Évaluation des annotations : ses principes et ses pièges

Yann Mathet* — **Antoine Widlöcher****

* *Université de Caen Normandie, GREYC UMR6072, yann.mathet@unicaen.fr*
** *Université de Caen Normandie, GREYC UMR6072, antoine.widlocher@unicaen.fr*

RÉSUMÉ. *Beaucoup de données sont produites par le TAL (systèmes automatiques) et pour le TAL (corpus de référence, pour la linguistique computationnelle ou pour l'apprentissage), et leur mise à disposition ne devrait se faire que dans la mesure où leur consistance est établie. Si l'on peut se réjouir de l'effort grandissant qui est fait en ce sens depuis une vingtaine d'années, par exemple par l'utilisation de plus en plus fréquente de mesures d'accord inter-annotateurs telles que le coefficient kappa, on constate cependant qu'il ne s'accompagne pas toujours d'une connaissance suffisante des principes sous-jacents à l'évaluation, ni de la rigueur nécessaire à l'application de ces derniers.*
L'objectif de cet article est d'une part de présenter et de questionner les concepts et les principes fondamentaux du domaine (faut-il par exemple « corriger par la chance » les mesures d'accord, et si oui, comment ?), et d'illustrer par des exemples concrets et chiffrés les conséquences d'une pratique approximative de l'évaluation.

ABSTRACT. *A lot of data is produced by NLP (automatic systems) and for NLP (reference corpus, for computational linguistics or for machine learning) and should be publicly released only if their consistency is proven. While the growing effort that has been made in this direction over the past two decades is encouraging, for example through the increasing use of inter-annotating agreement measures such as kappa, it is not always accompanied by sufficient knowledge of the principles underlying evaluation or the rigor required for their application. The aim of this paper is to present and question the basic concepts and principles of the domain (e.g., shall we use "chance correction" in agreement measures, and if so, how?), and to illustrate with concrete and quantified examples the consequences of an approximate practice of evaluation.*

MOTS-CLÉS : *accord inter-annotateurs,* gold standard*, évaluation d'annotations.*

KEYWORDS: *inter-annotator agreement, gold standard, annotation evaluation.*

1. Introduction

Le point de rencontre entre des préoccupations de nature proprement scientifique et des questionnements relevant de l'éthique est difficile à cerner. Abordant l'objet scientifique en termes éthiques, on pense immédiatement à des éléments situés soit en amont soit en aval de la démarche scientifique proprement dite, dans la sélection de ses objets, dans ses conséquences, dans les moyens qu'on y consacre. Mais au-delà de ces éléments en quelque sorte extérieurs à la démarche scientifique elle-même, une question demeure centrale d'un point de vue éthique : celle de la manière dont une communauté scientifique élabore les normes à l'aune desquelles la validité d'une production scientifique sera jugée. Et cette question impose à son tour une réflexion sur les moyens qu'on se donne pour vérifier l'adéquation entre les solutions mises en œuvre et les problèmes posés, pour vérifier que ces solutions sont adoptées pour de bonnes raisons (et non simplement par exemple parce qu'elles sont déjà sur le devant de la scène ou facilement disponibles). Plus précisément, dans cette perspective, le champ du scientifique rejoint selon nous celui de l'éthique, quand interviennent notamment des interrogations relatives à 1) la qualité des ressources produites, 2) l'impartialité dans l'interprétation des résultats et l'évaluation de ressources, 3) la transparence sur le cheminement suivi pour la constitution des données et sur leur interprétation.

Adopter un point de vue éthique sur nos disciplines, en TAL ou en linguistique computationnelle, conduit dès lors à accorder une place centrale à la question de la constitution de corpus annotés, à celle de leur évaluation et à celle de leur exploitation. À vrai dire, ce problème est si indissolublement lié à la validité de la démarche scientifique elle-même qu'on peut s'interroger sur cette dimension proprement éthique de la question. Elle n'échappera toutefois à cette dimension qu'une fois les mauvaises pratiques véritablement marginalisées. Certes, la question de l'évaluation de la qualité des données reçoit une attention grandissante depuis une quinzaine d'années dans notre communauté, et il est aujourd'hui presque impensable de publier des ressources sans que celles-ci ne s'assortissent de résultats chiffrés censés rendre compte de leur validité. À cette fin, un certain nombre de mesures d'accord, provenant souvent d'autres domaines que le TAL, sont employées, et leur usage pour les besoins de nos disciplines a déjà été largement décrit, notamment dans l'article de référence (Artstein et Poesio, 2008). Pour autant, ce consensus n'a pas encore conduit à la mise en place d'une réflexion suffisamment aboutie sur les moyens à mettre en œuvre, et beaucoup d'efforts restent à faire, tant pour faire mieux connaître et mieux appliquer les méthodes existantes, que pour comprendre leurs limites et les améliorer. Dans cet article, nous essayons de montrer les conséquences, en termes d'évaluation, d'un mauvais usage ou d'une mauvaise compréhension des mesures disponibles. Nous verrons que si la facilité d'accès de certaines métriques et leur popularité dans la communauté prévalent parfois sur la prise en compte de leurs champs d'application respectifs, l'intelligibilité des résultats d'évaluation peut s'en trouver sévèrement compromise. À cette double contribution mettant en lumière des points peu abordés frontalement dans la littérature, nous verrons que s'ajoute une réflexion sur le statut de la chance dans l'évaluation des accords inter-annotateurs, question centrale mais jamais étudiée de façon

systématique à notre connaissance. Cette nécessaire clarification, qui concerne notre communauté dans son ensemble, est une tâche complexe, qui comporte encore des questions ouvertes, comme en témoignent par exemple les débats constants entre les partisans de différentes mesures d'accord depuis des dizaines d'années (cf. l'emblématique débat entre Berry (1992) et Goldman (1992), ou les remarques de Di Eugenio et Glass (2004)).

Cet article ne peut bien sûr pas prétendre faire le tour de la question. Notre objectif est essentiellement de présenter les principaux concepts et méthodes de l'évaluation, et de montrer, de façon non exhaustive mais à chaque fois chiffrée, les conséquences liées à de mauvais choix ou à une utilisation approximative des outils disponibles[1]. Il est organisé de la façon suivante.Quatre concepts premiers sont présentés dans la section 2 : la notion d'annotations de référence, d'annotations multiples, d'accords inter-annotateurs et de validité. Les sections 3 et 4 se focalisent ensuite sur les mesures d'accord inter-annotateurs. La section 3 concerne la prise en compte par ces mesures de la part d'accord due à la « chance ». Si elle est nécessaire, elle reste difficile à formaliser, et controversée. La section 4 illustre les conséquences de l'utilisation détournée des mesures classiques pour tenter de les adapter à des configurations pour lesquelles elles ne sont pas conçues. Elle montre la nécessité de connaître les mesures adaptées aux différentes configurations, et aussi la nécessité d'en développer de nouvelles. La section 5 aborde la question de la nature des catégories : sont-elles toujours indépendantes les unes des autres ? Quelles conséquences en tirer ? La section 6 s'intéresse à différentes stratégies de constitution de corpus, et à leurs conséquences respectives. Enfin, la dernière partie de l'article concerne l'évaluation des systèmes, ses différences avec l'évaluation des annotations multiples, et les conséquences de la confusion que l'on constate parfois entre les deux.

2. Principaux concepts

Avoir une bonne pratique de l'évaluation passe tout d'abord par la compréhension de ses principes. Si cette affirmation semble aller de soi, on constate pourtant, dans différents travaux, une confusion parfois importante entre annotations de référence, annotations produites par des annotateurs, mesure de la validité et mesure d'accord. Nous consacrons donc cette première partie à la définition de ces concepts ainsi qu'aux liens entre ces derniers (en tentant de limiter la circularité).

2.1. *Continuum, annotation, localisation et caractérisation*

Commençons par préciser ce que nous entendrons ici par annotation (voir aussi (Fort, 2012)). Visant des travaux d'annotation en linguistique et en TAL, nous devons

1. Les exemples de mauvais usages des méthodes d'évaluation mentionnés dans cet article ont été constatés dans des publications avec comité de lecture, même si, pour des raisons évidentes, nous ne citons pas ces dernières.

bien entendu considérer avant tout des processus visant à caractériser des occurrences de phénomènes rencontrées en corpus. On suppose donc donné un continuum unidimensionnel textuel, audio ou vidéo. Ce continuum constitue le contexte qui permet à l'annotateur de déterminer l'occurrence et la valeur des phénomènes qu'il observe. Le processus d'annotation se ramène à deux phases qui ne sont pas systématiquement à la charge de l'annotateur : la localisation des occurrences au sein du continuum, et l'association aux occurrences de représentations permettant de les caractériser.

Pour un grand nombre de tâches d'annotation, le résultat de la première phase est fourni aux annotateurs, qui doivent alors simplement s'acquitter de la seconde . Ainsi, par exemple, certains verbes ont été préalablement identifiés, auxquels l'annotateur devra affecter une catégorie sémantique. Pour d'autres travaux, l'annotateur doit au contraire prendre en charge lui-même la délimitation des segments à annoter. Suivant Krippendorff (1995), nous parlerons d'*unitizing* pour désigner cette délimitation.

La caractérisation des segments occurrences peut prendre des formes très variées : association d'un *tag* choisi parmi un ensemble prédéfini et très limité, association d'un *tag* choisi librement, association d'une représentation structurée complexe (structure de traits par exemple), commentaire en texte libre... Dans cet article, nous nous en tenons au cas le plus simple, mais aussi le plus fréquent, en nous limitant à l'association aux occurrences de catégories choisies parmi un ensemble prédéfini pour laquelle nous parlerons simplement de catégorisation. Toutefois, on ne négligera pas la variété des rapports que les différentes catégories entretiennent entre elles, pour rendre justice au fait que les catégories ne sont pas toujours à égale distance les unes des autres et au fait que, conséquemment, les erreurs d'affectation ou les désaccords n'auront pas tous la même gravité.

2.2. *Référence ou* gold standard

La linguistique computationnelle ainsi que le TAL ont souvent besoin de disposer de données établies afin de servir de base à la vérification d'hypothèses ou à l'évaluation de différents systèmes automatiques. On appelle ces annotations la référence, ou *gold standard*. Cette référence est supposée refléter la « réalité » du phénomène linguistique visé, ou, à défaut de pouvoir atteindre cette réalité, refléter la compréhension consensuelle qu'en a la communauté, à un moment donné. C'est à l'aune de cette référence que la validité d'une annotation sera jugée, validité définie idéalement comme adéquation à la réalité. Comme l'indique Krippendorff (2013a), il est important que la référence rende compte de l'ensemble des phénomènes étudiés, ce dont il est difficile de s'assurer lorsque l'on a affaire à une tâche inédite.

2.3. *Annotations manuelles multiples*

L'évaluation d'une annotation impose la disponibilité d'une référence, mais lorsque l'on s'intéresse à une nouvelle tâche d'annotation, on ne dispose encore ni

de cette référence, ni même souvent d'experts pouvant la constituer. On ne sait même pas encore si la tâche est définie de façon suffisamment formelle (éviter les catégories floues) et consistante (pour une entrée donnée en contexte, une seule sortie possible) pour qu'une telle référence puisse un jour exister. Une pratique courante pour tout à la fois en vérifier la faisabilité et en obtenir une est de pratiquer une annotation manuelle multiple d'un même corpus : chaque annotateur annote de façon indépendante des autres annotateurs, avec comme hypothèse sous-jacente que si tous les annotateurs produisent (à peu près) les mêmes annotations, c'est-à-dire sont largement d'accord, c'est qu'il y a de grandes chances que ces dernières soient valides, et puissent donc, moyennant une éventuelle concertation finale pour les cas de désaccord, constituer une référence. L'étude de cette hypothèse sous-jacente est l'un des aspects importants du présent article. Notons que depuis une dizaine d'années, la montée en force du *crowdsourcing* permet de faire annoter un grand nombre de personnes non spécialistes avec l'idée que le grand nombre d'annotateurs puisse compenser leur manque d'expertise, et qu'en utilisant des techniques telles que le maximum *a posteriori* on puisse obtenir une référence de qualité. Si nous saluons l'idée d'utiliser un nombre important d'annotateurs, qui va dans le sens du principe même de l'annotation multiple, nous mettons la réserve suivante : ce n'est que dans le cas, invérifiable *a priori*, où les annotateurs ajoutent du bruit à l'annotation parfaite que l'on peut retrouver cette dernière par le calcul.

2.4. *Mesure d'accord inter-annotateurs*

Le but d'une mesure d'accord est d'indiquer le degré de consensualité atteint par des annotateurs multiples lors de l'annotation de mêmes documents. Une telle mesure compare des annotations dont on ignore le degré de validité à d'autres annotations dont la validité est tout aussi inconnue. En ne comparant pas des données à une référence, une telle mesure ne peut jamais garantir la validité des données : être d'accord ne signifie pas forcément avoir raison. En revanche, il est généralement admis, voir par exemple (Krippendorff, 2013a), qu'une bonne valeur d'accord garantit un fort taux de reproductibilité, c'est-à-dire que si plusieurs annotateurs sont globalement d'accord sur leurs annotations d'une partie d'un corpus, alors chacun d'eux devrait produire à peu près la même chose que ses collègues sur les autres parties du corpus. De la sorte, il n'est plus guère besoin de faire annoter l'intégralité du corpus par plusieurs annotateurs, mais on peut affecter chaque partie à un annotateur particulier. Comme nous le verrons, il est important que ces mesures prennent en compte le rôle de la chance (notion définie plus loin) dans le calcul de l'accord.

2.5. *Mesure de la validité des annotations*

Dès lors que l'on dispose d'une référence, il est possible d'évaluer la validité des productions d'un système (ou d'un humain) par différentes mesures, sortes de distances entre la référence et les productions évaluées. Cela peut être le pourcentage de

réponses valides dans le cas d'une tâche de catégorisation d'items, des mesures telles que rappel, précision et f-mesure dans le cas d'une tâche d'identification d'éléments particuliers, ou des mesures plus complexes pour des tâches telles que la segmentation thématique, la détermination des relations du discours ou encore la constitution de chaînes de référence. Lorsqu'une telle mesure indique un score de 100 %, le système répond parfaitement à la tâche demandée (si la référence rend compte de l'ensemble des phénomènes étudiés). Un système obtenant un meilleur score qu'un autre peut être qualifié de globalement plus performant.

Contrairement aux mesures d'accord, les mesures de validité n'ont, pour la tâche qui est originellement la leur, aucun besoin de prendre en compte le rôle éventuel de la chance : la comparaison étant faite par rapport à une référence, le score atteint est bien celui que l'on peut attendre sur d'autres corpus, et rend donc bien compte des capacités effectives du système. En revanche, il est bien sûr souhaitable de disposer d'une *baseline*, qui indique l'apport de ce système par rapport à un comportement naïf ou par rapport à d'autres systèmes existants.

2.6. *Liens entre les différents concepts*

Ces concepts étant posés, il est important de comprendre leurs corrélations. Tout d'abord, la figure 1, librement inspirée de (Krippendorff, 2013a) et où le centre de la cible est une métaphore de l'annotation correcte, illustre la différence entre accord et validité : le désaccord limite la validité (cf. B), mais l'accord ne l'implique pas (cf. C). Elle montre aussi que l'absence de référence ne permet pas de distinguer des configurations pourtant très différentes (A *versus* B et C *versus* D).

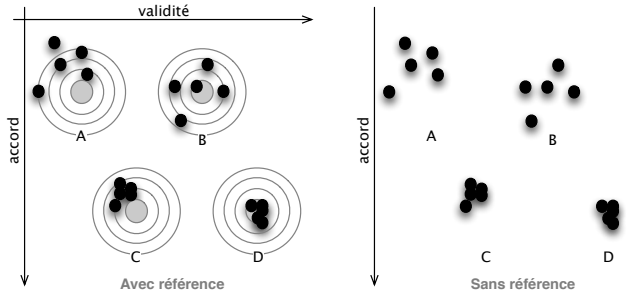

Figure 1. *Validité* versus *accord*

De façon plus générale, la figure 2 illustre les liens entre les quatre concepts. Dans le but d'obtenir une référence, il est possible et classique de procéder à une annotation multiple. L'indice de reproductibilité de cette dernière est estimé par une mesure d'accord inter-annotateurs. S'il est élevé, les annotations multiples serviront à établir la référence, *via* différentes stratégies possibles que nous verrons en dernière section (principe de la majorité, principe de l'unanimité, correction par un expert), illustrées

par la bulle « stratégie de constitution ». Cette référence est alors utilisée conjointement à une mesure de validité pour évaluer les productions d'un système. Insistons sur le fait que mesure d'accord et mesure de validité sont deux éléments distincts, tant dans leur fonctionnement que dans leurs entrées, point sur lequel nous reviendrons.

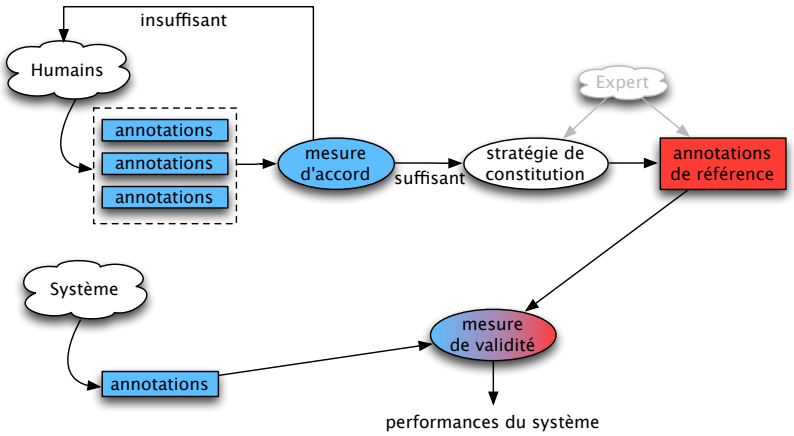

Figure 2. *Liens entre les différents concepts*

3. Correction par la chance : une nécessité et une difficulté

Une mesure de l'accord inter-annotateurs doit tenir compte de la part d'accord due à la « chance » (on parle de mesures « corrigées par la chance »). C'est un concept utilisé par nombre de mesures d'accord mais jamais clairement défini. Il ne s'agit pas d'un simple raffinement des méthodes de calcul, mais d'une condition nécessaire pour savoir dans quelle mesure des annotations multiples sont reproductibles.

Nous considérerons que la chance correspond à un accord entre deux annotateurs sans action maîtrisée d'au moins l'un d'entre eux : par exemple, l'un des annotateurs sélectionne par mégarde une autre catégorie que celle qu'il a choisie (et tombe « par chance » sur la catégorie retenue par l'autre, correcte ou non), ou les deux annotateurs font tous deux une erreur de jugement (la même ou non) menant au choix de la même catégorie (éventuellement la bonne catégorie, mais trouvée pour de mauvaises raisons). Nous parlerons d'accord fortuit, obtenu par hasard[2].

Ce que l'on cherche à calculer, c'est l'accord hors chance, c'est-à-dire la portion de la valeur d'accord qui provient de la faisabilité de la tâche, de la bonne compréhension de cette dernière, et de la bonne interprétation des données. En effet, imaginons deux annotateurs dont l'accord brut ne serait pas meilleur que celui qu'ils obtiendraient en

2. Dans une acception ordinaire, signifiant « sans logique apparente dans le contexte de la tâche d'annotation ».

jouant aux dés pour choisir leurs annotations : leur comportement serait totalement décorrélé de la tâche qu'ils ont à accomplir. Une mesure d'accord devrait donc leur attribuer le score de zéro. C'est un point délicat, qui a largement nourri la littérature des premières réserves de Feinstein et Cicchetti (1990) à la récente controverse entre Zhao *et al.* (2013) et Krippendorff (2013b), que nous allons étudier en détail. Comme nous le verrons, le niveau théorique où l'accord observé ne résulterait que de la chance (telle que nous venons de la définir) est couramment désigné « *expected agreement* » par les concepteurs de mesures d'accord. Le terme « *expected* », à forte connotation probabiliste, suggère que l'on peut alors assimiler les productions des annotateurs à des variables aléatoires. C'est une hypothèse forte, qui peut être intéressante pour tenter d'évaluer la valeur de l'accord par chance, mais qui ne rend pas forcément compte de la complète réalité de ce qu'est l'accord par chance : une action non maîtrisée ne se laisse pas forcément décrire par la notion d'aléatoire.

3.1. *Une irréfutable nécessité*

Souvent questionnée, cette conception de la mesure d'accord est même vigoureusement contestée par Zhao *et al.* (2013) qui affirment qu'elle suppose délibérément que les annotateurs sont malhonnêtes et annotent en partie au hasard. Avant même d'aborder la question de la chance, montrons tout d'abord un exemple où la valeur brute d'accord ne peut pas être interprétée directement, puisqu'elle ne peut pas être nulle : trois annotateurs annotent des items au moyen de deux catégories possibles A et B. Même si les annotateurs se concertaient pour essayer de ne jamais tomber d'accord, ils obtiendraient mécaniquement un accord minimal de 33,3 %, car si deux annotateurs sont en désaccord, le troisième n'a le choix que d'être en accord avec le premier ou le second. Le degré zéro de l'accord correspond donc ici à la valeur 0,33 et non à la valeur 0. Mais d'une façon plus générale, la prise en compte de la part d'accord résultant de la chance est nécessaire pour pouvoir interpréter les résultats : à partir du moment où les annotateurs ne sont pas en accord parfait, c'est qu'une partie de leurs annotations est produite de façon non maîtrisée, car ils ne peuvent être à la fois en désaccord et tous conformes à ce qui est attendu. Dès lors, il y a à chaque fois une certaine probabilité pour qu'ils tombent d'accord fortuitement, par exemple en faisant la même erreur simultanément. De façon assez parlante, s'il y a deux annotateurs et deux catégories, en supposant le degré théorique où les annotateurs agiraient aveuglément (par exemple sans lire le texte), ils arriveraient à un accord de 50 %. Dans un tel contexte, un accord brut de 75 % n'est situé qu'à mi-chemin entre le chaos et l'accord parfait. La prise en compte de la chance est donc indispensable : une campagne d'annotation ne peut s'appuyer sur des résultats d'accord brut.

3.2. *Principe premier : effectuer un changement de repère*

Les mesures d'accord effectuent tout d'abord une première mesure, que nous appelons ici valeur brute, de l'accord observé entre les différents annotateurs, désignée

A_o pour *observed agreement*. Dans le cadre le plus simple de la catégorisation d'items par deux annotateurs, il peut s'agir du pourcentage d'items sur lesquels les annotateurs ont choisi la même catégorie. Telle quelle, comme nous venons de le voir, sauf si elle est égale à 1, cette première valeur n'offre aucune indication interprétable quant au caractère reproductible du travail des annotateurs.

La prise en compte de la chance consiste à estimer la valeur A_e (pour *expected agreement*) supposée correspondre à la portion d'accord par chance, et effectuer un changement de repère tel que la valeur 0 indique une absence totale de corrélation entre les phénomènes étudiés et les annotations produites, et que la valeur 1 indique un accord parfait. Ce changement de repère est donné par l'équation 1 commune à toutes les mesures d'accord :

$$A = \frac{A_o - A_e}{1 - A_e} \quad [1]$$

Notons que certaines mesures d'accord, comme les α (Krippendorff, 2013a) ou γ (Mathet *et al.*, 2015) sont bâties à partir de calculs de désaccords plutôt que d'accords, et utilisent l'équation 2 qui est équivalente à l'équation précédente, en posant les désaccords D_o et D_e comme compléments respectifs des accords A_o et A_e. Compte tenu de cette équivalence, nous nous appuierons uniquement sur les valeurs A_o et A_e.

$$\left.\begin{array}{l} A_o + D_o = 1 \\ A_e + D_e = 1 \end{array}\right\} \Rightarrow A = 1 - \frac{D_o}{D_e} \quad [2]$$

Ce changement de repère peut être illustré par la figure 3. On constate que ce nouveau repère laisse inchangée la valeur 1 correspondant à l'accord parfait, mais déplace de façon plus ou moins importante la valeur 0, en la positionnant au niveau atteint par la « chance », c'est-à-dire, formellement, par A_e. Dans cette illustration, qui correspond au dernier exemple énoncé (deux catégories) on constate qu'alors que la valeur mesurée A_o est aux environs de 0,75 (aux trois quarts du segment [0 1] de l'axe brut), la valeur d'accord retenue A n'est qu'aux environs de 0,5 (vers le milieu du segment [0 1] de l'axe corrigé). Notons par ailleurs que des valeurs d'accord négatives sont formellement possibles, lorsque que A_o est inférieur à A_e, c'est-à-dire lorsque la mesure estime que les annotateurs ont fait moins bien que s'ils avaient procédé au hasard (en pratique, lorsqu'une part suffisante de désaccord systématique est présente, par exemple lorsqu'un annotateur inverse deux catégories). En réponse à la critique de

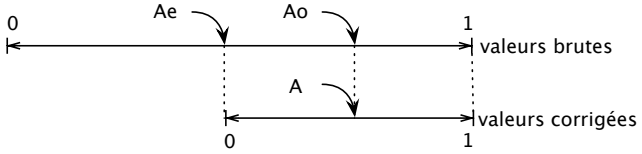

Figure 3. *La correction par la chance vue comme un changement de repère*

Zhao *et al.* (2013), indiquons tout d'abord que les mesures corrigées par la chance ne supposent pas un manque d'éthique de la part des annotateurs, qui confondraient leur

travail d'annotation avec une partie de dés : il est important de constater que la correction par la chance laisse invariante la valeur 1, c'est-à-dire que lorsque les annotateurs sont d'accord à 100 %, ces mesures estiment qu'ils sont parvenus à ce résultat sans que la chance n'intervienne (ce qui pourrait malgré tout arriver), c'est-à-dire de façon exemplaire. Ce n'est qu'à partir du moment où les désaccords se produisent que ces mesures tentent d'estimer la part d'accord fortuit (sans nécessairement supposer que les annotateurs jouent aux dés) qui peut en résulter.

3.3. *Estimer la part de « chance » : plusieurs conceptions, et un légitime débat*

Nous abordons à présent une partie cruciale et sans aucun doute la plus délicate de l'estimation de l'accord inter-annotateurs. Si on peut montrer, comme nous venons de le faire, la nécessité de prendre en compte une part de chance dans les valeurs brutes d'accord, il n'existe pour autant aucune certitude à ce jour sur la façon d'estimer cette dernière. À notre connaissance, aucune modélisation du comportement d'un annotateur n'est pour l'heure disponible, et nous doutons d'ailleurs qu'un tel modèle puisse correspondre à la diversité des annotateurs et des campagnes d'annotation. Pourtant, concevoir un modèle de la chance, c'est, en filigrane, faire des hypothèses sur la façon dont se comportent les annotateurs. Nous allons étudier et discuter les principales écoles. Notons que l'article de Krippendorff (2011) apporte un éclairage intéressant sur la question, tout en défendant la conception relative à ses propres mesures alpha.

3.3.1. *Approche équiprobable : ce que peut le hasard*

Cette approche, mise en œuvre dans la mesure S de Bennett *et al.* (1954) revient à considérer que l'on dispose d'une urne comportant une balle pour chaque catégorie disponible. Pour chaque item à annoter, on fait un tirage (avec remise) dans cette urne pour choisir une catégorie. L'objection majeure qui lui est faite est sa dépendance au nombre de catégories, avec la conséquence discutable que plus ce nombre augmente, plus la chance disparaît. Par exemple, l'ajout d'une catégorie inutile (jamais utilisée) à un modèle en comportant initialement deux fait passer la chance de 50 % à 33,3 %. Ce fait est d'après nous une conséquence d'une erreur conceptuelle plus grave de cette méthode : elle suppose que l'annotateur a un comportement schizophrénique. Pour chaque item, soit il annote avec sérieux et ne se trompe, dans ce cas, jamais, soit il annote au hasard en lançant un dé à n faces, n étant le nombre de catégories. En d'autres termes, il oscille entre l'annotateur parfait et l'annotateur le pire.

3.3.2. *Approches s'appuyant sur la distribution observée : une chance qui ne doit rien au hasard*

Dans cette conception de la chance, que nous nommerons distributionnelle, le principe du tirage est le même mais l'urne (ou les urnes, voir ci-après) comporte comme billes toutes les annotations effectuées par les annotateurs. Si les annotateurs utilisent la catégorie A plus fréquemment que la catégorie B, il y aura d'autant plus de billes A que de billes B. Elle traduit un comportement de l'annotateur qui nous apparaît plus

plausible : l'annotateur a ses défaillances, telles que la fatigue, l'ennui induit par une tâche répétitive, les habitudes qui en résultent (par exemple la plupart des items sont des A et non des B), ou tout simplement l'erreur d'interprétation des données, mais il essaye d'annoter correctement. De ce fait, lorsqu'il faillit, il est vraisemblable que l'annotation qui en résulte ne soit pas totalement décorrélée de sa tâche, mais que son (mauvais) choix résulte malgré tout des mécanismes qu'il emploie pour annoter correctement la plupart des autres items. Et l'on peut penser que ces mécanismes mènent à la distribution de catégories observée.

Cette conception de la chance qui est majoritaire parmi les mesures actuellement utilisées possède cependant deux versions qui s'opposent. Dans la première, que nous nommerons idiosyncratique, et sur laquelle repose notamment la mesure κ (Cohen, 1960), on considère que chaque annotateur dispose de sa propre urne correspondant à ses propres annotations, si bien que l'idiosyncrasie de chacun est préservée. Dans la seconde, que nous nommerons uniforme, et sur laquelle reposent notamment π (Scott, 1955) et α (Krippendorff, 1980), une seule urne est constituée, remplie par l'ensemble des annotations de tous les annotateurs, créant un annotateur virtuel moyen.

Le tableau 1 permet d'observer des différences typiques entre les conceptions équiprobable (*via* S), idiosyncratique (*via* κ) et uniforme (*via* π) de la chance, au moyen de quatre exemples. Dans les quatre cas, il y a six items à catégoriser *via* quatre catégories (de A à D), et les annotateurs sont d'accord uniquement sur deux d'entre eux (le premier et le dernier items), soit un accord brut toujours égal à 33,3 %.

Exemple 1

A	A	A	A	A	B
A	B	B	B	B	B

$\% = 0,333$
$S = 0,111$
$\kappa = 0,077$
$\pi = -0,333$

	A	B	
A	1	4	5
B	0	1	1
	1	5	6

Exemple 2

A	B	A	B	A	B
A	A	B	A	B	B

$\% = 0,333$
$S = 0,111$
$\kappa = -0,333$
$\pi = -0,333$

	A	B	
A	1	2	3
B	2	1	3
	3	3	6

Exemple 3

A	C	C	C	C	B
A	D	D	D	D	B

$\% = 0,333$
$S = 0,111$
$\kappa = 0,294$
$\pi = 0,077$

	A	B	C	D	
A	1	0	0	0	1
B	0	1	0	0	1
C	0	0	0	4	4
D	0	0	0	0	0
	1	1	0	4	6

Exemple 4

A	C	C	D	D	B
A	D	D	C	C	B

$\% = 0,333$
$S = 0,111$
$\kappa = 0,077$
$\pi = 0,077$

	A	B	C	D	
A	1	0	0	0	1
B	0	1	0	0	1
C	0	0	0	2	2
D	0	0	2	0	2
	1	1	2	2	6

Tableau 1. *Trois conceptions de la chance face à quatre exemples*

Le premier fait remarquable concerne S, qui contrairement aux autres ne tient pas compte du fait que seulement deux catégories sont utilisées par les annotateurs dans les exemples 1 et 2, et considère donc un accord identique de 0,111 dans les quatre exemples. Son score baisserait à – 0,333 dans les exemples 1 et 2, c'est-à-dire au niveau de π, si le modèle ne définissait que les deux catégories A et B.

Une différence importante entre les approches idiosyncratique et uniforme est illustrée par les exemples 1 et 2. Dans les deux cas, les annotateurs sont d'accord sur la qualification de deux items sur six, mais dans le premier exemple, l'un choisit essentiellement des A et l'autre essentiellement des B, alors que leurs choix sont équi-

librés dans le deuxième exemple. La distribution globale entre A et B est à chaque fois de 50 % / 50 %, ce qui mène à une valeur $\pi = -0{,}333$ identique. En revanche, l'importante différence de distribution de catégories dans l'exemple 1 augmente fortement κ, à 0,077 : les accords AA et BB étant, d'après les idiosyncrasies, moins probables, ils en deviendraient plus fiables. Le même phénomène se retrouve dans les exemples 3 et 4. On constate d'après ces deux couples d'exemples que κ est d'autant plus généreux (exemples 1 et 3) que la matrice de confusion n'est pas symétrique par rapport à sa diagonale. Nous reprenons à notre compte l'argument de Zwick (1988) qui est qu'à accord égal, κ rétribue le fait que les annotateurs soient en désaccord sur leurs distributions, et celui de Krippendorff (2011) qui est que ce coefficient est hybride, avec A_o bâti sur l'accord, et A_e bâti sur la corrélation, ce qui n'en fait pas une mesure adaptée à l'accord. Les mesures π et α reposent en revanche sur l'idée que les annotateurs sont interchangeables. Les conceptions idiosyncratique et uniforme deviennent équivalentes lorsque les annotateurs ont les mêmes distributions, mais l'idiosyncratique est plus optimiste, sans que l'on puisse asseoir cet optimisme sur des faits objectifs, lorsque celles-ci divergent.

3.4. *La question de la dépendance aux prévalences des catégories*

Un débat au long cours concerne la sensibilité des conceptions distributionnelles de la chance à la distribution des catégories (notamment autour de kappa, mais il concerne de façon plus générale toutes les mesures corrigées selon la distribution observée). Initialement, dans le domaine médical, Feinstein et Cicchetti (1990) ont relevé le paradoxe d'un kappa parfois faible malgré un fort taux d'accord observé, qui est dû à un fort déséquilibre dans la distribution des catégories (leurs prévalences respectives). Ce paradoxe a été repris dans le domaine de la linguistique computationnelle par Di Eugenio et Glass (2004).

Reprenons un exemple médical, assez parlant, celui d'une maladie rare, qui touche une personne sur 1 000. Deux médecins, traitant indépendamment les mêmes 10 000 cas, comparent leurs diagnostics : ils sont d'accord sur 99,85 % des patients, comme détaillé dans le tableau 2. Leur kappa est pourtant de seulement 0,5 (0,499 pour être précis), donc jugé médiocre. À quelle valeur se fier ?

	+	–	
+	5	5	10
–	5	9 985	9 990
	10	9 990	10 000

Tableau 2. *Forte prévalence d'une catégorie dans le cas d'une maladie rare*

Dans le détail donné dans le tableau 2, on constate que les médecins sont d'accord sur 9 985 cas qu'ils ont jugés sains, ainsi que cinq cas qu'ils ont jugés malades, mais il sont en désaccord sur cinq cas que l'un a jugé sains et l'autre malades, et cinq autres cas que l'un a jugé malades et l'autre sains. Quel patient serait rassuré d'être

jugé sain par l'un ou l'autre de ces médecins, qui ne sont d'accord que sur la moitié des patients supposément atteints par la maladie rare ? La valeur de kappa relativement faible correspond en fait, intuitivement, à une focalisation quasi totale de la mesure sur la catégorie rare, vu que les accords sur la catégorie dominante sont considérés, par son modèle de la chance, comme quasiment acquis : un médecin sans scrupule qui se contenterait, sans faire son travail de diagnosticien mais en se fondant sur la rareté de la maladie, de dire que tous les patients sont sains, obtiendrait un score de 99,9 % de réponses correctes, ce qui constitue un paradoxe inverse. La véritable qualité du diagnosticien, dans un tel cas, est d'identifier correctement les quelques cas rares, comme l'analyse Krippendorff (2013a). Dans notre exemple extrême, la valeur kappa de 0,5 correspond à une capacité d'accord de seulement 50 % sur la catégorie rare, en oubliant complètement, certes, la capacité d'accord sur la catégorie dominante. Ainsi, ce « paradoxe de la prévalence » ne doit pas être considéré comme une simple aberration, mais rend compte d'une défaillance manifeste des annotateurs. Il a au moins le mérite de pointer une faille potentiellement importante dans le jugement de ces derniers.

Pour autant, il serait absurde de balayer cette question d'un revers de main. Une expérience très éclairante a été menée par Bonnardel (1996) qui consiste à simuler le comportement de médecins en fixant leur capacité à faire des vrais positifs, *i.e.* leur « sensibilité », et des vrais négatifs, *i.e.* leur « spécificité », toutes deux à 90 %, et en faisant varier la prévalence de la maladie recherchée. On constate que la valeur de kappa varie de 0,64 lors de l'équilibre des catégories (autant de sains que de malades) à une valeur tendant vers 0 lorsque le déséquilibre fait qu'une catégorie tend à disparaître, ce qui démontre bien une dépendance de cette conception de la chance à la prévalence des catégories pour une qualité fixée des annotateurs.

Il est donc important non seulement de connaître, mais de savoir apprécier la dépendance de cette conception de la chance à la prévalence. De notre point de vue, elle n'est pas fortuite, puisqu'elle permet de pointer des défaillances éventuellement graves et laissées invisibles par la valeur d'accord brut. En revanche, si, pour le responsable de campagne, la rareté d'une catégorie ne la rend pas pour autant plus importante, cette sanction pourrait être jugée injustifiée. C'est pourquoi il est important de ne pas s'arrêter à la lecture de la seule valeur d'accord, mais de regarder le détail de ce qui y a conduit, et en particulier le détail de l'accord catégorie par catégorie, comme le proposent par exemple Krippendorff *et al.* (2016).

3.5. *Une approche contre-distributionnelle de la chance*

Un troisième type d'approche est porté par Aickin (1990) puis Gwet (2012). Il a pour objectif de prendre le contre-pied des approches distributionnelles que nous venons de voir, dans le but d'inverser le comportement des mesures dans le cas des distributions déséquilibrées. Même si son usage est plus rare, il nous semble nécessaire de la faire entrer dans le débat. Cette approche revient à considérer : (1) que chaque item est soit facile (E pour « *easy* »), soit difficile (H pour « *hard* ») à annoter ; (2) qu'un annotateur ne se trompe jamais sur un item E ; (3) qu'il procède par

pure chance sur un item H (avec la différence entre les deux approches que les E et les H sont communs à tous les annotateurs pour Aickin, tandis que Gwet considère que chaque annotateur a les siens propres). Comme bien sûr on ignore *a priori* si un item donné est E ou H, s'ensuit un modèle probabiliste assez élaboré censé en estimer les quantités respectives (sans pour autant les localiser) à partir de l'ensemble des annotations produites. Non seulement la pertinence de ce modèle probabiliste n'est pas établie, mais, en amont, l'hypothèse première n'est aucunement démontrée et nous semble peu probable. D'une part la dichotomie E *vs* H est simpliste (pourquoi n'y aurait-il pas d'items moyennement « difficiles » ?), et d'autre part le comportement de l'annotateur qui en résulterait serait, comme pour l'approche équiprobable, de type schizophrénique : pour un E, il ne se trompe jamais (comportement idéal), pour un H, il lance un dé. De façon plus précise, sans détailler les calculs et en s'en tenant au cas simple de deux catégories ayant les distributions p = x et q = 1 – x, le coefficient AC_1 de Gwet en vient à considérer que A_e = 0,5 lorsque x = 0,5, et A_e tend vers 0 lorsque x tend vers 0 ou vers 1 (p ou q vaut 1 et l'autre 0). Cela revient à dire que AC_1 se confond avec les autres approches lorsque les catégories sont parfaitement équilibrées, mais réfute toute possibilité d'accord par chance lorsque le déséquilibre devient extrême, et il prend alors le contre-pied de ces dernières. Gwet justifie cela en disant que si les annotateurs s'orientent vers une catégorie particulière, c'est qu'ils le font de façon parfaitement déterministe (et que, en agissant au hasard, ils aboutiraient à une distribution p = q = 0,5). Pourtant, il est impossible, dans le cas où les annotateurs choisiraient tout le temps la même catégorie, de distinguer s'il s'agit d'une action en lien avec ce qui est annoté, ou s'il s'agit de l'équivalent de thermomètres cassés qui donneraient toujours la même température.

3.6. *Critique et questions ouvertes sur les conceptions de la chance*

Notre objectif premier, dans cette section, est d'ouvrir un débat plus que d'apporter des réponses. Nous avons avancé un certain nombre d'arguments en faveur de l'approche distributionnelle uniforme, qui nous semble mieux rendre compte que les autres du travail d'annotateurs sérieux et censés être interchangeables. Nous la conseillons donc au vu des méthodes actuellement disponibles, mais nous interrogeons aussi sur ses limites, avec un certain nombre de points à notre connaissance jamais abordés.

Première limite : cette approche nous semble de moins en moins justifiable au fur et à mesure que l'accord faiblit[3]. À l'extrême, lorsque $A = 0$, c'est-à-dire lorsque $A_o = A_e$, les annotateurs sont considérés comme agissant sans rapport avec la tâche demandée. Dès lors, leurs annotations ne reflètent plus guère la distribution réelle des catégories, contrairement à ce que suppose le calcul de A_e, comme on le voit par exemple dans (Krippendorff, 2011) : « sans connaissance de la catégorisation correcte

3. Les valeurs hautes des mesures d'accord sont souvent à bon droit regardées comme les plus importantes et devant être les plus précises, puisqu'elles affirment la reproductibilité des annotations. Cela pondère la gravité de cette première limite, caractéristique des valeurs faibles.

des unités, cette conception considère les distributions de catégories que les codeurs ont utilisées comme meilleure estimation de la population réelle des catégories ».

Deuxième limite : en se fondant sur les distributions observées, cette hypothèse exclut les cas (il est vrai impossibles à quantifier) où les annotateurs auraient par exemple fait une erreur de saisie, qui n'a aucune raison de respecter de telles régularités. C'est plus vraisemblablement l'ordre de présentation des catégories dans l'interface de saisie qui influera (l'imprécision d'un clic amenant à cocher sur un choix contigu).

Troisième limite : cette conception fait l'hypothèse implicite que les distributions sont homogènes au sein du corpus. Cela pose problème si cette homogénéité n'est pas garantie. Imaginons une campagne d'annotation de textes où les items sont prédéfinis et où il existe deux catégories A et B. Supposons qu'un texte annoté comporte deux pages, le même nombre d'annotations et, pour simplifier la discussion, le même accord brut $A_o = 0,9$ pour chacune des deux pages. En revanche, les distributions sont très différentes entre les deux pages : 50 % A et 50 % B pour la page 1, et 10 % A et 90 % B pour la page 2, soit, pour l'ensemble du texte, 30 % A et 70 % B. Il en découle les valeurs A_e respectives de 0,5, 0,82 et 0,58, et donc les accords corrigés par la chance A respectifs de 0,8, 0,44 et 0,76 (on suppose que les annotateurs ont les mêmes distributions, pour éviter un débat entre κ, π et α). De ces valeurs, on peut déduire que : (1) la page 2 était beaucoup plus facile à annoter que la page 1 (A_e à 0,82 *vs* 0,5) ; (2) d'un accord important sur la moitié du corpus (0,8 en page 1) et médiocre sur l'autre moitié (0,44) résulte un accord global important (0,76). L'affirmation (1) n'a rien d'évident : l'annotateur, en première page, utilise moitié de A et moitié de B, et, sans qu'on le prévienne de quoi que ce soit, utilise de lui-même 10 % de l'un et 90 % de l'autre en page 2. Pourquoi, lorsqu'il se trompe en page 2, le fait-il tout d'un coup avec ces nouvelles proportions ? Nous touchons là une limite essentielle de cette conception : la distribution en catégories des annotations résulte, et de façon imbriquée, pour partie de la tâche d'annotation (par exemple, dans telle campagne, les objets A sont globalement plus fréquents que les objets B), et pour partie du passage en cours d'annotation (d'un passage à l'autre la distribution des catégories « réellement » présentes peut varier fortement). Il y a donc un double niveau global et local, et ce modèle de la chance n'en considère qu'un. L'affirmation (2) nous semble confirmer un manque de consistance de ces mesures face à ces variations locales. Qualifier de reproductible le travail des annotateurs d'après ce qu'ils font sur un texte alors qu'on affirme le contraire en les observant sur une moitié de ce dernier est discutable. Les travaux sur la mesure γ ont conduit ses auteurs à aborder cette question dans (Mathet *et al.*, 2015)[4], mais la question reste encore ouverte. De notre point de vue, il reste nécessaire de faire évoluer ce modèle dans le cas de distributions variables. *A minima*, les mesures devraient s'assortir d'indications sur ce degré d'homogénéité des distributions au sein du corpus, pour mettre en garde le responsable de campagne sur une possible défaillance de la mesure effectuée.

4. En proposant de calculer A_e soit au niveau local (*i.e.* pour chaque document annoté), soit au niveau global (en recommandant ce dernier) lorsque c'est possible, c'est-à-dire lorsque plusieurs documents ont été annotés pour une même campagne.

Pour conclure, la question de la chance est une question difficile, débattue, mais on ne peut pour autant s'en affranchir au motif que cette question est encore ouverte, et s'appuyer sur des valeurs brutes nettement avantageuses. En l'état actuel des connaissances, nous conseillons les méthodes distributionnelles uniformes telles que π ou α, tout en ayant connaissance de leurs présupposés et de leurs faiblesses. Même si les valeurs qu'elles fournissent peuvent paraître désavantageuses dans le cas de distributions déséquilibrées, le principe de précaution nous fait préférer de fausses alertes à de fausses garanties. Le cas échéant, une analyse détaillée des désaccords (par catégorie) permettra de voir si les données sont malgré tout suffisamment fiables pour les attendus de la campagne.

4. Kappa, le marteau de Maslow de l'évaluation ? Le cas de l'*unitizing*

La « théorie de l'instrument », définie par Maslow (1966) comme la tentation qui consiste à travestir la réalité d'un problème en le transformant en fonction des réponses dont on dispose, semble parfois s'appliquer dans le domaine de l'évaluation en TAL. À la diversité des structures de données considérées, devrait correspondre autant de méthodes de calcul, mais en pratique on constate que kappa (et les mesures du même type), qui est conçu dans le cadre bien circonscrit de la catégorisation d'items prédéfinis, voit son utilisation étendue bien au-delà de ce pour quoi il est prévu.

C'est notamment le cas lorsque l'annotateur doit librement déterminer, sur un continuum (texte, audio, vidéo), où sont situées des unités (en définissant leurs frontières), quel en est le nombre, et pour chacune d'elles, quelle est sa catégorie. Seul Krippendorff (1995) s'est attaqué frontalement à la question en proposant des mesures spécifiques, et en baptisant ce domaine « l'*uniziting* », suivi plus récemment par Mathet *et al.* (2015). Dans l'ignorance de ces travaux spécifiques, une stratégie courante est de déformer l'objet d'étude afin qu'il puisse se conformer à ce que kappa prend en entrée, c'est-à-dire un ensemble d'items prédéfinis. Pour cela, le continuum est atomisé en éléments unitaires (par exemple les lettres, les mots, les phrases ou les paragraphes, selon le type de campagne), et chacun d'eux se voit attribuer la catégorie de l'unité dans laquelle il se trouve, le cas échéant. Si un atome n'est recouvert par aucune unité, il se voit attribuer la catégorie artificielle « vide » (notons que cette déformation devient délicate dans le cas où des unités se recouvrent). Mathet *et al.* (2015) ont montré les écueils que comporte une telle transformation. Sans les reprendre en détail ici, nous allons illustrer par un exemple combien ce type d'approche est à proscrire.

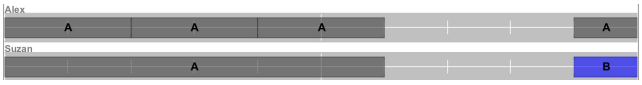

Figure 4. *Unitizing 1*

Les figures 4 et 5 reposent sur des continuums de longueur 10, annotés par deux annotateurs, Alex et Suzan. Dans le premier cas, Alex voit trois unités contiguës de

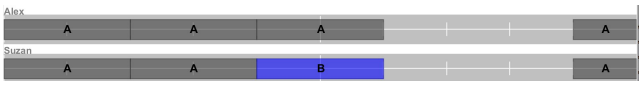

Figure 5. *Unitizing 2*

type A contre une seule pour Suzan, et ils sont en désaccord sur la catégorisation d'une unité à droite, A *versus* B. Dans le second cas, ils sont entièrement d'accord sur le nombre (4) et la position des unités, et divergent seulement sur la catégorisation de l'une d'elles (la troisième A versus B). Le tableau 3 indique les accords obtenus par deux méthodes : d'une part l'atomisation avec application de kappa, d'autre part la mesure γ spécifiquement développée pour l'unitzing. On constate que kappa juge excellente la configuration 1 (0,804), et même supérieure à la configuration 2 (0,648), car il y a un plus grand nombre d'accords sur les paires d'atomes, sans même se rendre compte qu'il y a des désaccords majeurs sur la détermination des unités (quatre unités contre deux). La mesure γ, qui prend en compte la nature réelle des unités sur le continuum, parvient naturellement à la conclusion contraire.

Il est donc fondamental que les utilisateurs connaissent parfaitement la finalité des mesures qu'ils utilisent, que l'utilisation détournée de méthodes soit proscrite, et qu'un effort soit fait pour développer et faire connaître les mesures adaptées à différents domaines (*unitizing*, segmentation, relations du discours, référence). C'est encore loin d'être le cas, le principal effort et les principaux débats, depuis plus de vingt ans, concernent la catégorisation d'items prédéfinis.

Expérience	Mesure kappa après atomisation	Mesure γ
unitizing 1	0,804	– 0,03
unitizing 2	0,648	0,69

Tableau 3. *Biais de l'atomisation d'un continuum*

5. Des catégories aux frontières perméables

5.1. *Des mesures « pondérées » pour des catégories contiguës*

Des responsables de campagne sont parfois étonnés par des valeurs d'accord proches de zéro, alors que leurs annotations leur semblent cohérentes. La raison est souvent qu'ils appliquent les versions classiques des mesures qui considèrent toutes les catégories comme orthogonales (on parle de catégories « nominales ») alors que ce n'est pas le cas de leur modèle. Le cas canonique est celui où les catégories correspondent en réalité à une échelle de valeurs, par exemple des notes allant de 0 (mauvais) à 9 (parfait). Une version « nominale » des mesures considérera un désaccord entre deux valeurs proches comme 0 et 1 comme aussi important qu'un désaccord entre les valeurs extrêmes 0 et 9, contrairement à une version pondérée. Par ailleurs, plus

le nombre de catégories est important (*i.e.* plus la résolution des jugements est fine), plus le score sera faible. Notons que la mesure emblématique kappa de Cohen (1960) bénéficie d'une version pondérée présentée dans (Cohen, 1968), tandis que les α de Krippendorff sont nativement pondérables depuis leurs premières versions.

Nous avons mené l'expérience suivante, reportée dans le tableau 4 : deux annotateurs ont des sensibilités légèrement différentes, le second étant légèrement plus indulgent que le premier et mettant généralement une unité de plus que celui-ci (0 –> 1, ... 8 –> 9, et enfin 9 –> 9). Ils obtiennent ainsi, avec des annotations variées, un accord de 0,04 dans la version « nominale » de α, et de 0,95 dans sa version « intervalle ». En supposant à présent que pour la même campagne, on réduise à 5 au lieu de 10 le nombre de catégories (*i.e.* la résolution de jugement) : 0 et 1 deviennent 0, ... , 8 et 9 deviennent 4. Avec les mêmes données initiales, $\alpha_{nominal}$ monte à 0,52 et $\alpha_{interval}$ baisse à 0,90. Enfin, en limitant le nombre de catégories à 2 (jugement binaire, 0 à 4 deviennent 0, et 5 à 9 deviennent 1), $\alpha_{nominal}$ et $\alpha_{interval}$ se rejoignent à 0,81. On constate que les deux versions tendent l'une vers l'autre lorsque le nombre de catégories tend vers 2, qui est le point où la pondération n'est plus possible.

Échelle	Annotations										$\alpha_{nominal}$	$\alpha_{interval}$
0 à 9	0	1	2	3	4	5	6	7	8	9	0,04	0,95
	1	2	3	4	5	6	7	8	9	9		
0 à 5	0	0	1	1	2	2	3	3	4	4	0,52	0,90
	0	1	1	2	2	3	3	4	4	4		
0 à 1	0	0	0	0	0	1	1	1	1	1	0,81	0,81
	0	0	0	0	1	1	1	1	1	1		

Tableau 4. *Échelle de valeurs continues considérée comme telle ou comme un ensemble de catégories orthogonales*

Ces chiffres doivent inciter à utiliser les mesures pondérées à chaque fois que les catégories ne sont pas nominales. Nous avons constaté qu'elles sont parfois inconnues des responsables de campagne (alors dépités d'avoir des accords faibles malgré de bons annotateurs), mais aussi qu'elles sont parfois jugées complexes : par exemple, pour une campagne donnée, faut-il pondérer par une échelle linéaire ou quadratique (*i.e.* la distance entre les catégories 0 et 2 vaut-elle 2 ou 4 fois la distance entre 0 et 1) ? C'est un point important sur lequel la communauté du TAL devrait se pencher et proposer ses recommandations pour différents domaines, sans doute à partir d'observations statistiques en corpus.

5.2. *Des catégories nominales non orthogonales ?*

L'expérience précédente permet de questionner une idée reçue qui est que plus on augmente le nombre de catégories, plus l'accord baisse. En revanche, si les catégories sont nominales, cette affirmation redevient souvent vraie, car en proposant plus de finesse de catégorisation, on augmente le risque de choisir une catégorie au lieu d'une

autre, mais on sanctionne toujours de façon binaire des désaccords devenus moins graves. Par exemple, si l'on décide de remplacer la catégorie « Nom » par « Nom propre » et « Nom commun », l'accord mesuré va certainement baisser, parce que certains annotateurs qui ne confondraient jamais un nom avec autre chose pourraient malgré tout confondre un « Nom propre » avec un « Nom commun ». Une expérience intéressante a été menée par Fort *et al.* (2010) qui consiste à tenter de dresser, à partir de la matrice de confusion des annotations (qui met en évidence les couples de catégories qui sont mises en association par différents annotateurs), des sortes de distances entre catégories *a priori* nominales. Cette idée mériterait d'être approfondie afin de donner davantage de souplesse à l'évaluation des annotations nominales, en permettant de configurer les mesures (sous contrôle du responsable de la campagne) afin que toutes les catégories ne soient pas forcément considérées orthogonales deux à deux. C'est une question délicate, car poussé à l'extrême, ce raisonnement revient à transformer en accords des désaccords fréquemment observés, et c'est pourquoi cette approche devrait passer par la validation manuelle des pondérations proposées par le système.

6. Construire une référence à partir d'annotations multiples

La construction d'une référence repose souvent, nous l'avons vu, sur une annotation multiple. Une fois l'accord jugé suffisant, il s'agit alors de constituer une référence unique à partir de plusieurs sources, en fort accord global, mais fatalement avec des désaccords locaux. Différentes stratégies peuvent être envisagées, donnant lieu à des résultats différents. Nous nous en tenons ici à la catégorisation d'items prédéfinis, la synthèse d'annotation avec *unitizing* étant nettement plus complexe.

6.1. *Principe du vote à la majorité*

Chaque annotateur est considéré comme ayant voté pour une certaine catégorie pour chacun des items. La catégorie retenue est celle ayant obtenu la majorité. Selon le nombre de catégories et d'annotateurs, il peut y avoir des cas d'égalité, où ce seul principe ne permet plus de choisir. D'autre part, une majorité faible est susceptible de donner lieu à un choix non valide. Notons que des techniques issues du *crowdsourcing*, telles que celle du *Minimax Entropy* (Zhou *et al.*, 2012), visent à améliorer la référence obtenue, en s'appuyant en particulier sur le degré d'expertise supposé des différents annotateurs (évalué à partir des observations, et correspondant plus, selon nous, à un degré de consensualité que d'expertise).

6.2. *Principe du vote à l'unanimité*

Cette stratégie, qui consiste à ne retenir dans la référence que les unités pour lesquelles l'accord est total, s'appuie sur le fait que d'une part, si l'accord est très élevé,

c'est qu'il y a beaucoup d'items ayant un accord total, et que par ailleurs, lorsque l'accord n'est pas total sur un item, c'est que le choix n'est pas sûr. Cette stratégie qui se veut prudente doit certes donner lieu à une référence plus valide que dans le cas de la majorité simple, mais avec un biais non négligeable dont on n'a pas forcément conscience : les items retirés sont ceux qui ont fait le moins consensus, et qui donc sont les cas les plus difficiles à annoter. On obtient ainsi une référence biaisée ne contenant que les items les plus faciles, et les systèmes seront donc jugés, *via* cette référence, de façon surévaluée.

6.3. *Principe de la révision collégiale*

Il s'agit ici de reprendre manuellement et collectivement l'observation de tous les items n'ayant pas fait l'unanimité, d'analyser pourquoi il y a eu désaccord et de prendre une décision collective. Cette méthode est de loin à la fois la plus coûteuse et la meilleure. Elle balaye les biais des deux précédentes méthodes, et elle permet le cas échéant de déceler des problèmes dans la tâche ou dans le modèle d'annotation (ambiguïtés, etc.). C'est une stratégie qui a été notamment utilisée lors de la constitution de la ressource Annodis (Péry-Woodley *et al.*, 2011).

7. Évaluation des systèmes

Comme nous l'avons vu, l'évaluation de la performance d'un système d'annotation est une tâche bien différente de l'évaluation de l'accord entre annotateurs. En effet, pour évaluer sa performance, il faut comparer ses productions au résultat exact auquel il est censé parvenir (la « référence »), et non à d'autres productions dont on ignore si elles sont valides. Nous allons voir dans cette partie que toute dérogation à ce principe constitue une erreur de méthode conduisant à des conclusions non validées.

7.1. *Une mesure d'accord pour évaluer un système ?*

Comme on le constate dans certaines publications, les chercheurs sont parfois tentés d'utiliser des mesures d'accord pour évaluer leurs systèmes, que ce soit pour des raisons relevant du « marteau de Maslow », ou dans le but d'intégrer une sorte de *baseline* (la chance) dans le résultat obtenu, afin d'estimer la plus-value réelle du système.

De notre point de vue, cette pratique constitue une erreur importante. Tout d'abord, si, comme nous l'avons vu, la prise en compte de la « chance » est fondamentale dans le cadre d'annotations multiples sans référence, elle n'est d'une part nullement nécessaire lorsque l'on dispose d'une référence, et d'autre part rend l'interprétation des résultats plus difficile. Il est plus utile de savoir qu'un système produit 92,3 % de résultats corrects (on sait clairement qu'il y a en moyenne 7,7 % d'erreurs, et on en tire les conséquences pratiques), plutôt que d'apprendre qu'il obtient un score de 0,712

d'accord κ avec la référence. Loin de nous l'idée de nier l'importance de disposer d'une *baseline*, point sur lequel nous ferons des propositions.

Mais l'erreur est bien plus profonde : en soumettant des systèmes aux mesures d'accord, chaque système va créer lui-même sa propre *baseline*. Ainsi, les scores obtenus par deux systèmes différents (ou par deux « *runs* » d'un même système) ne seront absolument pas comparables. En effet, dans la formule de calcul d'une mesure d'accord corrigée par la chance, deux valeurs interviennent : A_o, l'accord observé (le seul qui devrait retenir notre attention ici), et A_e, l'accord que l'on est censé obtenir par la chance. C'est cette dernière valeur qui pose problème ici : elle est calculée à partir des données observées, c'est-à-dire en l'occurrence à la fois de la référence et des sorties du système. On voit donc que la *baseline* proposée par les mesures d'accord est créée, pour moitié, par le système évalué lui-même. Pour mieux comprendre le phénomène et sa potentielle ampleur, nous avons construit un exemple extrême, reporté dans le tableau 5.

	1	2	3	4	5	6	7	8	Score	Chance	κ
Référence	A	A	A	A	B	B	B	B	-	-	-
Système 1	A	A	C	C	C	C	C	C	0,25	0,125	0,143
Système 2	A	A	B	B	B	B	A	A	0,50	0,50	0,0

Tableau 5. *Exemple d'évaluation biaisée de deux systèmes par la mesure d'accord κ*

Une tâche de catégorisation met en jeu trois catégories dénommées A, B et C. Nous disposons d'une référence pour huit items à annoter (numérotés de 1 à 8), et deux systèmes sont évalués. Pour chacun de ces derniers, trois valeurs sont calculées : le résultat brut, appelé « score » (exprimé entre 0 et 1 par souci d'homogénéité), la valeur A_e du κ, appelée « chance », et enfin la valeur κ. Le système 1 a un score de 0,25 (il ne trouve la bonne catégorie que pour les items 1 et 2), tandis que le système 2 a un score de 0,5 (items 1, 2, 5 et 6). Concernant la correction par la chance, le système 1 a le bonheur de proposer des prévalences bien différentes de la référence. κ va donc considérer qu'on ne peut avoir un accord par chance que dans un cas sur huit (deux chances sur huit que le système 1 produise A, multiplié par quatre chances sur huit que la référence produise A), soit 0,125. Le système 2 a le malheur de produire des prévalences identiques à la référence (quatre A et quatre B), ce qui induit un accord par chance de 0,5. Le système 2 sera donc jugé bien plus sévèrement que le système 1 par κ. Le résultat est sans équivoque : le système 2, bien que deux fois meilleur que le 1 (avec un score de 0,5 contre 0,25), obtient un κ nettement inférieur (0 contre 0,143).

Nous recommandons donc de ne pas utiliser les mesures d'accord pour évaluer la performance des systèmes par rapport à une référence, mais d'utiliser des distances directes par rapport à cette dernière. Il peut s'agir, de façon non limitative : 1) du pourcentage brut d'accord catégoriel dans le cas de la catégorisation d'items prédéfinis, 2) des valeurs de rappel, précision et f-mesure dans le cas où les éléments à caractériser ne sont pas prédéfinis, 3) de *window diff* ou mieux, de la distance de Hamming généralisée pour la segmentation, 4) ou plus généralement, de la valeur brute A_o calculée par une mesure d'accord (sans prendre en compte la valeur *expected* A_e) lorsque cette

dernière a été utilisée dans la phase manuelle de constitution de la référence (cela permet en effet généralement d'obtenir une valeur plus fine que les distances classiques, avec par exemple la prise en compte de la pondération des distances entre catégories, comme nous l'avons précédemment vu), 5) enfin, pour qui souhaite disposer d'une *baseline* à moindre frais, en l'absence d'autres systèmes auxquels se comparer, nous proposons d'utiliser la valeur A_e calculée lorsque l'on soumet la référence à elle-même. La mesure obtenue sera bien sûr de 1, puisque $A_o = 1$, mais la valeur A_e correspond à ce que l'on peut obtenir en considérant les prévalences observées dans la référence. C'est une proposition inédite à notre connaissance.

7.2. *Une référence qui n'en est pas une : comparer un système aux humains*

Nombre de travaux sur le développement de systèmes sont confrontés non seulement à l'évaluation de leur système, mais aussi, en amont, à la constitution manuelle des annotations qui permettront d'évaluer ce dernier. La raison est le plus souvent que ces systèmes se destinent à de nouvelles tâches pour lesquelles, par définition, aucune référence n'existe. Dans ce contexte, deux évaluations sont faites, celle concernant les évaluations manuelles (pour constituer la référence), puis celle concernant le système (par rapport à la référence). Il est dès lors tentant d'assimiler et de comparer les deux, puisque le système a pour but de faire aussi bien que la référence, c'est-à-dire l'humain. Nous allons voir deux variantes de ce principe, observées dans la littérature.

7.2.1. *Variante 1 : une annotation humaine non validée comme référence*

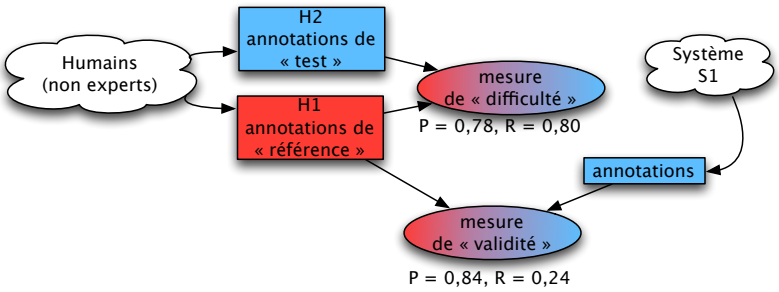

Figure 6. *Évaluation confuse par absence de vraie référence*

Pour une tâche donnée, un humain H1 annote un corpus, et constitue ainsi la référence. Un système S1 se voit soumettre le même corpus, pour la même tâche, et ses performances sont évaluées par rapport à H1 en termes de précision et de rappel, avec des scores respectifs de 84 % et de 24 %. Toutefois, les auteurs ont l'idée de soumettre ce corpus à un second humain H2 pour relativiser les performances plutôt moyennes de S1. H2 est lui aussi évalué par rapport à la référence H1, et obtient 78 % de précision et 80 % de rappel. Il en est conclu que la tâche est difficile, puisque même l'humain ne fait pas mieux que le système en ce qui concerne la précision des réponses.

Cette démarche, illustrée en figure 6, comporte d'importants biais méthodologiques, au premier rang desquels le fait de ne pas disposer d'une référence solide, H1 ayant été choisi arbitrairement comme tel alors qu'il n'est pas plus expert que H2. On dispose en fait de deux références humaines non validées, si bien que le « score » obtenu par H2 correspond plutôt à un accord entre annotateurs humains, qui est ici d'autant plus modeste que cette valeur est non corrigée par la chance. Supposons maintenant qu'un système S2 produise des valeurs de précision et de rappel de 100 %, (c'est-à-dire supposons qu'il produise exactement les mêmes sorties que H1). On doit en déduire que S2 est bien meilleur que l'humain (H2) pour faire ce que sait faire l'humain (H1), ce qui n'a aucun sens. Un score de 100 % n'est donc guère plus souhaitable que 84 %, ce qui montre combien les chiffres ne veulent ici rien dire. On peut seulement conclure de l'ensemble de ces données que le score obtenu entre annotateurs est trop faible pour constituer une référence solide. La tâche est probablement à revoir ou à préciser. On ne peut donc pas tirer de conclusion concernant les performances du système, puisqu'on ne peut le comparer à rien d'établi.

7.2.2. *Variante 2 : un système qui est globalement d'accord avec les humains en est-il pour autant performant ?*

Cette fois, une annotation multiple manuelle est tout d'abord effectuée, et un accord inter-annotateurs humains α_h est calculé sur ces données, conformément à la première étape classique. Ensuite, plutôt que de constituer un corpus de référence à partir de ce premier socle, c'est cet ensemble multi-annoté, sorte de produit mal fini, qui sert de pseudo-référence selon le principe suivant : les sorties d'un système à évaluer sont considérées comme le n+1 ième annotateur, et on applique de nouveau la mesure d'accord sur cet ensemble (humains + système) pour obtenir la valeur α_{h+s} . Si α_{h+s} est supérieur ou égal à α_h , alors le système peut être considéré comme un annotateur au moins du niveau des humains. L'idée est particulièrement tentante si l'on considère le temps et les moyens que demande la constitution d'une véritable référence. Pourtant, là encore, de tels chiffres n'ont aucune valeur. Considérons le cas d'un corpus de six items à annoter au moyen de trois catégories, et pour lequel nous disposons de trois annotateurs humains, afin d'évaluer les sorties de deux systèmes. Les annotations sont reportées dans le tableau 6. Comme précédemment, les annotations des trois humains et des deux systèmes sont reportées lignes 1 à 5. Nous avons ajouté deux lignes correspondant respectivement à la « référence par majorité » et à la « référence par révision collégiale » (qui correspond à ce que les annotateurs auraient dû idéalement faire). On en déduit les scores des différents annotateurs relativement à ces deux références : pour la « majorité », tous les humains ont un score de 83 %, et pour la « révision collégiale » les scores vont de 100 % pour le premier à 67 % pour les deux suivants, soit une moyenne humaine de 78 %. Si l'on considère la référence « par majorité », les deux systèmes font monter l'accord mais baisser le score (67 % *versus* 83 %), ce qui est contraire à l'intuition. Si l'on considère la référence « par révision collégiale », le système 2 qui fait un α_{h+s} non seulement supérieur à α_h, mais aussi supérieur au système1 (0,529 contre 0,503), n'a pourtant qu'un score de 50 %, largement inférieur à la moyenne des humains, mais encore plus largement inférieur

	1	2	3	4	5	6	α_h	α_{h+s}	Score M	Score RC
Humain 1	A	B	C	A	A	B		-	83 %	100 %
Humain 2	A	B	C	B	B	B	0,495	-	83 %	67 %
Humain 3	A	B	C	A	B	A		-	83 %	67 %
Système 1	A	B	C	A	A	C	-	0,503	67 %	83 %
Système 2	A	B	C	B	B	A	-	0,529	67 %	50 %
Référence M	A	B	C	A	B	B	-	-	-	-
Référence RC	A	B	C	A	A	B	-	-	-	-

Tableau 6. *Variante d'évaluation biaisée de deux systèmes utilisant α (M = par majorité, RC = par révision collégiale)*

au score du système 2 (83 %). Ainsi, dans un cas comme dans l'autre, il est possible de faire grimper l'accord tout en ayant un score inférieur à chacun des humains. Cela résulte d'un mélange des genres : dans α_{h+s}, le système et les humains jouent le même rôle, or on ne peut pas être juge et partie. Ce principe est donc, lui aussi, à proscrire.

8. Conclusion

Si notre communauté a déjà largement admis la nécessité de procéder à une évaluation des données annotées, ce consensus n'a pas conduit encore à la mise en place d'une réflexion suffisamment aboutie sur les moyens à mettre en œuvre pour vérifier et maximiser la pertinence des résultats d'évaluation. Au travers de cet article, nous espérons avoir contribué à éclairer les conséquences d'un mauvais usage ou d'une mauvaise compréhension des mesures disponibles et avoir contribué à montrer que, si la simple disponibilité de certaines métriques et la régularité de leur utilisation dans la communauté prévalent parfois sur la prise en compte de leurs champs d'application respectifs, l'intelligibilité des résultats d'évaluation présentés peut s'en trouver très largement compromise. Mais l'importance de cet effort d'élucidation des problèmes liés à l'évaluation nous interdit, dans le cadre limité de cet article, d'aller au-delà de réflexions sporadiques, peu systématisées, et d'illustrations non représentatives de l'extrême diversité des difficultés susceptibles d'être rencontrées dans ce domaine. Aussi, en conclusion, nous appelons de nos vœux la constitution, par notre communauté et pour notre communauté, d'un espace de réflexion sur ces pratiques de l'évaluation, d'un groupe de travail visant à établir un guide des bonnes pratiques et à répondre aux demandes de ceux d'entre nous qui, s'engageant dans des campagnes d'annotation pour le bénéfice de tous, souhaitent un éclairage sur ces questions d'évaluation, questions centrales mais souvent fatalement assez marginales par rapport à leurs préoccupations scientifiques premières.

9. Bibliographie

Aickin M., « Maximum Likelihood Estimation of Agreement in the Constant Predictive Probability Model, and Its Relation to Cohen's Kappa », *Biometrics*, vol. 46, p. 293-302, 1990.

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

Bennett E. M., Alpert R., C.Goldstein A., « Communications through Limited Questioning », *Public Opinion Quarterly*, vol. 18(3), p. 303-308, 1954.

Berry C. C., « The K statistic - to the editor », *Journal of the American Medical Association*, vol. 268, n° 18, p. 2513-2514, 1992. Letters.

Bonnardel P., Test statistique Kappa : programmation informatique et applications pratiques, Phd thesis, Université de Paris V, 1996.

Cohen J., « A Coefficient of Agreement for Nominal Scales », *Educational and Psychological Measurement*, vol. 20, n° 1, p. 37-46, 1960.

Cohen J., « Weighted kappa : Nominal scale agreement with provision for scaled disagreement or partial credit », *Psychological Bulletin*, vol. 70, n° 4, p. 213-220, 1968.

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 problems of Two Paradoxes », *Clin. Epidemiol.*, vol. 43, p. 543-548, 1990.

Fort K., Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus. Traitement du texte et du document, PhD thesis, Université Paris- Nord - Paris XIII, 2012.

Fort K., François C., Ghribi M., « Evaluer des annotations manuelles dispersées : les coefficients sont-ils suffisants pour estimer l'accord inter-annotateurs ? », *Traitement Automatique des Langues Naturelles (TALN)*, Montréal France, 2010.

Goldman R. L., « The K statistic - to the editor (in reply) », *Journal of the American Medical Association*, vol. 268, n° 18, p. 2513-2514, 1992.

Gwet K. L., *Handbook of Inter-rater Reliability*, third edn, Advanced Analytics, LLC, 2012.

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

Krippendorff K., « On the reliability of unitizing contiguous data », *Sociological Methodology*, vol. 25, p. 47-76, 1995.

Krippendorff K., « Agreement and Information in the Reliability of Coding », *Communication Methods and Measures*, vol. 5.2, p. 93-112, 2011.

Krippendorff K., *Content Analysis : An Introduction to Its Methodology*, third edn, Sage : Thousand Oaks, CA., 2013a.

Krippendorff K., « A dissenting view on so-called paradoxes of reliability coefficients », *C. T. Salmon (ed.), Communication Yearbook*, vol. 36, p. 481–499, 2013b.

Krippendorff K., Mathet Y., Bouvry S., Widlöcher A., « On the reliability of unitizing textual continua : Further developments », *Quality and Quantity*, 2016.

Maslow A. H., *The Psychology of Science*, New York : Harper Row, 1966.

Mathet Y., Widlöcher A., Métivier J.-P., « The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment », *Computational Linguistics*, vol. 41, n° 3, p. 437-479, 2015.

Péry-Woodley M.-P., Afantenos S., Ho-Dac L.-M., Asher N., « Le corpus ANNODIS, un corpus enrichi d'annotations discursives », *revue TAL*, vol. 52, n° 3, p. 71-101, 2011.

Scott W., « Reliability of content analysis : The case of nominal scale coding », *Public Opinion Quarterly*, vol. 19, n° 3, p. 321-325, 1955.

Zhao X., Liu J., Deng K., « Assumptions behind inter-coder reliability indices », *C. T. Salmon (ed.), Communication Yearbook*, vol. 36, p. 418-480, 2013.

Zhou D., Platt J. C., Basu S., Mao Y., « Learning from the Wisdom of Crowds by Minimax Entropy », *Advances in Neural Information Processing Systems (NIPS)*, vol. 25, p. 2204-2212, 2012.

Zwick R., « Another look at interrater agreement », *Psychological Bulletin*, vol. 103, p. 347-387, 1988.