linguistics and Natural Language Processing : From the Semantics of Space and Time to Annotations and Agreement Measures, Habilitation à diriger des recherches, Université de Caen Normandie, 2017a.

Mathet Y., « The Agreement Measure γcat a Complement to γ Focused on Categorization of a Continuum. », *Computational Linguistics*, vol. 43, n° 3, p. 0661-0681, September, 2017b.

Muzerelle J., Lefeuvre A., Schang E., Antoine J.-Y., Pelletier A., Maurel D., Eshkol I., Villaneau J., « ANCOR_Centre, a Large Free Spoken French Coreference Corpus : description of the Resource and Reliability Measures », *in* ELRA (ed.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reyjavik, Iceland, May, 2014. `https://hal.archives-ouvertes.fr/hal-01075679`.

Neuendorf K., *The content analysis guidebook.*, Sage, Thousand Oaks, CA, 2002.

Schuller B., Steidl S., Batliner A., « The Interspeech'2009 emotion challenge. », *Proceedings Interspeech'2009*, Brighton, UK, p. 312-315, 2009.

Sim J., Wright C., « The Kappa Statistic in Reliability Studies : Use, Interpretation, and Sample Size Requirements. », *Physical Therapy*, vol. 85, n° 3, p. 257-268, 2005.

Vanderheyden K., « Le Noel des animaux de la montagne. », , `http://www.momes.net/histoiresillustrees/contesdemontagne/noelanimaux.html`, 1995.

Vassallo R.-M., *Comment le Grand Nord découvrit l'été.*, Flammarion, Paris, France, 2004.

Vu H.-H., Villaneau J., Saïd F., Marteau P.-F., « Mesurer la similarité entre phrases grâce à Wikipédia en utilisant une indexation aléatoire », *TALN 2015*, Caen, France, 2015.