Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Évaluation de mesures d'accord sur des structures relationnelles par la dégradation contrôlée d'annotations

Antoine Boiteau[1]
(1) Normandie Université, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, FRANCE
antoine.boiteau@unicaen.fr

RÉSUMÉ

Les mesures d'accord inter-annotateurs sont essentielles pour évaluer la qualité des annotations humaines sur les corpus. Dans le cadre des structures relationnelles, la question de la qualité et de l'interprétabilité de ces mesures reste cependant ouverte. Cet article présente l'adaptation d'un outil déjà utilisé pour d'autres paradigmes d'annotation dont le but est de générer de manière contrôlée des annotations artificielles erronées. Les annotations obtenues sont fournies à des mesures d'accord adaptées aux structures relationnelles, permettant l'identification des comportements des mesures ainsi que les différences entre elles.

ABSTRACT

Controlled degradations on annotations for relational structures : utility and method

Inter-coder agreement measures are essential for assessing the quality of human annotations on corpora. In the context of relational structures, however, the question of the quality and interpretability of these measures remains open. This article presents the adaptation of a tool already used for other annotation paradigms. Its aim is to generate erroneous artificial annotations in a controlled manner. The resulting annotations are provided to agreement measures adapted to relational structures, allowing the identification of the behaviour of the measures as well as the differences between them.

MOTS-CLÉS : annotation de structures relationnelles, mesures d'accord inter-annotateurs, analyse de l'argumentation, évaluation.

KEYWORDS: annotation of relational structures, inter-coder agreement, argumentation analysis, evaluation.

1 Introduction

Les corpus textuels sont des jeux de données précieux pour la recherche scientifique. Déjà indispensables pour toute démarche expérimentale dans le domaine du TAL, leur intérêt a été encore grandissant ces dernières décennies avec les approches d'apprentissage machine qui ont révolutionné nos disciplines. Pour que ces jeux de données soient utilisables pour l'entraînement des systèmes d'apprentissage supervisés et semi-supervisés, et comme référence pour évaluer les performances de ces systèmes (Fort, 2016), il est nécessaire qu'ils fassent l'objet d'une étape supplémentaire d'annotation. Le procédé d'annotation consiste en l'identification et la mise en avant du phénomène recherché par la communauté. Sur ces corpus plusieurs types d'annotation, éventuellement complémentaires, sont possibles. Selon (Leech *et al.*, 1997), l'annotation peut être vue comme une information interprétative rajoutée à un jeu de données brutes, ou déjà précédemment annotées. Une annotation de référence est

Actes de la 26ème Rencontre des Étudiants Chercheurs en Informatique
pour le Traitement Automatique des Langues, pages 57–68.

un jeu de données enrichi de connaissances que l'on juge fiables (ce jeu de données constitue alors une référence ou *gold standard*). La littérature présente de nombreux exemples de campagnes d'annotation sur des corpus textuels ou oraux retranscrits afin de mettre en lumière les phénomènes linguistiques. Si l'annotation de référence peut parfois s'obtenir au moyen d'un unique expert, dans de nombreux cas, notamment pour les tâches inédites ou difficiles, la méthode classique consiste à procéder à une annotation multiple du jeu de données. L'hypothèse sous-jacente à cette annotation multiple est que plus l'accord entre les annotateurs est important et plus l'annotation se rapproche de la vérité. Pour l'annotation multiple, différents annotateurs produisent chacun leur propre annotation, indépendante des autres. Les annotateurs peuvent lors de cette étape avoir des productions différentes. Une mesure de l'accord inter-annotateurs adaptée à la tâche d'annotation permet alors d'évaluer le degré d'accord de la multi-annotation. La qualité et le sens des mesures d'accord est une problématique étudiée depuis plusieurs décennies par la communauté. Ce questionnement s'illustre par la diversité des mesures d'accord inter-annotateurs, nombreuses pour la tâche de catégorisation (où l'annotateur doit typer des unités déjà identifiées), plus rares pour la tâche d'identification d'unités (où l'annotateur doit de surcroît identifier et positionner les unités dans le *continuum* textuel), tâche souvent dite d'*unitizing* (Krippendorff, 2019). Bien que l'état de l'art soit riche pour ces deux types de tâches d'annotation, les travaux sur l'accord pour d'autres tâches d'annotation telles que les structures relationnelles sont plus rares. C'est dans ce contexte que nos présents travaux se positionnent. Nous préciserons tout d'abord le cadre de l'annotation des structures relationnelles dans les corpus textuels. Nous reprendrons à notre compte le principe du Corpus Shuffling Tool (Mathet *et al.*, 2012), dédié initialement à l'identification d'unités, en l'appliquant à ce type de structures. Cet outil permet en effet d'évaluer et de mieux comprendre les comportements des différentes mesures et de pouvoir les comparer entre elles. Enfin, nous présenterons nos résultats expérimentaux suite à la mise à l'épreuve de mesures d'accord adaptées aux structures relationnelles grâce à notre outil de dégradation contrôlée.

2 Cadre des structures relationnelles

Nous nous intéressons ici à des structures relationnelles, c'est-à-dire des configurations au sein desquelles deux ou plusieurs unités d'intérêt sont mises en relation. Le TAL et la linguistique offrent de nombreux exemples de telles structures, parmi lesquelles on peut citer par exemple les structures syntaxiques, les structures rhétoriques ou encore les structures argumentatives.

2.1 Exemple de structure relationnelle

Par la suite, nous illustrerons nos propos en nous appuyant sur un modèle de structuration qui nous permettra de mettre en lumière certaines difficultés posées par la comparaison de structures relationnelles : les structures argumentatives telles que décrites par (Putra *et al.*, 2022). Cette modélisation est adaptée aux besoins du corpus ICNALE (Ishikawa, 2013, 2018). Pour l'annotation de cette structure, les unités de bases sont les phrases du corpus. Ce choix de granularité est déjà présent dans d'autres études de l'argumentation (Teufel *et al.*, 1999; Wachsmuth *et al.*, 2016). L'annotation de la structure se fait alors en deux étapes. Vient en premier lieu l'*identification des composants de l'argumentation*. Il s'agit ici de catégoriser chaque unité comme étant Composant de l'Argumentation (CA) ou détachée de la structure argumentative (non-CA). La seconde étape est celle de l'*identification de la structure argumentative* qui consiste à identifier et positionner les relations existant entre les CA. Dans ce

cas d'étude, les relations sont toutes étiquetées par l'un des trois types de relation binaire et dirigée proposés : *soutien* (*support*), *attaque* (*attack*) et *détail* (*detail*). En plus de ces trois types fréquemment rencontrés dans la littérature, sous ces appellations ou d'autres (Kirschner *et al.*, 2015), les auteurs ajoutent la relation de type *reformulation* (*restatement*) qui est bidirectionnelle et signale qu'un CA vient reprendre l'argument d'un autre CA pour le réitérer et l'évoquer stratégiquement à un moment différent du corpus (Skeppstedt *et al.*, 2018).

Dernière caractéristique de la structure de (Putra *et al.*, 2022), chaque unité CA a toujours une et unique relation sortante mais peut-être la cible de plusieurs relations entrantes, sauf pour le cas particulier des CA de type *proposition principale* (*main claim*) qui n'ont aucune relation sortante et désignent l'argument principal défendu ou attaqué par un ensemble de CA, voire par l'ensemble du texte. Ainsi pour un texte à annoter avec n CA et m CA de type *proposition principale*, il y a exactement $n - m$ relations dans l'annotation. Les structures d'argumentation résultantes peuvent être considérées comme des graphes (Park & Cardie, 2018). La structure sur laquelle nous portons notre attention est plus précisément une forêt d'arbres dont les racines sont les CA de type *proposition principale*.

2.2 Des mesures d'accord inter-annotateurs peu adaptées au cadre relationnelle

Certaines études actuelles continuent d'utiliser des mesures qui n'ont pas été spécifiquement pensées pour des structures de cet ordre, au risque de générer des biais. (Putra *et al.*, 2022), avant de présenter des mesures adaptées à des structures argumentatives, illustrent ce phénomène en reformulant le résultat de l'annotation par trois niveaux de catégorisation dans le but d'appliquer des mesures prévues pour ce paradigme d'annotation :

— la catégorisation binaire de chaque unité du texte, en l'espèce les phrases du corpus, comme étant, ou non, un composant de l'argumentation ;
— la catégorisation binaire de chaque paire d'unités non-identiques a et b telle qu'il existe, ou non, une relation entre a et b ;
— la catégorisation par l'étiquetage d'un type de relation pour chaque paire d'unités a et b où l'annotateur a annoté l'existence d'une relation entre a et b.

Cette approche permet aux auteurs d'utiliser deux mesures très connues pour la catégorisation, l'accord observé et le κ de Cohen (Cohen, 1960).

(Kirschner *et al.*, 2015) identifie certains inconvénients à l'utilisation de ces mesures dans le cadre des structures relationnelles : l'accord observé et le κ fonctionnent correctement lorsque l'annotation porte sur des objets indépendants les uns des autres et distribués uniformément dans le corpus ; or dans le cadre des structures relationnelles l'annotation d'un objet peut dépendre de l'annotation d'un autre objet du corpus, ceci pouvant entraîner des variations de la mesure d'accord et des scores erronés. De plus, dans les corpus présentant des structures relationnelles, les cas où les unités sont fortement connectées entre elles par des relations font figure d'exception. Ainsi, pour n unités on observe de l'ordre de n relations alors qu'étiqueter toutes les paires possibles d'unités distinctes revient à en annoter de l'ordre de n^2. Ainsi plus le corpus présente d'unités et plus on peut être amené à considérer un grand nombre de paires d'unités comme non-connectées, ce qui, selon les besoins de l'étude, apporte peu sur la compréhension du phénomène que l'on veut étudier. Ajoutons que plus deux unités sont éloignées dans l'ordre de lecture dans un même corpus et moins les chances qu'une relation entre ces unités existent sont grandes. (Kirschner *et al.*, 2015) propose de pallier cet effet

en pondérant l'importance des paires en fonction de la distance entre les unités de la paire, ainsi qu'en retirant les paires dont les unités sont trop éloignées l'une de l'autre. Ces réserves conduisent donc naturellement à s'interroger sur la pertinence d'encoder l'annotation de toutes les paires en deux catégories, porteuse ou non d'une relation. Cet exemple illustre les biais qui peuvent résulter de l'utilisation forcée d'une mesure pré-existante sans tenir compte des spécificités des objets annotés. Comme pour les autres domaines d'annotation, il apparaît donc nécessaire de disposer de mesures d'accord spécifiquement pensées pour les structures relationnelles. Un cadre formel commun est sous-jacent aux exemples présentés ci-dessus. Ces structures peuvent être représentées et analysées comme des graphes diversement contraints, ces contraintes pouvant s'exprimer sur les propriétés telles que l'orientation des arrêtes, la présence d'une arborescence ou la connectivité du graphe. Pour ce paradigme de représentation sous forme de graphe, il est dès lors nécessaire de disposer de mesures d'accords adaptées aux particularités de ces structures. C'est ce que fait (Kirschner *et al.*, 2015) en présentant ce qui est à notre connaissance la première mesure d'accord inter-annotateurs spécifiquement créée pour les graphes. Nous comparerons cette mesure aux variantes de *Mean Average Recall* (Putra *et al.*, 2022). Nous détaillerons le fonctionnement de ces mesures en 4.3.

3 Principe des dégradations contrôlées

Pour aider à la comparaison des mesures existantes et le cas échéant à l'élaboration de nouvelles mesures, nous proposons un environnement dans lequel on peut étudier le comportement des mesures en les confrontant à des données dont on contrôle la fiabilité.

3.1 Benchmark et interprétabilité des mesures d'accord inter-annotateurs

Nous l'avons vu plus tôt, les mesures d'accord inter-annotateurs sont des outils cruciaux pour l'établissement de corpus annotés de référence. La question de la qualité des mesures que nous employons est ainsi une problématique d'intérêt pour la communauté scientifique. (Artstein & Poesio, 2008) posent aussi la question de l'interprétabilité des mesures et du sens que portent les scores donnés par ces mesures. Prenant l'exemple de la famille des mesures κ, (Mathet *et al.*, 2012) posent quelques questions qui paraissent triviales en apparence mais qui sont loin de l'être : « Un score κ de 0,75 indique-t-il un bon résultat ? Un score κ de 0,8 est-il deux fois meilleur qu'un score de 0,4 ? Un score de 0,6 obtenu avec une première mesure est-il meilleur qu'un score de 0,5 obtenu avec une autre mesure, et pour quelle tâche d'annotation ? ». Pour répondre à ces questions, sur ces mesures et sur d'autres, les auteurs ont proposé un outil nommé le Corpus Shuffling Tool (CST) qui a pour but de comparer l'évolution des scores de différentes mesures d'accord inter-annotateurs face à des multi-annotations générées de manière contrôlée et paramétrable. Ces données d'annotations sont spécifiquement fabriquées par l'outil pour évaluer les réactions des mesures.

Décrivons étape par étape le fonctionnement du CST. Pour fonctionner, le CST nécessite tout d'abord un jeu d'annotations à dégrader. Nous appellerons ce jeu d'annotations *référence* pour la suite de cette section. Il n'est pas nécessaire que ce jeu soit réellement le *gold standard* issu d'une campagne d'annotation, il suffit qu'il soit représentatif des phénomènes qu'on retrouverait dans une telle annotation de référence. Pour démarrer son processus de dégradation, en plus d'une *référence*, le CST a besoin de deux autres paramètres : le nombre n d'annotateurs à simuler et une magnitude m dont la valeur est comprise entre 0 et 1 inclus. Lors de ce processus, le CST va, pour chacun des

n annotateurs, faire une copie de la *référence* et dégrader aléatoirement les annotations de la copie selon la magnitude m. L'ensemble des copies dégradées ainsi générées forment une multi-annotation que nous pouvons fournir aux mesures que l'on souhaite évaluer. Afin d'observer comment les mesures d'accord inter-annotateurs réagissent face à des multi-annotations de plus en plus dégradées, le CST exécute son processus de dégradation en faisant varier incrémentalement son paramètre de magnitude de 0 à 1 d'un pas paramétrable. Pour chaque valeur de la magnitude on obtient ainsi une multi-annotation composée de jeux d'annotations de plus en plus éloignés de la *référence*. Le désaccord entre les annotateurs simulés est donc théoriquement de plus en plus grand. Enfin, on calcule le score des mesures d'accord inter-annotateurs pour chaque multi-annotation et on trace le graphique résultant en prenant la magnitude m en abscisse du graphique et les scores en ordonnées. Pour que l'évaluation des mesures soit possible avec cet outil, il faut que les jeux d'annotations dégradés soient vraisemblables. Pour se faire, (Mathet *et al.*, 2012) présentent 5 types de dégradations différentes dont le but est de générer des erreurs qui sont observées dans les campagnes d'annotation.

3.2 Types de dégradation

Lorsque (Mathet *et al.*, 2012) présentent leur outil, celui-ci est restreint à deux tâches d'annotation : la catégorisation et l'identification d'unités. Cela les mène à présenter des types de dégradations spécifiques à ces tâches comme la fragmentation : le fait de séparer une unité de référence en plusieurs nouvelles unités plus petites. Dans le cadre de la structure relationnelle que nous étudions et que nous avons précisée en 2.1, plusieurs de ces types de dégradations ne sont pas adaptées à notre cas, et d'autres dégradations nouvelles et propres aux structures relationnelles méritent d'être explorées. Contrairement aux précédents travaux sur le CST, les dégradations proposées ici ne sont pas motivés par un recensement des erreurs rencontrées fréquemment dans les campagnes d'annotation. En effet, nos dégradations se reposent sur l'ensemble des paramètres qui définissent une relation de structure relationnelle, c'est-à-dire : l'origine de la relation, sa cible, son orientation et sa catégorie. Les dégradations ainsi proposées sont donc issues de considérations théoriques et peuvent ne pas pouvoir s'appliquer totalement à toutes les modélisations et campagnes d'annotation, ou alors nécessiter des adaptations substantielles pour être utilisées. L'objectif premier de cet article est de démontrer la possibilité de l'utilisation du CST pour des structures relationnelles et de fournir une *boîte d'outils* divers et indépendants de dégradations pour les futurs travaux s'intéressant à la pertinence des mesures d'accord inter-annotateurs employées. Pour illustrer ces outils, nous préciserons ci-après les types de dégradation que nous proposons dans le cadre particulier des structures argumentatives.

3.2.1 Dégradations élémentaires pour les structures argumentatives

La magnitude m est une valeur réelle telle que $0 \leq m \leq 1$ qui reflète un niveau de dégradation. Au niveau 0 aucune dégradation n'est subie. Au niveau 1 la dégradation est maximale, ce qui simule la perte par un annotateur de toute compétence. Sauf dans le cas particulier du faux positif, la magnitude correspond ici à la probabilité pour chaque relation de la *référence* de subir une dégradation.

Changement de cible : L'identification du rattachement d'un argument pour soutenir ou attaquer un autre peut être source d'erreur. Dans le cadre de cette dégradation l'annotateur simulé assigne comme cible de la relation une autre unité. Chaque unité a une probabilité pondérée d'être choisie comme la nouvelle cible. La structure argumentative peut être vue comme un arbre dont la racine est

la *proposition principale*. Les unités qui se trouvent sur le chemin entre la cible originale et la racine (incluse) ont une probabilité importante d'être sélectionnées. Les unités se trouvant entre l'origine de la relation et les feuilles de l'arbre ont un poids moyen. Enfin les autres unités sur cet arbre reçoivent une pondération faible.

Changement d'origine : Cette dégradation utilise le même principe que le changement de cible, mais c'est ici l'origine de la relation qui est modifiée. En considérant le graphe sous forme d'arbre, la distribution des pondérations pour la nouvelle origine est la suivante : les unités entre l'origine et les feuilles ont un poids élevé, les unités entre la cible et la racine (incluse) reçoivent une pondération moyenne et les autres unités ont un poids faible.

Permutation d'orientation : La permutation d'orientation est le fait pour l'annotateur simulé d'identifier la présence d'une relation entre deux unités en se trompant sur le sens de la relation.

Changement d'étiquetage : Le changement d'étiquetage intervient lorsque l'annotateur se trompe sur la catégorie à attribuer à une relation. Ce phénomène est particulièrement fréquent quand deux catégories de relation sont proches ou qu'une catégorie est rare dans le corpus. Lorsqu'il y a changement de catégorie, une nouvelle catégorie est attribuée au hasard tout en tenant compte la fréquence des catégories dans la *référence*.

Faux négatif : Le faux négatif consiste en le retrait d'une annotation présente dans la *référence*.

Faux positif : En miroir du faux négatif, le faux positif est l'ajout d'une annotation erronée absente de la *référence*. Pour a annotations dans la *référence* et une magnitude m, $a * m$ relations sont crées et ajoutées au jeu de données. Pour ne pas dénaturer substantiellement le sens de la *référence*, les nouvelles relations sont crées en reprenant les caractéristiques statistiques de la *référence*.

3.2.2 Combinaison de dégradations

Lors des campagnes d'annotation, les différents types d'erreurs d'annotation qui peuvent intervenir ne s'excluent pas forcément mutuellement. Pour restituer ce phénomène nous prévoyons de rendre possible la combinaison des dégradations précitées lors du processus de génération des multi-annotations. Dans le cas de la combinaison de dégradations, chacun des t types de dégradation sélectionnés est appliqué successivement sur les jeux d'annotations ; soit avec une magnitude t/m si l'on souhaite une distribution uniforme des dégradations, soit avec une pondération visant à reproduire une distribution observée dans des annotations réelles.

4 Expérimentations sur des structures argumentatives

4.1 Jeu d'annotations et instance étudiée

Pour évaluer les métriques de l'accord inter-annotateurs avec le CST il faut en premier lieu lui fournir une *référence annotée* que l'outil viendra dégrader. Afin de nous procurer une structure argumentative annotée que nous pourrions employer à ces fins, nous nous sommes tout d'abord tourné vers le corpus ICNALE (Ishikawa, 2013, 2018) qui est utilisé par (Putra *et al.*, 2022) pour introduire les mesures MAR. Néanmoins, ce corpus propose des structures d'argumentation avec peu d'unités argumentatives puisque nous y retrouvons environ 13,9 phrases par texte argumentatif, dont toutes ne sont pas CA. Or, nous pensons qu'apporter une simple dégradation aléatoire à des objets argumentatifs de si petite taille modifierait déjà substantiellement la structure argumentative et son sens. Pour opérer nos dégradations sur des objets dont le sens global n'est pas totalement altéré après quelques dégradations successives, et ainsi pouvoir comparer nos mesures sans craindre de trop grandes variations entre chaque itération des dégradations, tout en conservant un coût de calcul acceptable, nous proposons comme *référence* du CST une structure argumentative de taille suffisante avec 101 CA et 100 relations, représentée sur la figure 2. La structure argumentative de référence que nous fournissons à notre implémentation du CST représente l'exemple d'une structure que nous attendons d'une argumentation fournie avec de nombreux arguments et contre-arguments dirigés vers un argument principal. Sur l'illustration, les relations sont dépourvues d'étiquettes puisque les catégories des relations n'influent pas sur les mesures que nous évaluons dans notre étude (cf. 4.3). Chaque relation a néanmoins été annotée avec une étiquette lors de l'élaboration par nos soins de cette structure.

4.2 Dégradations retenues pour cette campagne

Dans la section 3.2 nous avons présenté les types de dégradations envisageables pour des objets tels que la structure argumentative. Pour nos expérimentations nous excluons de cette étude trois types de dégradations. Tout d'abord, nous écartons l'*étiquetage des relations* car les mesures que nous étudions ne prennent pas en compte ce paramètre. Nous n'incluons pas non plus le paradigme du *faux positif* car son implémentation concrète dans le cadre décrit par (Putra *et al.*, 2022) nous interroge encore. En effet, nous attendons $n-1$ relations *explicites* pour une structure à n CA ; or les exemples que nous traitons possèdent déjà $n-1$ relations. Nous ne pouvons donc pas rajouter de relations supplémentaires. Enfin, la *combinaison de dégradations* sera pour l'heure mise de côté car nous cherchons ici à étudier les comportements des mesures en fonction de dégradations précises. L'étude de l'impact d'une combinaison de facteurs serait intéressante mais dépasse notre objectif actuel.

4.3 Mesures d'accord inter-annotateurs implémentées

Comme nous l'avons vu, il existe des méthodes pour traduire les annotations relationnelles en plusieurs étapes successives de catégorisation (Putra *et al.*, 2022) afin de pouvoir utiliser des mesures comme le κ de Cohen sur des structures d'annotations qui ne s'y prêtent pas originellement. Pour les raisons que nous avons évoquées ci-dessus, nous pensons néanmoins qu'il faut privilégier des mesures spécifiquement adaptées à ces structures. Avec le CST, nous cherchons à comparer ici quatre mesures dédiées pour l'accord sur les structures argumentatives : la première *mesure d'accord*

inter-annotateurs basée sur les graphes de (Kirschner *et al.*, 2015) et les trois variantes de *mean average recall* (*MAR*) [1] de (Putra *et al.*, 2022).

Mesure de Kirschner : Soient deux structures argumentatives A et B. Cette mesure adaptée aux graphes dirigés évalue le taux d'inclusion du graphe A dans le graphe B, puis du graphe B dans le graphe A et donne finalement la moyenne (arithmétique ou harmonique) des deux taux.
Pour E_A l'ensemble des relations (x,y) dans le graphe A où x est l'origine de la relation et y la cible et $SP_B(x,y)$ la longueur du plus court chemin entre les nœuds x et y dans le graphe B, l'inclusion de A dans B est donnée par la formule 1 :

$$\text{Mesure de Kirschner} = \frac{1}{|E_A|} \sum_{(x,y)\in E_A} \frac{1}{SP_B(x,y)} \tag{1}$$

MAR^{link} : Cette mesure est la moyenne du rappel des relations de deux structures relationnelles A et B. Soient E_A l'ensemble des relations dans A et E_B l'ensemble des relations dans B, la moyenne des rappels de ces ensembles est donnée par la formule 2.

$$\text{MAR}^{link} = \frac{1}{2}\left(\frac{|E_A \cap E_B|}{|E_A|} + \frac{|E_A \cap E_B|}{|E_B|}\right) \tag{2}$$

MAR^{path} : Plutôt que de s'appuyer sur les relations elles-mêmes, cette variante cherche à mesurer l'accord sur les chemins créés par la mise bout à bout des relations dans la structure argumentative. Pour deux relations dirigées (a,b) et (b,c) où le premier élément est la source et le second l'origine, on obtient trois chemins : [a,b,c], [a,b] et [b,c].
On note respectivement P_A et P_B les ensembles des chemins pour les graphes A et B. L'accord entre ces ensembles est donné par la formule 3 :

$$\text{MAR}^{path} = \frac{1}{2}\left(\frac{|P_A \cap P_B|}{|P_A|} + \frac{|P_A \cap P_B|}{|P_B|}\right) \tag{3}$$

MAR^{dSet} : Cette variante mesure l'accord entre les ensembles de descendants de chaque unité dans la structure argumentative en partant de la racine (*i.e.* le CA *proposition principale*). L'ensemble des descendants d'une unité a est composé de l'unité a elle-même et de toutes les unités qui sont ses descendantes dans l'arbre (c'est-à-dire toutes les unités qui peuvent être atteintes en remontant en sens inverse les relations dirigées en partant de l'unité a). Soient deux jeux d'annotations A et B contenant respectivement les ensembles d'unités (a_1,a_2,...a_n) et (b_1,b_2,...b_n). Une fonction f est définie et prend en paramètre le jeu d'annotations A et retourne un vecteur comprenant les scores de correspondance entre les ensembles de descendants de a_i et de b_i pour $a_i \in$ A et $b_i \in$ B. N_A et N_B correspondent respectivement au nombre d'unités dans A et B. L'accord est donné par la formule 4 :

$$\text{MAR}^{dSet} = \frac{1}{2}\left(\frac{\sum f(A)}{|N_B|} + \frac{\sum f(B)}{|N_A|}\right) \tag{4}$$

1. Au sein même des variantes de MAR, on peut trouver des sous-variantes qui apprécient différemment les relations si elles ont pour trait d'être *explicit* ou *implicit*. Cette distinction, qui est engendrée par le cas particulier de la catégorie *reformulation* peut faire sens dans le cadre de la campagne ICNALE. Nous faisons cependant le choix d'écarter cette catégorie de relation si particulière à une étude pour privilégier la généralisation de nos résultats ; ainsi lorsque nous évoquons par la suite les variantes MAR, nous ne les décrivons que sous leur forme utilisant des relations uniquement *explicit*.

Le score de correspondance entre deux ensembles de descendants peut être *exact* (la valeur est de 1 si les deux ensembles de correspondants pour a_i et b_i sont identiques, 0 sinon) ou *partial* (la valeur est égale au nombre d'unités présentes dans les deux ensembles divisée par le nombre d'unités dans l'ensemble des descendants de b_i). Dans toutes les configurations possibles, le score de correspondance *partial* est toujours supérieur ou égal au score de correspondance *exact*. Nous implémentons dans nos expériences les versions *exact* et *partial* de MAR^{dSet} afin de les comparer.

4.4 Environnement expérimental

Il n'existe pas à notre connaissance d'outil en libre accès pour établir les scores des quatre mesures que nous tentons d'évaluer. Le CST tel que présenté par (Mathet *et al.*, 2012) n'est pas outillé pour des structures relationnelles. Face à ce double manque de logiciels à notre disposition nous avons fait le choix de créer nos propres outils de dégradation d'une référence et de calcul de mesures d'accord pour des structures relationnelles. L'ensemble de ces programmes est réalisé avec les langages *Python 3* et *DOT* et sera rendu accessible à la communauté [2].

4.5 Observations et commentaires

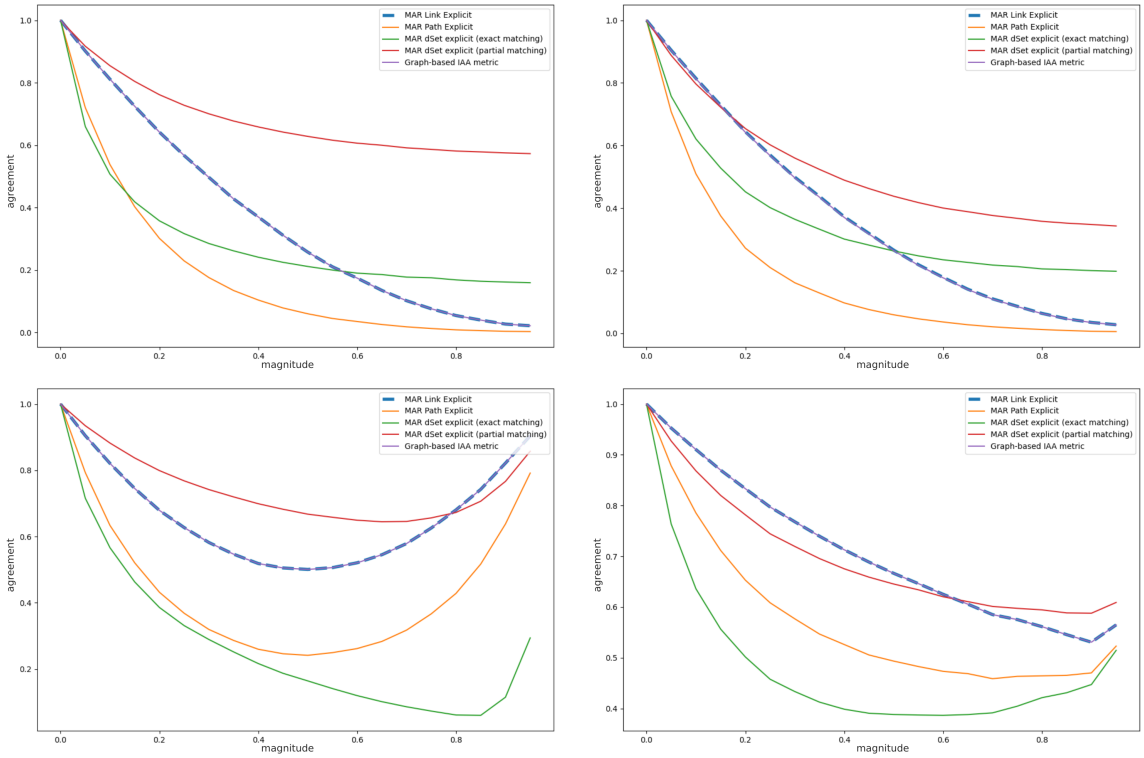

FIGURE 1 – *Changement de cible* (en haut à gauche), *changement d'origine* (en haut à droite), *permutation d'orientation* (en bas à gauche) et *faux négatif* (en bas à droite)

2. https://www.greyc.fr/equipes/codag/#ressources

Nous comparons ici les cinq mesures décrites ci-dessus : la mesure d'accord inter-annotateurs basée sur les graphes (*graph-based IAA metric* (GBM)) (Kirschner *et al.*, 2015), MAR^{link}, MAR^{path}, MAR^{dSet} *exact matching* et MAR^{dSet} *partial matching* (Putra *et al.*, 2022). La figure 1 présente le comportement de ces mesures pour 4 paradigmes de dégradation différents. Dans tous ces paradigmes on remarque que les scores de GBM et MAR^{link} sont toujours identiques alors que ces deux mesures ont des méthodes de calcul différentes. Nous avons remarqué lors de nos expérimentations que ces deux mesures donnent le même score lorsque deux jeux d'annotations comparés possèdent le même nombre de relations, comme c'est le cas dans nos dégradations contrôlées, mais que leurs valeurs divergent lorsque le nombre de relations des deux jeux de données diffère. Pour la suite de nos observations, ce que nous commenterons sur GBM s'appliquera donc aussi à MAR^{link}.
Pour le *changement de cible* et le *changement d'origine*, toutes les mesures suivent une évolution décroissante monotone, ce qui est attendu. Les deux variantes de MAR^{dSet} semblent être limitées chacune par une asymptote, tandis que GBM et MAR^{path} montrent une étendue complète de 1 (accord parfait) à 0 (absence d'accord).
Dans le paradigme de *permutation d'orientation*, aucune des mesures n'est monotone malgré l'augmentation des dégradations. MAR^{path} et GBM atteignent leur minimum lorsque la magnitude est à 0,5 (*i.e.* lorsque la moitié des relations sont inversées dans la structure argumentative) et sont symétriques à l'axe $x = 0.5$. Les variantes de MAR^{dSet} ne présentent pas cette symétrie. Cela s'explique car elles sont les seules mesure à avoir un *sens de lecture particulier* de la structure argumentative, de la racine vers les feuilles. Mécaniquement, leurs scores ne remontent que lorsque les ensembles de descendants se vident et deviennent de moins en moins différents les uns des autres.
Pour le *faux négatif*, aucune des mesures n'est pleinement monotone. Les mesures GBM, MAR^{path} et MAR^{dSet} *partial matching* sont décroissante monotone jusqu'à la magnitude 0,9 ; le score de MAR^{dSet} *exact matching* semble atteindre un plateau minimum dès la magnitude 0.45 avant de croître. Ce phénomène de non-monotonie des courbes pour le *faux négatif* après la magnitude $m = 0.9$ est déjà présent dans les premiers travaux sur le CST (Mathet *et al.*, 2012) pour des paradigmes d'annotation et des mesures différents. Cela est manifestement lié à l'implémentation du *faux négatif* qui est similaire dans les deux études. Nous pouvons l'expliquer simplement, moins il reste d'annotations à comparer dans les simulations générées par le CST et moins les mesures peuvent y trouver du désaccord.

5 Conclusion et perspectives

De manière similaire à (Mathet *et al.*, 2012), notre implémentation du CST pour des structures relationnelles présente des résultats exploitables pour les mesures d'accord inter-annotateurs et les paradigmes de dégradations étudiés. Les comparaisons offertes par cet outil mettent en évidence des similarités entre certaines méthodes, mais surtout les différences et les écueils qui caractérisent les mesures les plus récentes dans le domaine de l'annotation de l'argumentation.

Au rang des perspectives nous prévoyons dans de futurs travaux avec notre outil d'utiliser des références issues directement de campagnes d'annotation et d'améliorer la simulation des dégradations en tentant de nous rapprocher des erreurs commises par les annotateurs humains et de leur fréquence. De plus, l'implémentation du CST que nous proposons aujourd'hui est conçue et spécialisée pour la question de l'accord dans les structures argumentatives. Il serait intéressant de travailler sur d'autres paradigmes de structures relationnelles tels que celui de la coréférence qui possèdent un grand panel de mesures d'accord inter-annotateurs dédiées.

Remerciements

Nous remercions grandement Yann Mathet et Antoine Widlöcher pour leur aide précieuse en tant qu'encadrants, mais aussi pour leurs encouragements et leurs nombreux conseils tout au long de la réalisation de ce travail. Nous remercions également les trois relecteurs anonymes pour leurs commentaires très utiles à l'amélioration de ce document.

Références

ARTSTEIN R. & POESIO M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, **34**(4), 555–596. DOI : 10.1162/coli.07-034-R2.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46. Place : US Publisher : Sage Publications, DOI : 10.1177/001316446002000104.

FORT K. (2016). *Collaborative Annotation for Reliable Natural Language Processing : Technical and Sociological Aspects*. Wiley, 1 édition. DOI : 10.1002/9781119306696.

ISHIKAWA S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner corpus studies in Asia and the world*, **1**, 91–118.

ISHIKAWA S. (2018). The ICNALE Edited Essays : A Dataset for Analysis of L2 English Learner Essays Based on a New Integrative Viewpoint. *English Corpus Linguistics*, **25**, 1–14.

KIRSCHNER C., ECKLE-KOHLER J. & GUREVYCH I. (2015). Linking the Thoughts : Analysis of Argumentation Structures in Scientific Publications. In *Proceedings of the 2nd Workshop on Argumentation Mining*, p. 1–11, Denver, CO : Association for Computational Linguistics. DOI : 10.3115/v1/W15-0501.

KRIPPENDORFF K. (2019). *Content Analysis : An Introduction to Its Methodology*. SAGE Publications, Inc. DOI : 10.4135/9781071878781.

LEECH G. N., GARSIDE R. G. & MCENERY T. (1997). *Corpus annotation : linguistic information from computer text corpora / Roger Garside, Geoffrey Leech, Tony McEnery.* London : Longman.

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 M. KAY & C. BOITET, Éds., *Proceedings of COLING 2012 : Posters*, p. 809–818, Mumbai, India : COLING 2012 Organizing Committee.

PARK J. & CARDIE C. (2018). A Corpus of eRulemaking User Comments for Measuring Evaluability of Arguments. In N. CALZOLARI, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, K. HASIDA, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS & T. TOKUNAGA, Éds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : European Language Resources Association (ELRA).

PUTRA J. W. G., TEUFEL S. & TOKUNAGA T. (2022). Annotating argumentative structure in English-as-a-Foreign-Language learner essays. *Natural Language Engineering*, **28**(6), 797–823. Publisher : Cambridge University Press, DOI : 10.1017/S1351324921000218.

SKEPPSTEDT M., PELDSZUS A. & STEDE M. (2018). More or less controlled elicitation of argumentative text : Enlarging a microtext corpus via crowdsourcing. In N. SLONIM & R. AHARONOV, Éds., *Proceedings of the 5th Workshop on Argument Mining*, p. 155–163, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5218.

TEUFEL S., CARLETTA J. & MOENS M. (1999). An annotation scheme for discourse-level argumentation in research articles. In H. S. THOMPSON & A. LASCARIDES, Éds., *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, p. 110–117, Bergen, Norway : Association for Computational Linguistics.

WACHSMUTH H., AL-KHATIB K. & STEIN B. (2016). Using Argument Mining to Assess the Argumentation Quality of Essays. In Y. MATSUMOTO & R. PRASAD, Éds., *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, p. 1680–1691, Osaka, Japan : The COLING 2016 Organizing Committee.

A Annexes

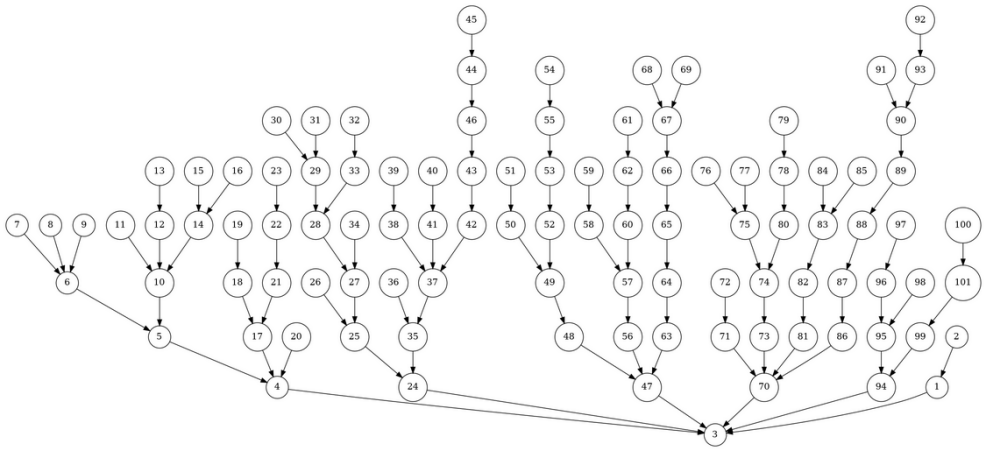

FIGURE 2 – Structure argumentative de référence fournie à notre implémentation du CST, l'AC 3 est de type *proposition principale*