TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Une approche holiste et unifiée de l'alignement et de la mesure d'accord inter-annotateurs

Yann Mathet[1] Antoine Widlöcher[1]
(1) GREYC, UMR CNRS 6072, Université de Caen, 14032 Caen Cedex
{prenom.nom}@unicaen.fr

Résumé. L'alignement et la mesure d'accord sur des textes multi-annotés sont des enjeux majeurs pour la constitution de corpus de référence. Nous défendons dans cet article l'idée que ces deux tâches sont par essence interdépendantes, la mesure d'accord nécessitant de s'appuyer sur des annotations alignées, tandis que les choix d'alignements ne peuvent se faire qu'à l'aune de la mesure qu'ils induisent. Nous proposons des principes formels relevant cette gageure, qui s'appuient notamment sur la notion de désordre du système constitué par l'ensemble des jeux d'annotations d'un texte. Nous posons que le meilleur alignement est celui qui minimise ce désordre, et que la valeur de désordre obtenue rend compte simultanément du taux d'accord. Cette approche, qualifiée d'holiste car prenant en compte l'intégralité du système pour opérer, est algorithmiquement lourde, mais nous sommes parvenus à produire une implémentation d'une version légèrement dégradée de cette dernière, et l'avons intégrée à la plate-forme d'annotation Glozz.

Abstract. Building reference corpora makes it necessary to align annotations and to measure agreement among annotators, in order to test the reliability of the annotated ressources. In this paper, we argue that alignment and agreement measure are interrelated : agreement measure applies to pre-aligned data and alignment assumes a prior agreement measure. We describe here a formal and computational framework which takes this interrelation into account, and relies on the notion of disorder of annotation sets available for a text. In this framework, the best alignment is the one which has the minimal disorder, and this disorder reflects an agreement measure of these data. This approach is said to be holistic insofar as alignment and measure depend on the system as a whole and cannot be locally determined. This holism introduces a computational cost which has been reduced by a heuristic strategy, implemented within the Glozz annotation platform.

Mots-clés : Alignement d'annotations, mesure d'accord inter-annotateurs, linguistique de corpus.

Keywords: Alignment, inter-coder agreement measure, corpus linguistics.

1 Contexte

La multiplication des travaux sur corpus, en linguistique computationnelle et en TAL conduit naturellement à la multiplication des campagnes d'annotation et rend nécessaire la mise en place de méthodes et d'outils permettant d'interpréter le fruit de ces campagnes. Pour établir des corpus annotés de référence, ou simplement pour mieux comprendre les phénomènes linguistiques que ces campagnes prennent pour objets, il est notamment nécessaire de mettre en correspondance (d'aligner) les annotations produites par différents annotateurs (humains ou automatiques), sur un même jeu de données, et de prendre la mesure de leurs accords et désaccords.

Dans cet article, nous nous intéressons aux questions d'alignement et d'accord inter-annotateurs, en nous limitant à des annotations de textes consistant, de façon très générale, à délimiter et à catégoriser des unités. Il est important de noter que la méthode que nous cherchons à définir doit permettre d'aligner et de comparer des objets textuels relativement variés, distribués dans le texte de manières elles aussi variées, et qu'à ce titre, nous devons nous écarter de nombreux travaux eux aussi consacrés à l'alignement et à la mesure d'accord (*cf.* section 2).

Nous cherchons à aligner et à comparer des *unités*, segments de texte commençant et s'achevant en des positions déterminées. Insistons sur le fait que la segmentation du texte, *i.e.* le positionnement des unités, n'est pas considérée comme acquise. En effet, dans certains cas, les annotateurs n'auront pas exclusivement à caractériser des données déjà délimitées, mais devront également déterminer leur position dans le texte et leur taille. Concernant ce

positionnement des unités, précisons de plus qu'il ne conduit pas nécessairement à un pavage complet du texte, la sporadicité des phénomènes étant même parfois assez grande. Concernant leur taille, ajoutons que celle-ci pourra varier fortement d'une unité à l'autre, et cela, éventuellement, pour un même type d'objet linguistique. Pour le positionnement relatif des unités, nous souhaitons de plus offrir une grande souplesse : les unités pourront se succéder, s'inclure, se chevaucher. Chaque unité possède par ailleurs une *catégorie* choisie parmi un ensemble prédéfini pour une campagne d'annotation donnée. Ajoutons que les couches d'annotations correspondant aux différentes catégories ne doivent pas être regardées comme indépendantes, l'attribution d'une mauvaise catégorie à un objet pouvant être parfois, dans une certaine mesure, acceptable.

Les raisons pour lesquelles nous devons privilégier la tolérance tiennent dans une large mesure à la nature des objets linguistiques sur lesquels nous travaillons par ailleurs. En effet, nous opérons souvent dans le champ disciplinaire de l'analyse du discours et explorons des structures textuelles variées, telles qu'envisagées par des approches aussi hétérogènes que l'*Argumentative Zoning* (Teufel *et al.*, 1999), l'*encadrement du discours* (Charolles, 1997) ou encore la SDRT (*Segmented Discourse Relation Theory*) (Asher, 1993). À titre d'exemple, précisons que ce travail prend place dans la continuité du projet ANR Annodis (Péry-Woodley *et al.*, 2009), qui vise la mise en place d'un corpus de référence pour le français, en analyse de discours. Comme on le verra, il entretient par ailleurs de nombreuses relations avec la plate-forme d'annotation et d'exploration de corpus Glozz (Widlöcher & Mathet, 2009), plate-forme permettant de produire des annotations hétérogènes exigeant cette tolérance.

La méthode que nous proposons ici n'est néanmoins pas dédiée à l'évaluation d'annotations discursives. Elle se veut aussi générique que possible et nous pouvons résumer ainsi son objectif : nous recherchons à la fois un alignement et une mesure d'accord multi-annotateurs portant sur des annotations composées d'unités marquées par leur possible variété de grain, leur possible variété catégorielle, leur possible sporadicité et la souplesse de leurs distributions relatives.

2 État de l'art

Parmi les travaux dans la continuité desquels notre étude prend position, nous pouvons distinguer ceux qui portent leur attention sur la question de l'attribution de catégories à des unités prédéfinies (la caractérisation) d'une part et ceux qui privilégient la question de la segmentation des unités d'autre part.

Pour les premiers, l'accord entre annotateurs concerne principalement l'affectation, par chacun, d'une catégorie choisie parmi un ensemble défini pour une campagne d'annotation donnée, à des unités dont la délimitation est considérée comme non problématique, souvent le mot. Dans cette perspective, de nombreux travaux se réfèrent notamment aux coefficients que sont π (Scott, 1955) et κ (Cohen, 1960) ainsi qu'à des variantes multi-annotateurs tel le K de (Siegel & Castellan, 1988) et aux coefficients pondérés α (Krippendorff, 1980) et κ_w (Cohen, 1968)[1]. Non spécifiquement issues du TAL ou de la linguistique, ces différentes approches de la mesure d'accord font l'objet de travaux qui visent à étudier leur pertinence et leurs limites dans ces domaines d'accueil et à en comparer les retombées. Nous pensons ici en particulier à l'excellente synthèse de (Artstein & Poesio, 2008), sur laquelle nous nous appuyons fortement ici. Si la présente étude apporte, comme on le verra, au problème de la caractérisation une réponse provisoire relativement légère, la fréquentation de ces travaux nous ouvre toutefois d'ores et déjà des perspectives essentielles, dont l'influence sera encore accrue dans nos travaux futurs. En particulier, l'importance que ces travaux accordent à la confrontation entre les accords observés et ceux que le seul hasard peut engendrer est tout à fait éclairante, de même que l'est leur réflexion sur l'obtention du « meilleur hasard possible », qui tiendra compte des propriétés particulières de la campagne engagée et de son corpus, ainsi, le cas échéant, que des tendances des annotateurs. Une autre avancée importante concerne la proposition de solutions, intimement liée à la réflexion sur les coefficients pondérés, permettant de rendre compte du fait important que tous les désaccords ne se valent pas. Ce point sera évidemment au cœur de la question de la segmentation, mais, dans la continuité de ces travaux, nous y accorderons aussi une large place en ce qui concerne la caractérisation des unités. Enfin, mentionnons la place méritée que ces travaux accordent à la délicate question de l'interprétation qualitative des résultats quantitatifs. Dans le prolongement de leur effort, nous serons aussi amenés à envisager des « grilles » permettant l'interprétation des mesures que nous proposons.

Naturellement, la principale limite de ces travaux, du point de vue qui nous occupe, est le fait que la segmentation y

[1] Comme le note (Artstein & Poesio, 2008), certains flottements de dénomination perturbent souvent les discussions relatives à ces coefficients. Nous retenons la clarification qu'ils proposent.

soit globalement considérée comme acquise. Il convient toutefois de remarquer, et (Artstein & Poesio, 2008) nous y invitent, que ces approches peuvent fournir un cadre pour l'estimation de l'accord sur des tâches de segmentation. Ainsi, (Teufel *et al.*, 1999) envisagent par exemple l'accord obtenu sur l'attribution de rôles argumentatifs à des phrases, en utilisant le coefficient de (Siegel & Castellan, 1988), phrases dont l'ajacence conduit à l'émergence de segments. Les mesures d'accord sur l'attribution de catégories peuvent encore être utilisées, comme c'est le cas dans (Hearst, 1997), non plus sur le contenu des segments, mais pour mesurer l'accord sur l'identification des bornes, c'est-à-dire sur l'attribution d'une catégorie *borne*. Toutefois, l'utilisation de ces approches pour des tâches de segmentation se heurte à la difficulté majeure suivante : une délimitation d'unité n'est regardée comme faisant consensus que si les annotateurs sont parfaitement d'accord sur le positionnement exact des bornes. Or en la matière, et en particulier aux échelles discursives, une plus grande souplesse est nécessaire, pour que de légers désaccords dans le positionnement des bornes soient moins lourdement pénalisés.

La méthode proposée par (Grouin *et al.*, 2011), adossée à la mesure d'erreur *slot error rate* (Makhoul *et al.*, 1999), permet de combiner alignement et mesure d'erreur et d'aborder simultanément positionnement des unités et attribution de catégories. Certes, elle permet d'aligner des unités dont les positions ne sont pas identiques, mais les différentes raisons suivantes la rendent peu adaptée à notre perspective : tous les écarts sont sanctionnés de manière identique ; elle est prévue pour comparer seulement deux annotations, dont l'une fait office de référence ; enfin, elle opère à l'échelle de la phrase et non du texte.

D'autres travaux visent à prendre spécifiquement en charge les problèmes de segmentation, en particulier dans le domaine de la segmentation thématique. Dans ce domaine, un consensus s'est établi autour de la mesure WindowDiff (Pevzner & Hearst, 2002), qui consiste à déplacer une fenêtre glissante le long du texte, et à comparer le nombre de ruptures présentes dans une annotation considérée comme référence et dans une annotation évaluée. Aux limites de cette méthode évoquées par exemple par (Lamprier *et al.*, 2007) et (Bestgen, 2009) (difficulté à interpréter les résultats, dépendance à l'égard de la taille de la fenêtre glissante, erreurs pénalisées différemment selon leur position dans le texte, erreurs légères parfois trop pénalisées...) s'ajoute dans notre perspective le fait que cette méthode ne fournit pas à proprement d'alignement, limite qui s'applique également aux aménagements de WindowDiff proposés par (Lamprier *et al.*, 2007). (Bestgen, 2009) préconise pour sa part le recours à la distance de Hamming généralisée (DHG) (Bookstein *et al.*, 2002), distance d'édition entre deux annotations, qui ajoute à la distance de Hamming l'opération de déplacement qui permet de donner la souplesse nécessaire à la prise en compte d'erreurs légères dans le positionnement des bornes. Offrant un résultat plus facile à interpréter que d'autres indices, cette méthode souffre selon nous de limites qui s'appliquent du reste également à WindowDiff. Pensées (dans le cas de WindowDiff) ou détournées (dans le cas de DHG) pour l'évaluation de la segmentation thématique, ces méthodes sont intimement liées d'une part à l'idée de pavage complet du texte (ce qui enfreint notre contrainte d'éventuelle sporadicité) et d'autre part à l'unicité du phénomène envisagé, *i.e.* le phénomène de rupture thématique (ce qui enfreint notre contrainte de prise en charge d'annotation multi-catégorielles). Ajoutons que ces méthodes n'intègrent pas de correction par le hasard. La solution α_U proposée par (Krippendorff, 1995), qui repose sur la mesure du recouvrement entre les annotations de différents annotateurs répond à beaucoup des exigences que nous avons fixées et nous devrons l'évaluer davantage. Elle impose toutefois, comme l'indiquent (Artstein & Poesio, 2008), que les annotations d'un même annotateur ne se recouvrent pas, ce qui contredit la contrainte de souplesse positionnelle que nous nous sommes fixée.

3 Quelle mesure d'accord ?

3.1 Difficulté majeure : interdépendance de la mesure d'accord et de l'alignement

Pour un « même phénomène » repéré par plusieurs annotateurs, il est nécessaire de prévoir une mesure d'accord suffisamment souple pour pouvoir rendre compte d'une double divergence, la première portant sur le choix de catégorie attribuée au phénomène, la seconde portant sur son positionnement. Il n'est pas rare, en particulier, que le positionnement diffère de façon substantielle sur l'une, l'autre, ou même les deux bornes. Du fait de ces divergences de positionnement, la mesure d'accord est assujettie à la détermination d'un alignement inter-annotateurs, un tel alignement consistant à déterminer quelle unité de tel annotateur correspond à telle autre de tel autre annotateur. Si l'on dispose d'un alignement complet des annotations de l'ensemble des annotateurs, il est possible, pour chaque unité repérée, de déterminer dans quelle mesure les annotateurs se sont entendus sur son positionnement et sur sa qualification. Cette quantification sera établie au moyen d'une mesure de « dissimilarité » entre

unités annotées : plus les unités seront considérées comme « proches », plus cette mesure devra être faible. Des propositions relatives à l'établissement de telles mesures seront faites ci-après.

Mais comment obtenir un tel alignement ? Aligner une unité u_a de l'annotateur A avec une unité u_b de l'annotateur B consiste à considérer que les propriétés (catégorie, position) de u_a et de u_b sont suffisamment « proches » pour pouvoir être assimilées : l'annotateur A et l'annotateur B ont rendu compte d'un « même phénomène », bien que de façon éventuellement (et légèrement) différente. La méthode d'alignement doit donc pour sa part s'appuyer sur une « distance » entre unités pour pouvoir opérer.

Dès lors, mesure d'accord et alignement sont inter-dépendants : on ne peut mesurer sans disposer d'un alignement, ni l'on ne peut aligner sans disposer d'une mesure, si bien que ces deux processus ne peuvent constituer deux étapes successives. Cette interdépendance renvoie simplement à l'unicité de l'objectif effectivement posé : établir dans quelle mesure des éléments éventuellement différents peuvent malgré tout être considérés comme semblables, soit pour quantifier ces différences (dans le cas de la mesure), soit pour assimiler des unités « similaires » (dans le cas de l'alignement). Il est donc nécessaire de disposer d'une méthode unifiée pour la mesure et l'alignement.

3.2 Le désaccord comme créateur de désordre

Considérons un ensemble de n annotateurs travaillant sur un même texte et une même tâche d'annotation. Idéalement, si la tâche d'annotation était rigoureusement établie, et si elle relevait de phénomènes ne prêtant pas à confusion, les n annotateurs devraient délivrer le même ensemble d'éléments annotés. C'est cependant bien entendu un constat que nous ne faisons jamais. Pour autant, les différentes propositions des annotateurs devraient en grande partie converger, à défaut de quoi la tâche proposée devrait être considérée comme un échec (tâche trop peu spécifiée, phénomènes étudiés ne permettant aucun consensus...). Ainsi, pour une campagne d'annotation donnée, on constatera un « taux d'accord » inter-annotateurs situé entre l'idéal constitué par une annotation unique (les n annotateurs ont annoté exactement le même ensemble d'unités) et le cas le pire constitué par n générateurs aléatoires d'annotations. L'enjeu de la mesure d'accord est de situer ce jeu d'annotations entre ces deux extrêmes.

Notre proposition est de considérer que l'annotation multiple est potentiellement génératrice de désordre. Le cas idéal (dans lequel tous les annotateurs ont délivré exactement le même jeu d'annotations) peut être considéré comme parfaitement ordonné : l'information portée par les annotations d'un annotateur donné est parfaitement confirmée par les annotations de chacun des autres annotateurs. Par rapport à cette situation idéale, opérons un ensemble de transformations élémentaires sur un certain nombre d'unités : déplacement de l'une des deux bornes d'une unité, requalification de sa catégorie, ou encore, suppression pure et simple. Chacune de ces transformations va engendrer un certain désordre au sein de ce système. Le désordre total obtenu pour un ensemble de transformations élémentaires sera la résultante de l'ensemble des désordres élémentaires ainsi créés. Nous définirons ci-après un cadre formel et une méthode de calcul de ce désordre, et poserons que le taux d'accord inter-annotateurs correspond au niveau d'ordre du système relativement au niveau d'ordre d'un système construit aléatoirement.

4 Dissimilarité, alignement, entropie et accord

4.1 Définitions : unité, annotateur, jeu d'annotations

Nous définissons tout d'abord $\mathcal{A}$ l'ensemble des annotateurs, $\mathcal{T}$ l'ensemble des textes et $\mathcal{U}$ l'ensemble des unités.

Unité : une unité u possède une catégorie notée $cat(u)$, et une position donnée par ses deux bornes, correspondant chacune à un indice de caractère du texte, notées respectivement $start(u)$ et $end(u)$, $start$ et end étant donc des fonctions de $\mathcal{U}$ vers $\mathbb{N}^+$. Nous définissons l'égalité entre deux unités comme suit :

$$\forall (u,v) \in \mathcal{U}^2, u = v \Leftrightarrow ((cat(u) = cat(v)) \wedge (start(u) = start(v)) \wedge (end(u) = end(v)))$$

Une unité est produite par un annotateur donné, et est relative à un texte donné (dans le cadre d'une campagne donnée). L'unité émanant de l'annotateur a et de rang i est notée u_a^i.

Jeu d'Annotations : un jeu d'annotations j est un ensemble d'unités relatives à un même texte et produites par un ensemble donné d'annotateurs. Un tel jeu est dit aléatoire quand ses annotateurs sont des processus aléatoires.

4.2 Dissimilarité entre deux unités

Une dissimilarité est une fonction $d : \mathcal{U}^2 \rightarrow \mathbb{R}^+$, telle que :

$$\forall (u,v) \in \mathcal{U}^2, \begin{cases} d(u,v) = d(v,u) \text{ (d est symétrique)} \\ d(u,v) = 0 \Leftrightarrow u = v \end{cases}$$

Une dissimilarité n'est pas nécessairement une distance au sens mathématique dans la mesure où l'inégalité triangulaire n'est pas imposée. Nous verrons pourquoi.

4.2.1 Dissimilarité positionnelle d_{pos}

Il est possible de proposer différentes mesures de dissimilarités positionnelles pour différents paradigmes d'annotation. Nous nous en tiendrons ici à la dissimilarité suivante, bien adaptée à des annotations sporadiques :

$$d_{pos-sporadique}(u,v) = \left(\frac{|start(u) - start(v)| + |end(u) - end(v)|}{\left(\frac{(end(u)-start(u))+(end(v)-start(v))}{2} \right)} \right)^2 \quad (1)$$

Cette dissimilarité rend compte des différences entre les bornes gauches des deux unités ainsi qu'entre leurs bornes droites. Sa croissance est quadratique par rapport à la somme de ces différences, si bien que l'on pénalise d'autant plus les écarts importants. Elle ne respecte pas l'inégalité triangulaire pour cette raison. Par ailleurs, le fait de diviser les différences par la moyenne des deux longueurs des unités (*cf.* dénominateur) permet de rendre la dissimilarité insensible aux changements d'échelle. C'est un choix qui peut être discuté selon la campagne d'annotation envisagée. A titre d'exemple, une seconde dissimilarité positionnelle est actuellement expérimentée pour émuler la distance de Hamming généralisée, basée sur la longueur moyenne des unités notée k :

$$d_{pos-Hamming}(u,v) = \frac{|end(u) - end(v)|}{k/2} \quad (2)$$

4.2.2 Dissimilarité catégorielle d_{cat}

Soit C l'ensemble des catégories. Pour une campagne d'annotation donnée, n catégories distinctes sont définies.

Nous définissons tout d'abord la distance catégorielle entre catégories $dist_{cat}$ au moyen d'une matrice carrée de taille n, prenant l'ensemble des catégories à la fois sur les lignes et sur les colonnes. Chaque case indique la distance entre deux catégories par une valeur située dans l'intervalle $[0, 1]$. La valeur 0 signifie l'égalité des catégories (du fait des propriétés d'une distance), tandis que la valeur 1, maximale, signifie que les deux catégories sont incompatibles (l'une ne peut en aucun cas se substituer à l'autre). Une telle matrice est nécessairement symétrique et possède une diagonale nulle, du fait, là encore, des propriétés d'une distance. Voici un exemple de matrice rendant compte d'un ensemble de 3 catégories. Elle permet une correspondance possible entre une unité de type cat_1 avec une unité de type cat_2, avec un coût de 0.5 (qui sera à mettre en balance avec les coûts issus des dissimilarités positionnelles), et elle interdit les autres correspondances :

	cat_1	cat_2	cat_3
cat_1	0	0.5	1
cat_2	0.5	0	1
cat_3	1	1	0

TAB. 1 – Exemple de matrice pour 3 catégories

On définit alors la dissimilarité catégorielle entre deux unités par :

$$d_{cat}(u,v) = dist_{cat}(cat(u), cat(v)).\Delta_{\emptyset} \tag{3}$$

$\Delta_{\emptyset}$ est une constante qui sera définie ultérieurement et qui assure ici notamment que deux unités de catégories distinctes ne seront jamais alignées.

4.2.3 Dissimilarités combinée d_{combi}

Soient deux dissimilarités d_1 et d_2 données. On définit $d_{combi(d_1,d_2,\alpha,\beta)}$ par :

$$d_{combi(d_1,d_2,\alpha,\beta)}(u,v) = \alpha.d_1(u,v) + \beta.d_2(u,v) \tag{4}$$

Cette combinaison linéaire de dissimilarités est elle-même une dissimilarité. Elle permet notamment, dans le cas où $\alpha = 0.5$ et $\beta = 0.5$, de donner un poids égal à deux dissimilarités (par ex. positionnelle et catégorielle).

4.3 Alignement

4.3.1 Alignement unitaire $\dot{a}$

Un alignement unitaire $\dot{a}$ est un i-uplet, i étant compris entre 1 et n, n étant le nombre d'annotateurs, contenant au plus une unité de chaque annotateur. Pour des raisons d'homogénéité facilitant notamment son implémentation informatique, nous créons une unité fictive vide, notée $u_{\emptyset}$, correspondant à la réification du fait qu'un alignement unitaire ne contienne aucune unité pour un annotateur donné. Nous ferons comme si cet alignement contenait cette unité fictive pour cet annotateur là, si bien que tout alignement unitaire devient finalement, dans tous les cas, un n-uplet, contenant au moins une unité non vide, et, pour chaque annotateur, soit l'une de ses unités, soit $u_{\emptyset}$. Pour n annotateurs numérotés de 1 à n, et ayant respectivement annoté $card_i$ unités, le nombre d'alignement unitaires qu'il est possible de générer est de $(\prod_{i=1}^{n} card_i) - 1$ (en retirant l'alignement ne contenant que des $u_{\emptyset}$).

4.3.2 Alignement $\bar{a}$

Pour un jeu d'annotations donné, un alignement $\bar{a}$ est défini comme un[2] ensemble d'alignements unitaires tel que chaque unité de chaque annotateur apparaît dans un et un seul de ses alignements unitaires.

4.4 Alignement et entropie[3]

4.4.1 Entropie d'un alignement unitaire

L'entropie d'un alignement unitaire $\dot{a}$, notée $\dot{e}(\dot{a})$, est définie pour une dissimilarité d_x donnée comme la valeur moyenne des dissimilarités deux à deux de ses unités constituantes :

$$\dot{e}(\dot{a}) = \frac{1}{C_n^2} \cdot \sum_{(u,v)\in\dot{a}^2} d_x(u,v) \tag{5}$$

Cependant, étant donné qu'un alignement unitaire peut comporter des unités fictives $u_{\emptyset}$, il est nécessaire de définir la dissimilarité entre une unité réelle et l'unité fictive $u_{\emptyset}$.

Pour toute dissimilarité d_x, pour toute unité u, $d_x(u,u_{\emptyset}) = d_x(u_{\emptyset},u) = \Delta_{\emptyset}$, constante qui est à définir pour une campagne donnée. En effet, cette valeur indique jusqu'à quel seuil de dissimilarité il convient de préférer aligner une unité avec une autre plutôt qu'avec $u_{\emptyset}$. Par exemple, si on choisit comme dissimilarité positionnelle

[2]Notons que, pour n annotateurs qui auraient chacun créé le même nombre p d'unités, le nombre d'alignements qu'il est possible de générer est supérieur à $(p!)^{n-1}$, ce qui dépasse très rapidement les capacités de stockage et de traitement des machines.

[3]Le terme entropie est ici quelque peu usurpé, pris seulement pour évoquer la notion de désordre.

$d_{pos-sporadique}$, et que l'on souhaite que deux unités de même longueur soient alignables tant qu'elles se touchent (et en faisant abstraction de la dissimilarité catégorielle éventuelle), on calcule la dissemblance d'une telle configuration (unité u positionnée de x à $x + l$ et v positionnée de $x + l$ à $x + 2l$, soit $d_{pos-sporadique}(u, v) = ((l + l)/l)^2 = 2^2 = 4$) et on pose donc $\Delta_\emptyset = 4$. Par ailleurs, concernant la dissimilarité catégorielle, la formule (3) montre qu'une valeur de 1 dans la matrice des distances fait systématiquement préférer le choix de $u_\emptyset$ (valeur $\Delta_\emptyset$) à une unité de la catégorie concernée (valeur $\Delta^+_\emptyset$), même si la dissimilarité positionnelle est nulle. Bien sûr, lorsque l'on combine d_{pos} et d_{cat}, les écart positionnels et catégoriels s'ajoutant, on en vient d'autant plus rapidement à dépasser la valeur $\Delta_\emptyset$ et à préférer $u_\emptyset$ à une unité réelle.

Enfin, le choix de la valeur moyenne des dissimilarité plutôt que leur somme permet de s'abstraire du nombre d'annotateurs.

4.4.2 Entropie d'un alignement

L'entropie d'un alignement $\bar{a}$, notée $\bar{e}(\bar{a})$, est la valeur moyenne de l'entropie de ses alignements unitaires :

$$\bar{e}(\bar{a}) = \frac{1}{|\bar{a}|} \cdot \sum_{i=1}^{|\bar{a}|} \dot{e}(\dot{a}_i) \tag{6}$$

Nous faisons le choix de considérer la valeur moyenne des entropies unitaires plutôt que leur somme afin par exemple que l'entropie d'un jeu multiple d'annotations qui serait constitué de la duplication d'un jeu donné possède la même entropie que ce dernier et non pas son double.

4.5 Alignement idéal et mesure d'accord

Alignement idéal $\hat{a}$. Un alignement $\bar{a}$ d'un jeu d'annotation j est considéré comme idéal vis-à-vis d'une fonction de dissimilarité d_x donnée s'il minimise son entropie parmi tous les alignements possibles de j. Il est alors noté $\hat{a}$.

Entropie d'un jeu d'annotations $e(j)$. L'entropie d'un jeu d'annotations j, notée $e(j)$, pour une fonction de dissimilarité d_x donnée, est définie comme l'entropie de son ou de ses alignements idéaux $\bar{e}(\hat{a})$. Par prudence, nous sommes contraints de parler de « ses alignements idéaux » et non pas de son alignement idéal car, même si c'est peu probable, plusieurs alignements distincts peuvent minimiser l'entropie d'un jeu d'annotations.

Nous venons d'établir les deux définitions cruciales de notre approche, qui rendent compte en particulier de son caractère unifié. En effet, le choix de l'alignement idéal se fait sur la base de l'entropie, donc de la mesure d'accord (*cf.* ci-dessous) entre annotateurs, et, réciproquement et parallèlement, la mesure d'accord se fait sur la base de l'alignement idéal.

Corpus : un corpus c est un ensemble donné de textes et l'ensemble des jeux d'annotations relatifs à ces textes.

Entropie aléatoire $e_{aleatoire}$. Pour un corpus c donné, soit P l'ensemble des processus aléatoires d'annotation actuellement disponibles[4]. $\forall p \in P$, soit $eAvg(p)$ la moyenne des entropies obtenues sur un ensemble significatif de jeux d'annotations produits par p. L'entropie aléatoire de ce corpus, notée $e_{aleatoire}(c)$, est définie comme $min(\{eAvg(p)/p \in P\})$. C'est une valeur qui sera susceptible de s'améliorer (diminuer) au fil des avancées en termes de génération aléatoire astucieuse.

Mesure d'accord. La mesure d'accord inter-annotateurs d'un jeu d'annotations est alors donnée par :

$$\forall j \in c, accord(j) = \frac{e_{aleatoire}(c) - e(j)}{e_{aleatoire}(c)} \tag{7}$$

Si les annotateurs sont parfaitement d'accord, comme dans le cas idéal évoqué au début de cet article, l'entropie résultante est nulle, si bien que la mesure d'accord est égale à 1. Au contraire, si les annotateurs ne font pas mieux que le hasard, leur taux d'accord est nul, voire négatif.

[4]Nous proposons deux tels processus de génération aléatoire en section suivante.

La méthode proposée peut être qualifiée d'holiste dans la mesure où c'est la considération de l'ensemble des annotations qui permet de déterminer les alignements unitaires. Il est impossible de partir d'alignements unitaires « sûrs » pour constituer, de façon ascendante, l'alignement idéal complet.

5 Opérationalisation : vers une méthode d'implémentation de l'alignement et de la mesure d'accord holistes

Pour toute la suite de cet article, nous allons utiliser la dissimilarité positionnelle $d_{pos-sporarique}$, et une dissimilarité catégorielle définie par une matrice remplie de 1, à l'exception de la diagonale qui est, comme toujours, nulle. Son rôle est ici limité à l'interdiction des couplages entre unités de catégories distinctes, afin de limiter les phénomènes entrant en jeu dans le cadre de cet article. Enfin, nous combinons ces deux dissimilarités en les sommant via $d_{combi(d_{pos-sporadique},d_{cat},1,1)}$.

Les définitions que nous venons de poser ont valeur d'un point de vue théorique, mais leur implémentation informatique pose un important problème de complexité, en raison du caractère holiste de la méthode proposée. Un parcours de toutes les possibilités est en effet inenvisageable, l'espace de recherche étant minoré par $(p!)^{n-1}$. A titre d'illustration, mentionnons que pour 5 annotateurs ayant simplement annoté chacun 5 unités, cette minoration est de $(5!)^4 = 120^4$, soit plus de 207 millions.

Nous allons établir des principes permettant de réduire cet espace de recherche de façon à obtenir une méthode utilisable avec des jeux de données réels, tels que 4 annotateurs ayant chacun annoté une centaine d'unités.

5.1 Une réduction de l'espace de recherche

Parmi les innombrables possibilités d'alignements qu'offre l'espace de recherche, une immense majorité reposent sur des alignements unitaires improbables. Nous allons démontrer qu'il est possible d'éliminer un grand nombre d'entre eux sans écarter l'alignement idéal.

En effet, considérons l'alignement idéal $\hat{a}$, de cardinalité m. Soit $\dot{a}$ l'un quelconque de ses alignements unitaires. Par commodité, nous lui donnons l'indice 1 ($\dot{a} = \dot{a_1}$), les autres ayant donc les indices de 2 à m. Cet alignement unitaire $\dot{a}$ contient n unités (réelles ou $u_\emptyset$). Pour chacune de ces unités u_i (avec $1 \leq i \leq n$), créons l'alignement unitaire $\dot{a}_{m+i} = (u_i, u_\emptyset, ..., u_\emptyset)$ de cardinalité n. Il est possible de créer un alignement $\bar{a}$ constitué de l'ensemble des alignements unitaires de $\hat{a} \setminus \{\dot{a}\}$, auquel on ajoute les alignements unitaires $\dot{a}_{m+1}$ à $\dot{a}_{m+n}$ que l'on vient de créer[5]. Il est de cardinalité $m + n - 1$. On a, du fait que $\hat{a}$ minimise l'entropie :

$$\bar{e}(\hat{a}) \leq \bar{e}(\bar{a}) \Rightarrow \frac{1}{m}\sum_{i=1}^{m}\dot{e}(\dot{a}_i) \leq \frac{1}{m+n-1}\sum_{i=2}^{m+n}\dot{e}(\dot{a}_i) \Rightarrow \sum_{i=1}^{m}\dot{e}(\dot{a}_i) \leq \frac{m}{m+n-1}\sum_{i=2}^{m+n}\dot{e}(\dot{a}_i) \leq \sum_{i=2}^{m+n}\dot{e}(\dot{a}_i)$$

$$\Rightarrow \dot{e}(\dot{a_1}) \leq \sum_{i=m+1}^{m+n}\dot{e}(\dot{a}_i)$$

et comme $\forall i > m, \dot{a}_i = \Delta_\emptyset$, et que l'on a posé $\dot{a} = \dot{a_1}$,

$$\Rightarrow \dot{e}(\dot{a}) \leq n \cdot \Delta_\emptyset \tag{8}$$

On a donc majoré l'entropie unitaire de tout alignement unitaire candidat à l'alignement idéal. Ainsi, sur exemple réel, pour 3 annotateurs créant chacun 25 unités, on passe d'environ 19000 à 1000 alignements unitaires.

5.2 Un algorithme rapide pour l'obtention d'une solution approchée

Une fois l'ensemble des alignements unitaires (restreint aux seuls alignements susceptibles d'appartenir à la solution, *cf.* point précédent) généré, trions ce dernier, et obtenons ainsi la liste $L_{initial}$. L'algorithme 1 permet

[5] $\bar{a}$ est bien un alignement, puisque chacune des unités apparaît dans un et un seul alignement unitaire.

d'obtenir une solution approchée de la recherche d'un alignement idéal. Sa complexité observée est telle qu'il est utilisable en des temps raisonnables, comme le montre le tableau 2.

Algorithm 1 Algorithme rapide pour une solution approchée

Require: $L_{initial}, L, L^-$ des listes, $i \in \mathbb{N}$, $\dot{a}$ un alignement unitaire
1: $L \leftarrow L_{initial}$
2: $i \leftarrow 0$
3: **while** $i < size(L) - 1$ **do**
4: $\quad \dot{a} \leftarrow L\,[i]$
5: $\quad L^- \leftarrow L\,[i+1, (size(L)-1)]$
6: $\quad$ Retirer de L^- tous les alignements contenant (au moins) l'une des unités de $\dot{a}$
7: $\quad i \leftarrow i + 1$
8: **end while**

	Nb. d'annotateurs	Nb. d'unités par annotateur	Espace de recherche (alignements unitaires après filtrage)	Temps d'exécution
Cas 1	3	25	1145	Instantané
Cas 2	3	100	4347	$<$ 1 sec.
Cas 3	4	100	38624	5 sec.
Cas 4	4	200	69994	16 sec.
Cas 5	5	25	96794	9 sec.

TAB. 2 – Temps d'exécution de l'algorithme 1

Les alignements résultant de cette méthode ont été évalués qualitativement, par observation graphique de sorties telles que sur la figure 1. Si l'on observe localement quelques croisements, on notera toutefois qu'ils sont rares et de faible amplitude. Nous considérons que le degré de précision obtenu sur le calcul d'entropie qui en résulte suffira aux expériences menées dans la présente étude.

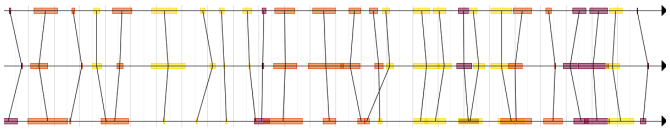


FIG. 1 – Exemple de résultat d'alignement avec l'algorithme 1 (un annotateur par ligne)

5.3 Vers une valeur minimale de $e_{aleatoire}$

En l'absence provisoire de l'alignement idéal, nous nous contenterons, pour la suite de cette étude, de l'approximation obtenue ci-dessus. Il reste à calculer la mesure d'accord, c'est-à-dire à situer l'entropie des annotations observées par rapport à ce que pourrait produire le hasard. Pour ce faire, nous chercherons à simuler le travail d'un observateur judicieux qui annoterait en aveugle un nouveau texte (en connaissant sa taille, mais sans pouvoir le lire), en s'inspirant des annotations faites par d'autres sur d'autres textes et en cherchant minimiser l'entropie.

Une première stratégie (*random1*) de génération aléatoire d'annotations a consisté, tout simplement, à observer des données réelles, en tenant compte du nombre (plages) d'annotations par annotateur, puis à générer des annotations conformes à ces observations, et de taille aléatoire. Il nous a toutefois semblé qu'un hasard bien entendu devrait prendre en compte des régularités plus fines observables sur un corpus donné, telles que (et de façon non limitative) : telle catégorie conduit à un pavage complet du texte, telle autre donne lieu à des tailles relativement stables, les textes commencent ou se terminent toujours par tel type particulier... La stratégie *random2* vise à prendre en compte de telles régularités. Elle consiste à assembler, au sein d'un même texte virtuel, des annotations émanant de textes réels différents, en respectant la règle suivante : pour générer une annotation virtuelle de n annotateurs, on utilise n textes différents dans chacun desquels on puise les annotations d'un et d'un seul annotateur[6]. Nous obtenons ainsi un jeu d'annotations qui respecte par construction la distribution des unités dans

[6]Une méthode permettant de pallier les différences de taille entre les différents textes a été mise en place.

les textes (régularité de position, de quantité...). Il est d'autre part effectivement aléatoire, puisque les annotations sont assemblées en aveugle : elles sont issues de textes sans rapport entre eux et ni avec le texte virtuel[7]. L'expérience rapportée ci-dessous confirme la supériorité de cette stratégie *random2* : nous obtenons en effet des tirages aléatoires ayant une entropie plus faible que celle provenant d'un tirage réalisé selon *random1* (3.48 au lieu de 3.67 pour un maximum possible de 4, soit 57 % de désordre en moins).

6 Observations expérimentales

6.1 Expérimentations sur un jeu de données factices

Afin de disposer d'une grille de lecture permettant de savoir à quoi correspondent les valeurs d'accord situées entre 0 et 1 (1 étant l'accord parfait, et 0 ce qu'est capable de faire le hasard), nous avons procédé à l'établissement empirique de deux courbes correspondant à deux modes de lecture parallèles. En premier lieu, nous cherchons à savoir comment fluctue le taux d'accord d'un jeu de données à partir d'un état parfait vers un état de plus en plus dégradé du seul point de vue du placement des unités. Pour ce faire, un algorithme crée aléatoirement les annotations d'un premier annotateur (25 unités réparties aléatoirement et de tailles aléatoires), puis crée des annotations pour deux autres annotateurs par dégradation du premier jeu d'annotation selon un facteur k, selon le principe suivant : chaque unité de l'annotateur 1 est dupliquée pour chacun des annotateurs 2 et 3, et modifiée aléatoirement par translation de chacune de ses deux bornes d'un vecteur compris entre $-\frac{k}{2}$ et $+\frac{k}{2}$ fois la longueur de l'unité dupliquée. En second lieu, nous nous intéressons aux fluctuations relatives à un jeu de données possédant de plus en plus de faux négatifs. Cette fois, l'algorithme crée les annotations des annotateurs 2 et 3 en dupliquant chacune des unités de l'annotateur 1 avec une probabilité p d'oubli (faux négatif). Avec $p = 0.5$, il y aura en moyenne une unité sur deux non présente dans les jeux 2 et 3 par rapport au jeu 1. Par contre, les entités conservées restent parfaitement alignées avec les originales. Les fluctuations obtenues sont reportées dans les graphes suivants.

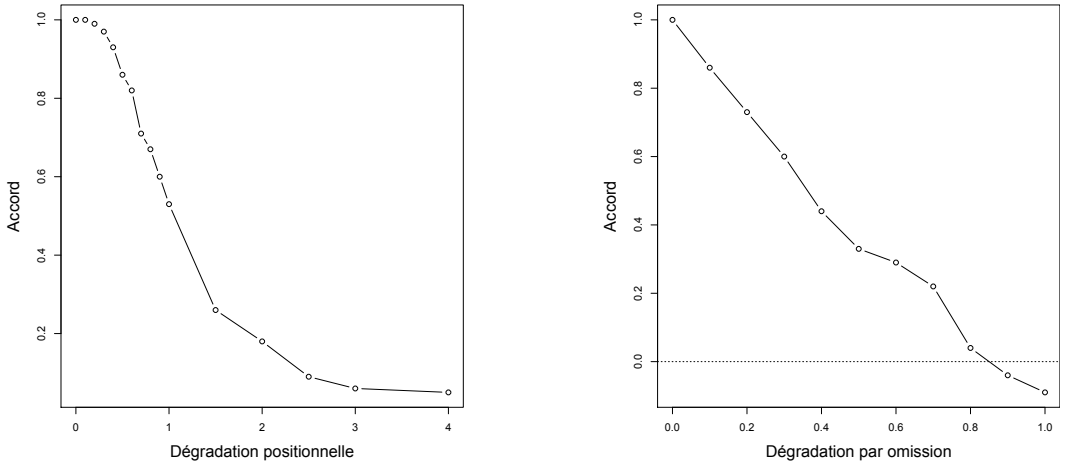


FIG. 2 – Évolution du taux d'accord selon la dégradation appliquée

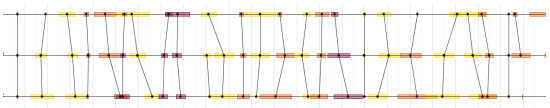
Il est ainsi possible de voir quel est l'équivalent d'une valeur d'accord donnée soit en termes de dégradation positionnelle, soit en termes d'oublis (faux négatifs). À titre d'illustration, la figure 3 donne deux exemples issus de ces deux paradigmes d'interprétation correspondant à la même valeur d'accord de 0.5.

6.2 Expérimentation sur un jeu de données réelles

Une première expérimentation sur des données réelles a été réalisée sur la base du corpus établi par (Labadié *et al.*, 2010), dont les annotations portent sur la segmentation thématique des textes, comportent plusieurs catégories, et comprennent des unités superposées au sein même des annotations d'un annotateur.

[7] Ajoutons que cette méthode permet de générer à moindre coût des données aléatoires en quantité importante. Pour a annotateurs et t textes ($t \geq a$), on peut en effet générer jusqu'à $C_t^a \cdot a^a$ combinaisons différentes.

Accord de 0.5 par dégradation positionnelle



Accord de 0.5 par dégradation par omissions

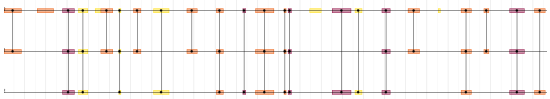


FIG. 3 – Illustrations d'un accord de 0.5

	Nb. annot.	e_{obs}	$e_{random1}$	a_1	$e_{random2}$	a_2
Texte 31	3	3.44	3.76	0.09	3.48	0.01
Texte 121	3	3.20	3.76	0.15	3.48	0.08
Texte 20	4	2.81	3.72	0.24	3.34	0.16
Texte 14	3	2.79	3.76	0.25	3.48	0.20
Texte 1	3	2.67	3.76	0.29	3.48	0.23
Texte 13	3	2.37	3.76	0.36	3.48	0.32
Texte 6	3	2.18	3.76	0.42	3.48	0.37

TAB. 3 – Résultats obtenus sur données réelles

Le tableau 3 rapporte les résultats obtenus sur 7 de ces textes (en utilisant la même dissimilarité que précédemment), en donnant pour chacun son entropie (e_{obs}), celles obtenues par $random1$ ($e_{random1}$) et par $random2$ ($e_{random2}$) pour le corpus correspondant, et les mesures d'accord respectives qui en résultent.

On constate que $random2$ est sensiblement plus efficace que $random1$, faisant passer le meilleur accord de 0.42 à 0.37, et que la baseline ainsi obtenue est relativement sévère dans la mesure où dans le cas le pire, l'accord obtenu n'est que de 0.01, soit d'un niveau équivalent à $random2$. Une observation visuelle, donc qualitative, des 7 textes annotés, confirme l'ordre obtenu par le calcul. Mentionnons que le texte 6 donnant lieu au meilleur accord $a_2 = 0.37$ correspond respectivement à une valeur $k = 1.2$ (translation des bornes jusqu'à ± 0.6 fois la largeur des unités associées, mais aucune omission) et $p = 0.46$ (omission d'une unité dans 46% des cas, mais positionnements parfaits) de nos deux grilles de lecture.

7 Conclusions et perspectives

Nous avons proposé une méthode réalisant alignement et mesure d'accord dans un processus commun. Très générale, cette méthode autorise les variations positionnelles et catégorielles, et ne dépend pas de la taille des entités annotées. Elle peut être configurée pour chaque paradigme d'annotation, par le choix des formules de calcul de la dissimilarité positionnelle et par l'indication des similarités entre catégories. Nous la disons holiste dans la mesure où elle s'appuie sur l'intégralité des données pour faire des choix, là où des méthodes fenêtrées comme WindowDiff opèrent localement. Par ailleurs, elle n'exige pas de jeu d'annotations de référence pour fonctionner ou s'amorcer. Nous travaillons actuellement à la mise en place de méthodes permettant de calculer automatiquement les valeurs optimales de $\Delta_\emptyset$ et des coefficients des matrices catégorielles (pour ce second point, voir aussi (Fort *et al.*, 2010)). Un algorithme permettant d'approcher la solution optimale a été défini et nous avons procédé à une implémentation complète du système, avec rendu graphique des alignements effectués. Cette implémentation sera distribuée avec la prochaine version de la plate-forme Glozz (1.1.0), devenant ainsi publiquement accessible, et directement exploitable par les utilisateurs de cette plate-forme. Un travail est actuellement mené, en partenariat avec Jean-Philippe Métivier (GREYC), qui vise à calculer dans un temps raisonnable la solution optimale. À cette fin, un système à base de CSP (*Constraint Satisfaction Problem*) prend en entrée l'ensemble des alignements unitaires candidats à la solution. Nous pourrons ainsi d'une part quantifier la différence d'entropie entre la solution approchée et la solution idéale, et d'autre part proposer, mais dans un temps éventuellement beaucoup plus long, la solution idéale.

Remerciements

Nous tenons à remercier Jérôme Chauveau qui a récemment rejoint l'équipe de développement de Glozz et contribué aux développements nécessaires aux expérimentations dont il est fait mention ici.

Références

ARTSTEIN R. & POESIO M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, **34**(4), 555–596.

ASHER N. (1993). *Reference to Abstract Objects in Discourse*. Kluwer.

BESTGEN Y. (2009). Quel indice pour mesurer l'efficacité en segmentation de textes ? In *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

BOOKSTEIN A., KULYUKIN V. A. & RAITA T. (2002). Generalized Hamming distance. *Information Retrieval*, (5), 353–375.

CHAROLLES M. (1997). L'encadrement du discours : Univers, champs, domaines et espaces. *Cahier de Recherche Linguistique*, (6).

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46.

COHEN J. (1968). Weighted kappa : Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**(4), 213–220.

FORT K., FRANÇOIS C. & GHRIBI M. (2010). Evaluer des annotations manuelles dispersées : les coefficients sont-ils suffisants pour estimer l'accord inter-annotateurs ? In *Traitement Automatique des Langues Naturelles (TALN) Traitement Automatique des Langues Naturelles (TALN)*, p.Õ, Montréal France. Quaero.

GROUIN C., GALIBERT O., ROSSET S., QUINTARD L. & ZWEIGENBAUM P. (2011). Mesures d'évaluation pour entités nommées structurées. In *Évaluation des méthodes d'Extraction de Connaissances dans les Données*, Brest, France.

HEARST M. (1997). Texttiling : segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, **23**(1), 33–64.

KRIPPENDORFF K. (1980). *Content Analysis : An Introduction to Its Methodology*, chapter 12. Sage : Beverly Hills, CA.

KRIPPENDORFF K. (1995). On the reliability of unitizing contiguous data. *Sociological Methodology*, (25), 47–76.

LABADIÉ A., ENJALBERT P., MATHET Y. & WIDLÖCHER A. (2010). Discourse structure annotation : Creating reference corpora. In *Workshop on Language Resource and Language Technology Standards - state of the art, emerging needs, and future developments*, La Valetta, Malta : Conference LREC 2010.

LAMPRIER S., AMGHAR T., LEVRAT B. & SAUBION F. (2007). On evaluation methodologies for text segmentation algorithms. In , Ed., *Proceedings of ICTAI 2007*, p. 19–26 : .

MAKHOUL J., KUBALA F., SCHWARTZ R. & WEISCHEDEL R. (1999). Performance measures for information extraction. In *Proceedings of DARPA Broadcast News Workshop*, p. 249–252.

PÉRY-WOODLEY M.-P., ASHER N., ENJALBERT P., BENAMARA F., BRAS M., FABRE C., FERRARI S., HO-DAC L.-M., LE DRAOULEC A., MATHET Y., MULLER P., PRÉVOT L., REBEYROLLE J., TANGUY L., VERGEZ-COURET M., VIEU L. & WIDLÖCHER A. (2009). ANNODIS : une approche outillée de l'annotation de structures discursives. In *Actes de la 16e Conférence Traitement Automatique des Langues Naturelles (TALN'09), session posters*, Senlis, France.

PEVZNER L. & HEARST M. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, **28**(1), 19–36.

SCOTT W. (1955). Reliability of content analysis : The case of nominal scale coding. *Public Opinion Quarterly*, **19**(3), 321–325.

SIEGEL S. & CASTELLAN N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, 2nd edition.

TEUFEL S., CARLETTA J. & MOENS M. (1999). An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of Ninth Conference of the EACL*, p. 110–117, Bergen.

WIDLÖCHER A. & MATHET Y. (2009). La plate-forme Glozz : environnement d'annotation et d'exploration de corpus. In *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis, France : ATALA LIPN.