Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2940–2948
Marseille, 20-25 June 2022

Validity, Agreement, Consensuality and Annotated Data Quality

Anaëlle Baledent, Yann Mathet, Antoine Widlöcher,
Christophe Couronne, Jean-Luc Manguin
Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, FRANCE
{anaelle.baledent, yann.mathet, antoine.widlocher, christophe.couronne, jean-luc.manguin}@unicaen.fr

Abstract

Reference annotated (or gold-standard) datasets are required for various common tasks such as training for machine learning systems or system validation. They are necessary to analyse or compare occurrences or items annotated by experts, or to compare objects resulting from any computational process to objects annotated (selected and characterized) by experts. But, even if reference annotated gold-standard corpora are required, their production is known as a difficult problem, from both a theoretical and practical point of view. Many studies devoted to these issues conclude that multi-annotation is most of the time a necessity. Measuring the inter-annotator agreement, which is required to check the reliability of data and the reproducibility of an annotation task, and thus to establish a gold standard, is another thorny problem. Fine analysis of available metrics for this specific task then becomes essential. Our work is part of this effort and more precisely focuses on several problems, which are rarely discussed, although they are intrinsically linked with the interpretation and the evaluation of metrics. In particular, we focus here on the complex relations between agreement and reference (of which agreement among annotators is supposed to be an indicator), and the emergence of a consensus. We also introduce the notion of *consensuality* as another relevant indicator.

Keywords: Annotations, Agreement, Validity, Consensuality

1. Introduction

There is a broad consensus concerning the need for gold-standard annotated datasets, and many annotation campaigns aim at producing such datasets. Especially in the Natural Language Procession (NLP) community we come from, they provide a crucial element to study a phenomenon, to evaluate or to train a system devoted to the detection or the analysis of that phenomenon. However, the creation of such datasets is accompanied by theorectical and practical issues, especially when annotated phenomena are complex and require fine interpretation. When it is considered as impossible to reach directly, the establishment of the reference is therefore in practice often postponed, in favor of a confrontation of multiple points of view on the same data. Indeed, it is assumed that a significant agreement among annotators is a necessary condition for the emergence of a reference. A more debatable hypothesis assumes that such an agreement is also a sufficient condition.

In order to clarify the methodological framework in which reference data are produced, communities promoting data-centric approaches tend to support works directly devoted to these questions (Aroyo and Welty, 2015; Oortwijn et al., 2021). For example, in the NLP community, we can mention (Fort, 2016) or (Mathet et al., 2015), where it appears that measuring agreement among annotators – a necessary prerequisite to the establishment of a reference – is a difficult problem. These issues make it necessary to analyze available metrics, to compare them and to study their conditions of interpretation, as in works such as (Artstein and Poesio, 2008) or (Mathet et al., 2012).

In this paper, our main goal is to refine our understanding of the way a consensus emerges and of its complex relationship to the truth. In particular, we focus here on the two following problems:

P1 Does a good inter-annotator agreement necessarily give access to the truth? How to deal with the risk of annotators agreeing on errors?

P2 How does the consensus emerge? If it is clear that the overall agreement obviously depends on the individual performance of each actor, how to measure its influence?

In Section 2, we present both the main goal and the methodology of our approach. Section 3 introduces the annotation task and collected data on which our experiments were carried out. We define the notions of agreement, validity and consensuality in Section 4. Section 5 reports our analysis of the relationship between consensuality and performance. Finally, Section 6 provides a few words of conclusion and presents some of our future works concerning consensuality.

2. Goals and methodology

2.1. Guidelines for annotation methodology

In response to questions **P1** and **P2** supra, our main goal is here to provide methodological recommandations for annotation campaign managers concerning i) the interpretation of agreement measures and ii) the tracking and, if at all possible, the filtering of unreliable or less reliable data (and/or annotators). For some annotation tasks, a partial reference (the ground truth for a part of the dataset) may be available and we aim at exhibiting useful controls to do in this case. Furthermore, we want to determine what consequences can be

drawn from this when no reference is availble, which is a very common case.[1]

In short, our main goal is to highlight, by isolating them, various parameters and problems little commented in the literature although intimately linked to manual annotation and to the establishment of a gold standard. The analysis of theses problems and biaises should allow us to "integrate as soon as possible the compensation or control processes" (Braffort et al., 2011), and therefore to improve annotation quality and reliability.

And even if we carry out our expriments on a very specific dataset resulting from a very specific annotation task, our main contribution concerns methodology: we aim at providing generic and reusable guidelines and verification procedures for controlling the process of various annotation campaigns.

2.2. Global roadmap

In order to address problem **P1**, we have to carry out experiments in a suitable context where the ground truth is available. It is indeed necessary to set up ways of studying the correlations beetween this reference and the individual and collective productions, and ways of observing the *performance* understood as the degree of proximity to the ground truth.

In order to address problem **P2**, we introduce the notion of *consensuality*, understood as the participation of each annotator in the emergence of a consensus. And we propose ways of observing the contribution of an annotator to the agreement of a group.

Then we investigate the relationship between consensuality and performance, to find out, in particular, whether the consensuality may be a performance predictor, useful even when no reference is available.

2.3. Requirements concerning the dataset

To carry out these observations, limiting biases and hidden parameters, we need an annotation campaign and a dataset fulfilling the following requirements :

- **An indisputable reference:** Working with an available and indisputable reference was our primary concern. Indeed, we want to observe the delicate problem of the relationship between inter-annotators agreement and ground truth reference to which agreemeent is supposed to give access.
- **A task with interpretation:** To mimic literary or linguistic annotation tasks, the selected task has to imply a significant amount of interpretation, so that annotations differ sufficiently among annotators.[2]
- **A task requiring no special training:** For this experiment, we want to limit the effects of annotators' skills on their annotations and we do not want to deal with their training level (subject already discussed in (Dandapat et al., 2009; Bayerl and Paul, 2011)). The annotation should therefore be an easy task which should not require special training or advanced skills.
- **Scalar annotations:** Scalar numerical annotation (vs. nominal values), for each item, makes it possible to finely quantify a) the performance for each annotated item (i.e. the distance between an annotated value and the reference value) and b) the similarity between annotations produced by several annotators.[3]
- **Aggregatable annotations:** We have to compute and to compare performances within groups of annotators but also between such groups. Aggregatable annotations enable the computation of a collective annotation, based on individual ones.

3. Annotation task and data

3.1. Annotation task

To take the constraints mentionned in Section 2.3 into account, we have chosen the task of estimating the age of human based on photographs. This task fulfills all the above mentionned requirements. Indeed, it is easy to have access to the exact age at the time the photograph was taken and then to an indisputable reference. Age estimation relies on a sometimes difficult interpretation of face's visual characteristics and the resulting annotations may consequently vary a lot for the same photography from one annotator to another. Besides, this task is one that everyone has faced before and it does not require any special training. Finally, the numerical nature of the annotations makes it possible to finely quantify the difference between the truth and the annotations, and to perform different mathematical operations on them.

Admittedly, the age estimation does not immediately concern the NLP domain we come from. However,

[1] For some annotation tasks, having a gold standard is sometimes not necessary, possible or desirable, especially when access to the *truth* is regarded as impossible. In this study, we mainly focus on annotation tasks for which such a reference is considered both possible and desirable, even if it is not already available.

[2] It is important to point out that making all these requirements compatible may be a challenging problem. For example, having an indisputable reference is all the more difficult as the annotation process requires a significant amount of interpretation. However, a compromise has to be found to limit biases in our observations as much as possible.

[3] This constraint also meets the requirements of an other work devoted to scalar annotations, involved in the same project but not presented in this paper. Widely used in NLP, for example for sentiment or opinon analysis (Bregeon et al., 2019; Bradley and Lang, 1999; Kang et al., 2018), such scalar annotations make it for example possible to express polariry and intensity. They are little studied from the inter-annotator agreement point of view on wich our project focuses.

we selected this specific task because it is possible and quite easy to build the corpus and the reference and to drive the annotation process without hidden parameters and with few biases, making our observations easier and more transparent. We also hope that our methodology and our protocols of observation can be generalized on other annotation tasks (in particular for textual datasets to which NLP tasks are mainly devoted).

3.2. Biases for the age estimation

Even if age estimation is not, as such, the focus of this study (devoted, more generally, to validity, agreement and consensuality), it is nevertheless necessary to take into account biases related to this specific task that could affect our observations.

At first, the biases that annotators may face because of the use of photographs have to be mentionned. Indeed, the photograph's quality (in color or black and white, varied deteriorations, out of focus, etc.) may affect the interpretation of the age.

More specifically, the use of photographs of famous people may introduce other biases. The subjects are usually well made up and know how to pose, and look more younger or older according to their intentions. The annotator can also recognize the celebrity and/or an event, which sometimes makes her task easier... but sometimes misleads her.

Concerning biases related to the annotation process, (Clifford et al., 2018) mention two common important biases. The first one is the serial dependency, whereby the annotator tends to bring the person's age on the actual photograph towards the age of the previous photograph. The second biais concerns the fact that young people seem to be older than they are, whereas the older seem to be younger.

(Vestlund et al., 2009) also notice the same second biais observed by (Clifford et al., 2018), whereby we overestimate the age of younger people and underestimate the age of older people. (Voelkle et al., 2012) mention a better age estimation when we are confronted with portraits that are close to our own age group. Finally, (Watson et al., 2016) note a tendency to bring the person's age closer to our own age.

3.3. Collecting the data

For our experiment, we built a corpus from 100 photographies (collected from WIKIMEDIA COMMONS) of persons whose ages are within the range 3 months - 97 years.

A questionnaire, with each question displaying a portrait, was open for one day. The time required for annotating the 100 images was approximately 30 minutes, but each annotator was given one hour to complete the task. Annotators were non-expert students, in their twenties, who never participated in any annotation campaign before, and who were unaware of the goal of our study. The provided instruction was:

> For each question, you will be presented with a photograph of a human individual and you will have to determine the age as accurately as possible, which you will simply indicate in the input field provided for this purpose by entering a simple integer number corresponding to your estimate.

52 annotators have taken part in the campaign, for a total of 4850 usable annotations. For this experiment, we keep only the annotators who answered all the questions: in total, we keep 42 annotators.

3.4. First approach of annotations

Figure 1 provides a general overview of the productions of all the annotators, and makes it possible to compare their answers to the reference. It should be emphasised that we model here the performance of a mean annotator; we do not compare the annotators with each other. On this graph, for each photography, the real age is a dot and the average of estimated age is represented by a square marker. Thus, it allows us to see the difference with the reference and to see if annotators think people are younger or older than they really are.

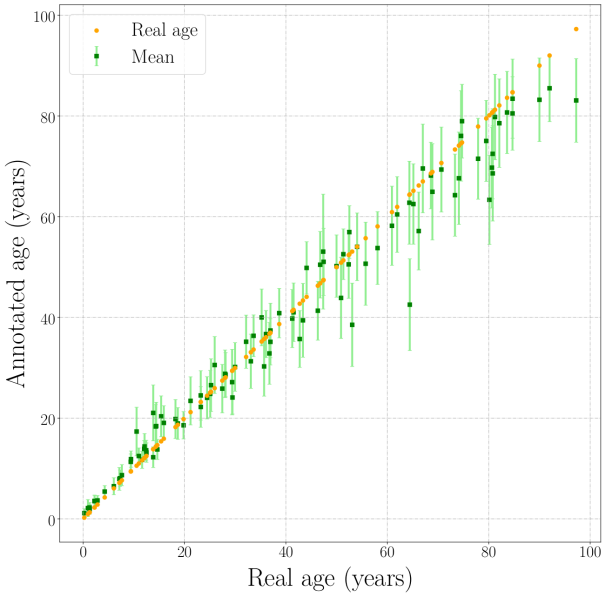

Figure 1: Age given by annotators for each photography

Two trends have to be noticed: (1) the annotators tend to wrongly increase the age of young people; (2) they tend to wrongly decrease the age of old people. This observation confirms the conclusions of (Clifford et al., 2018). It is not surprising that the estimated age of children and teenagers remains fairly close to their actual age, whereas estimated age of older people (beyond 30 years old) tends to be less accurate. Indeed, it is hard to get several years wrong for a baby or child, but it is more difficult to be that accurate for older people. We also plot error bars, which support this interpretation: while the annotations for the first photographs tend to

be more clustered around the actual age, they seem to be more disparate as the pictures illustrate older peoples.

4. Agreement, validity, consensuality

On Figure 1, two main observations stand out:

1. There is a difference between the reference age and the mean annotation, sometimes important;
2. There is variance among annotators, as revealed by the error bars.

In this Section, we lay the foundations to examine theses two points more closely.

4.1. Agreement and validity

It is important to distinguish between the notions of agreement and validity, which are sometimes misunderstood or even mixed up:

- **Agreement**: in the context of multiple annotation, an inter-annotator agreement measure attempts to propose a degree of similarity between annotations from distinct annotators;
- **Validity**: in the context of an automatic annotation process, a validity measure attempts to provide a degree of similarity between a candidate annotation and a reference (generally an annotation judged valid by an expert). This terminology is in accordance with (Krippendorff, 2013).

If these two types of measures can be similar since they pronounce on similarities between sets of annotations, let us mention a first fundamental difference: in case of agreement, there is no reference, the entries all have the same status, whereas there is a deep asymmetry in case of validity, where we compare a candidate to the "truth".
Other differences may exist, such as the fact that in agreement, there can be as many annotation sets as desired (e.g. 10 annotators), while there are systematically 2 in validity. Besides, for agreement, measures generally try to remove the part of chance involved in the agreement among annotators. It follows that if several annotators all annotate perfectly, their annotations will be similar, and their agreement will be total. But the reciprocal is unfortunately not established. For example, annotators can agree (on all or part of the annotated elements) without their annotations being valid, because they can make the same mistakes.

4.2. Towards the definitions of consensuality

We will try to clarify the links between agreement and validity. For an image i, μ_i denotes the reference and $x_{i,a}$ refers to the annotation of the annotator a over it. N is the total number of images. Given a subgroup of annotators, $G = \{a_1, ..., a_n\}$, we denote as $|G|$ its cardinality and $\sigma_i(G)$ the variance of this subgroup over image i. We define:

- the **annotator's imperfection** (the contrary of the **annotator's performance**) is calculated based on formula 1:

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

  This formula is similar to standard deviation. When it is zero, corrresponding annotations are all valid.

- the **annotators' group imperfection** is the average of all the annotator's imperfections in the group. See formula 2:

$$\text{imperfection}(G) = \frac{1}{|G|}\sum_{j=1}^{|G|}\text{imperfection}(a_j) \quad (2)$$

  When it is zero, all the annotations of all the annotators of the group are valid.

- the **annotators' group disagreement**: we take the variance of annotators on each image. The disagreement is the average of this value for all photos. See formula 3:

$$\text{disagreement}(G) = \frac{1}{N}\sum_{i=1}^{N}\sigma_i(G) \quad (3)$$

  When this disagreement is zero, the annotators have all, for each photo, given the same age (but not necessarily the right one).

- the **annotator's consensuality degree regarding a group** (to which she belongs) is given by the algebraic difference between the disagreement of this group deprived of this annotator a, and the disagreement of this group. See formula 4, for the annotator $a \in G$:

$$\text{consensuality}(a, G) = \text{disagreement}(G \setminus a) - \text{disagreement}(G) \quad (4)$$

  If this algebraic value is positive, it means that this annotator generates agreement, therefore that she is consensual. Warning, the annotator's consensuality degree is relative to the group considered. We distinguish two ways of establishing individual consensuality:

  - **Initial Consensuality**: for each annotator of the considered group, we compute her consensuality regarding the whole group, and we sort the annotators according to their consensuality value.
  - **Progressive Consensuality**: for this variant, we proceed in an iterative way. From the initial consensuality, we remove the least consensual annotator. We repeat the process

from the remaining subgroup, recalculating at each iteration all the consensualities from the remaining group. Thus, we try to keep, little by little, the most consensual annotators among those who are already the most consensual.

These measures do not incorporate the notion of chance, like the inter-annotator agreement measures generally use. Indeed, we do not compare annotations in order to establish a reference, but to observe the agreements among different annotators and groups of annotators and to sort them. Integrating a chance correction would not alter ordering and ranks.

5. Consensuality analysis

Unless otherwise indicated, figures read from right to left, and from top to bottom.

5.1. Consensuality ranking versus performance ranking

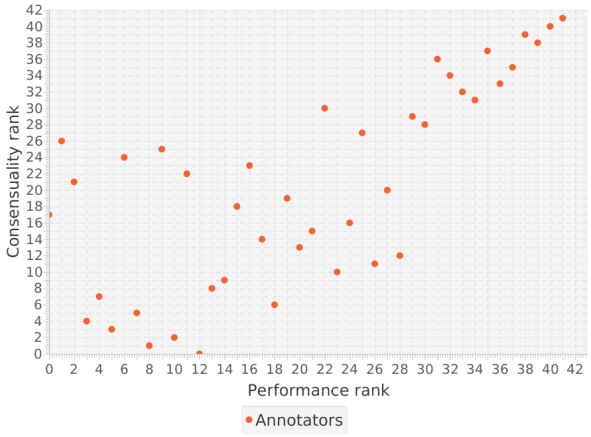

Figure 2: Annotators' ranks according to their performance and their consensuality (here with progressive consensuality, but the phenomenon is quite the same with initial consensuality)

At first, we have computed the annotators' consensuality ranking. Figure 2 presents each annotator (materialized by a dot), arranged along two axes: her performance rank is graduated on abscissa axis and her consensuality rank on ordinate axis. Thus, the dots in top right represent the least performing and least consensual annotators; the dots in bottom left are the most performing and most consensual annotators.
Ideally, what we would expect is that the dots take place on a perfect straight line from top right to bottom left, which would indicate that consensuality and performance are directly correlated. In this ideal situation, we would have the opportunity to know how performs an annotator compared to others even in the absence of a reference. Of course, and unfortunately, this is not the case here. It is about to be the case in the top right corner, where we can see that the 5 or 6 worst annotators are also almost the least consensual, but the more we go to the bottom left, the more scattered the dots.

From this first graph, we conclude that there is no strong correlation between the annotators consensuality and a good annotators' performance.

5.2. Removing least consensual annotators

If we progressively discard least performing annotators from our set, of course the global score will increase.

But in the absence of a reference, is it relevant to remove least consensual annotators to increase the global score? In this section, we make a series of experiments to understand what happens when progressively removing annotators according to their lack of consensuality.

In the following graphs, the first dot (in top-right) corresponds to the whole set of annotators, then the second dot corresponds to the whole set minus the least consensual annotator, and so on till the last point which corresponds to singleton containing the most consensual annotator (at the left of the figure, but not necessarily the bottom).

5.2.1. Group disagreement and annotator's performance evolution

First, we observe in figures 3 (for initial consensuality) and 4 (for progressive consensuality) the evolution of group disagreement (abscissa axis) and individual imperfection (ordinate axis) of the annotator removed from the group. In the top right of the figures, all annotators are present, the disagreement is at its top, and we are about to remove the least consensual annotator (who is also the least performing one), and so on.

What we would ideally expect here is that we remove annotators from worst to best, and so to get descreasing curves. Of course, what we observe is different, but we can see that it is the case at the beginning of the curves (in top-right part). The curve of progressive consensuality has a better behavior, which removes mainly worst annotators till rank 15, whereas it is about rank 8 for initial consensuality. We also see that the best annotator is removed at rank 25 for progressive consensuality, whereas it is at rank 10 for intial consensuality.

To conclude this part, we can see in these figures that consensuality provides interesting clues about performance at a global point of view (the curves are gloablly descreasing). At an annotator level, we can only observe that first least consensual are also the least performing annotators, but after rank 10, some really performing annotators may unfortunately be removed.

For this reason, it is interesting to focus now on what happens at group level rather than at annotator level and ask the following question: does removing least consensual annotator improve group performance? The next section is devoted to this question.

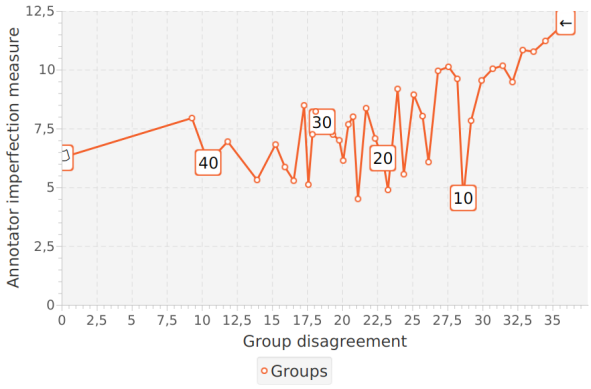

Figure 3: Removing the least consensual annotator each time with *initial consensuality*

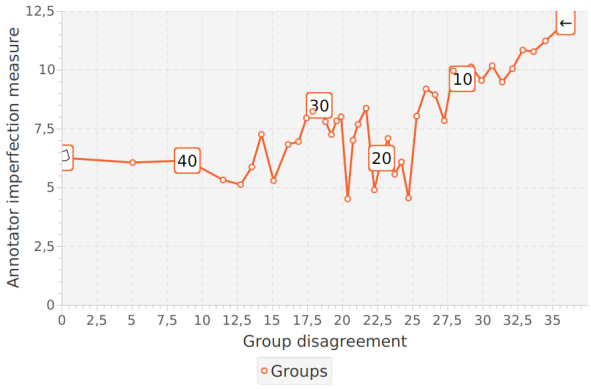

Figure 4: Removing the least consensual annotator each time with *progressive consensuality*

5.2.2. Group disagreement and group performance evolution

Figure 5 follows the same principle as figure 4, except that instead of displaying the individual imperfection, it displays the group imperfection on the vertical axis. Initially, we notice a slow decrease in group imperfection: the removal of least consensual annotators allows the overall imperfection to decrease. However, after the removal of the sixteenth annotator (from the right), we observe a small rise in group imperfection. Other rises in imperfection can be seen on the curve. In the light of figure 4, we can link these rises in group imperfection to the withdrawals of the best performing annotators. Despite these withdrawals, group imperfection of the most consensual annotators remains acceptable: we could consider keeping only the 10% most consensual annotators and still get good annotations.

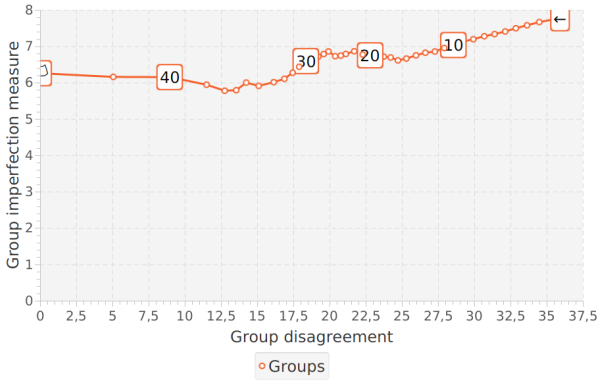

Figure 5: Group imperfection by removing the least consensual annotator each time (progressive consensuality)

5.3. Initial consensuality versus progressive consensuality

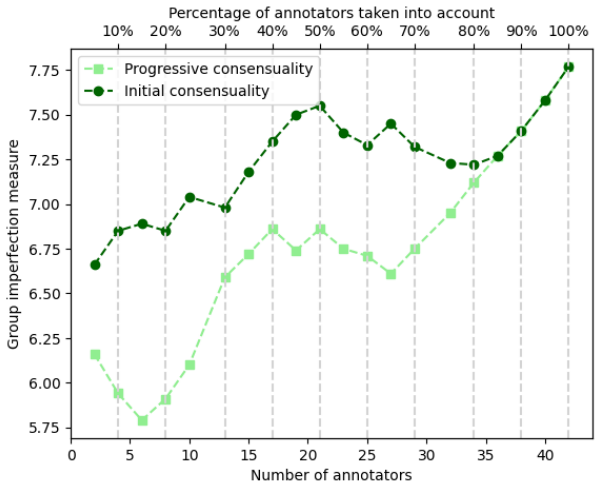

Figure 6: Imperfection of most consensual annotators

To support this last statement, we wanted to compare the average imperfection of the most consensual annotators found by each consensuality's type. This comparison can be seen in figure 6. For a given percentage of the most consensual annotators, we computed the average of their imperfection (in other words, the more their average tends towards 0, the better their group performance). From this, a first tendency emerges from the graph: the most consensual annotator groups according to the progressive consensuality (curve with square markers) obtain better averages than the groups with the same number of annotators found with the initial consensuality. Although the curves are not strictly monotonic, we also observe better group performance when the number of selected annotators is lower.

Thus, in light of the previous analyses and of this graph, one would have to prefer a small group of the most consensual annotators, selected through progressive consensuality, in order to collect better quality annotations. Of course, such a conclusion cannot be generalised without further similar experimentations on other data. However, we can already point to comparable experiments conducted in the context of crowdsourcing, where collecting annotations from a crowd of diverse annotators often leads to questions about the quality of the data produced. For example, (Passonneau et al., 2012) identify annotators whose annotations are most

distinct from other annotators and then removes them from the experiment; inter-annotator agreement is then significantly improved, enough to indicate reliable annotation. In (Inel et al., 2014), the authors prefer to study annotation disagreements in order to identify ambiguous items, but disagreements also allow them to identify annotators who stand out too much from other annotators.

6. Conclusion and future works

6.1. Answers to problems P1 and P2

This article is an attempt to clarify the problems P1 and P2 introduced in the first section. To enable progress towards their resolution, we proposed observation methods and tools, as well as an illustrative dataset, all of which will be made available to the community.

P1 – Does a good inter-annotator agreement necessarily give access to the truth? The answer is No. A good inter-annotator agreement does not necessarily imply access to the ground truth, contrary to a common hypothesis. This assumption should therefore be applied with caution. However, the consensuality, observed with care, may improve the establishment of the reference from multiple annotations.

P2 – How does the consensus emerge? Answering this question is the key to correctly exploit consensuality. A good agreement value means that annotators converge towards a consensusal annotation. In this paper, we have defined the notion of consensuality, which clarifies how an individual annotator behaves with respect to the rest of her group. While it is clear that individual consensuality does not necessarily depicts performance, we have nevertheless shown that, for this corpus and this group of annotators, the least consensual are also the least performant annotators.

6.2. Recommendations concerning annotation methodology

Through this study, we observed the relationships a) between inter-annotator agreement and access to the truth and b) between the performance of an annotator and her consensuality in relation to a group of annotators. From these experiements, we would like to highlight the following conclusions and recommendations:

- **Reusable ways of observations:** Of course, these experiments concern a very specific task on very specific data (age estimation of human based on photographs). Therefore, there is obviously no guarantee that our observations on these data also apply to other datasets (even if a counter example on a specific dataset may be sufficient to invalidate a general hypothesis). Let us emphasize that our observations are less important than the ways to achieve them, which are reproducible and constitue the main contribution of this paper. We recommend that managers of annotation campaign carry out similar observations on their specific datasets.

- **Partial reference as a necessity:** Unsurprisingly, we can confirm that agreement and consensuality are not systematically a guarantee of performance. Indeed, agreeing with others is only desirable insofar as the others are themselves efficient, which is unknown when a reference is missing. If a reference is available for the full dataset, their is obviously no need for multi-annotation, agreement or consensuality measure. But for all other cases (the most frequent ones), working without any partial reference would be a very risky business. A partial reference makes it possible to observe correlations between performance and agreement, performance and consensuality, for a given task and a given group of annotators, and to make decision for the whole dataset. Using iterative cycles of multi-annotation, agreement measure and deliberation, a partial reference should be established as a gold standard as soon as possible.

- **Consensuality measure to improve the annotation process:** Even if consensuality is not systematically a guarantee of performance, it makes it possible, in the context of our experiment, to some extent (which should be better specified with other experiments), to filter out some of the annotators (the least consensual ones) in order to increase the overall performance. It would obviously be premature and incorrect to conclude that a universal protocol may be defined to filter out undesirable annotators. Nevertheless, consensuality can be used to identify the least efficient annotators quite efficiently, in order to check their productions more carefully, to better understand their misunderstaning, to further train them and to update annotation instructions consequently.

- **Progressive consensuality better than initial one:** To identify the least consensual annotators, we have proposed and compared two types of consensuality. In the context of this campaign, progressive consensuality is clearly more efficient than initial consensuality. Insofar as the data coming from a not consensual annotator alter the whole dataset and the whole consensuality network, it seems more adequate, a priori, to compute progressive consensuality. We suggest however, pending further experimentation, to compare both consensualities to confirm that.[4]

- **Identify problematic items:** Agreement among annotators may vary significantly from one item

[4] It may be useful to note that they are very easy to implement, since they only require a measure of agreement (not necessarily chance corrected).

to another. It is then necessary to setup methods to monitor these variations and to identity problematic items. An interesting question concerns their presence in a reference corpus. In all cases, we recommand to keep track of the disagreement they are subject to. A subsequent training or evaluation system will thus be able to modulate the confidence of decisions made from these items.

6.3. Future works

The main goal of this paper was to highlight different methods and tests for measuring annotators' performance and consensuality. But this paper only provides a first approach of their complex relationship, which requires further investigations. Our next works will follow two main directions.

- **From partial reference to global reference thanks to homogeneous consensuality:** A first work will concern the possible contribution of consensuality when a partial reference is available. Indeed, in some annotation campaigns, it is possible but expensive to establish a reference. We wish to see which are the least expensive strategies allowing to obtain a reference on the whole corpus from a multiple annotation on the whole corpus and a reference on a fraction of it. In particular, we will study the following question: does a homogeneous consensuality on the whole corpus allows us to know which annotators are the most efficient on the whole corpus?

  To measure homogeneity of consensuality, we will split the corpus into several parts, and analyze to what extent the ordering of consensuality of the annotators remains the same on these different parts (e.g. by the average of the variance of the annotators' rank). We wish to see if a strong homogeneity allows us to ensure that the n% of annotators who are both the best performing and the most consensual on the part for which we have a reference are also the best performing n% on the whole corpus.

- **Collective intelligence and consensuality:** The dataset we are working on seems to be a good candidate to test the hypothesis of "collective intelligence", where the performance of a group may outperform the average performance of its contributing annotators. For exemple, let us consider a group of 2 annotators, and let us assume that for an item whose true age is 50, annotator 1 estimates 48, and annotator 2 estimates 52. The average absolute error is 2 years. But for a virtual collective annotator which averages the values estimated by the group, this item will be given the true value. Hence, we may expect a group response far better than what individual scores would suggest, in particular if some of the annotators mutually compensate their errors (for instance when one of them is prone to underestimate ages, whereas another one overestimates them).

  More generally, we intend to work with two kinds of group performance: the current one which averages individual performances, and a new one, corresponding to so-called collective intelligence, defined as the performance of the average virtual collective annotator. In this way, thanks to these two complementary kinds of performance, we hope to better understand the links between consensuality and group performance and to improve the methods for comparing initial and progressive consensualities (including observation tools of which Figure 6 gave a first simple overview).

7. Bibliographical References

Aroyo, L. and Welty, C. (2015). Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine*, 36(1):15–24, March.

Artstein, R. and Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Bayerl, P. S. and Paul, K. I. (2011). What Determines Inter-Coder Agreement in Manual Annotations? A Meta-Analytic Investigation. *Computational Linguistics*, 37(4):699–725, December.

Bradley, M. M. and Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report, The Center for Research in Psychophysiology, University of Florida.

Braffort, A., Chételat-Pelé, E., and Segouat, J. (2011). Corpus de langue des signes : situer les biais des méthodes d'annotation et d'analyse. *Corpus*, 10:25–40, November.

Bregeon, D., Antoine, J.-Y., Villaneau, J., and Lefeuvre-Halftermeyer, A. (2019). Redonner du sens à l'accord interannotateurs : vers une interprétation des mesures d'accord en termes de reproductibilité de l'annotation. *Traitement Automatique des Langues*, 60(2):23, September.

Clifford, C. W. G., Watson, T. L., and White, D. (2018). Two sources of bias explain errors in facial age estimation. *Royal Society Open Science*, 5(10), October.

Dandapat, S., Biswas, P., Choudhury, M., and Bali, K. (2009). Complex Linguistic Annotation – No Easy Way Out! A Case from Bangla and Hindi POS Labeling Tasks. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 10–18, Suntec, Singapore, August. Association for Computational Linguistics.

Fort, K. (2016). *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. Wiley.

Inel, O., Khamkham, K., Cristea, T., Dumitrache, A., Rutjes, A., van der Ploeg, J., Romaszko, L., Aroyo, L., and Sips, R.-J. (2014). CrowdTruth:

Machine-Human Computation Framework for Harnessing Disagreement in Gathering Annotated Data. In *The Semantic Web – ISWC 2014*, pages 486–504. Springer, Cham, Switzerland, October.

Kang, D., Ammar, W., Dalvi, B., van Zuylen, M., Kohlmeier, S., Hovy, E., and Schwartz, R. (2018). A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana, June. Association for Computational Linguistics.

Krippendorff, K. (2013). *Content analysis: An Introduction to its Methodology*. Sage publications.

Mathet, Y., Widlöcher, A., Fort, K., François, C., Galibert, O., Grouin, C., Kahn, J., Rosset, S., and Zweigenbaum, P. (2012). Manual Corpus Annotation: Giving Meaning to the Evaluation Metrics. In *Proceedings of COLING 2012 - Posters*, pages 809–818, Mumbai, India, December. The COLING 2012 Organizing Committee.

Mathet, Y., Widlöcher, A., and Metivier, J.-P. (2015). The Unified and Holistic Method Gamma for Inter-annotator Agreement Measure and Alignment. *Computational Linguistics*, 41(3):437–479.

Oortwijn, Y., Ossenkoppele, T., and Betti, A. (2021). Interrater Disagreement Resolution: A Systematic Procedure to Reach Consensus in Annotation Tasks. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 131–141, Online, April. Association for Computational Linguistics.

Passonneau, R. J., Bhardwaj, V., Salleb-Aouissi, A., and Ide, N. (2012). Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46(2):219–252, June.

Vestlund, J., Langeborg, L., Sörqvist, P., and Eriksson, M. (2009). Experts on age estimation. *Scandinavian Journal of Psychology*, 50(4):301–307, August.

Voelkle, M. C., Ebner, N. C., Lindenberger, U., and Riediger, M. (2012). Let me guess how old you are: effects of age, gender, and facial expression on perceptions of age. *Psychology and Aging*, 27(2):265–277, June.

Watson, T. L., Otsuka, Y., and Clifford, C. W. G. (2016). Who are you expecting? Biases in face perception reveal prior expectations for sex and age. *Journal of Vision*, 16(3):5, February.