Analyzing the Impact of Prevalence on the Evaluation of a Manual Annotation Campaign

Karën Fort[*,⋆], **Claire François**[*], **Olivier Galibert**[†], **Maha Ghribi**[*]

[*]INIST - CNRS
2 alle de Brabois, 54500 Vandoeuvre-ls-Nancy, France
{karen.fort,claire.francois,maha.ghribi}@inist.fr
[⋆]LIPN, Université Paris 13 & CNRS
99 av. J.B. Clément, 93430 Villetaneuse, France
[†] LNE
29 avenue Roger Hennequin, 78190 Trappes, France
olivier.galibert@lne.fr

Abstract

This article details work aiming at evaluating the quality of the manual annotation of gene renaming couples in scientific abstracts, which generates sparse annotations. To evaluate these annotations, we compare the results obtained using the commonly advocated inter-annotator agreement coefficients such as S, κ and π, the less known R, the weighted coefficients κ_ω and α as well as the F-measure and the SER. We analyze to which extent they are relevant for our data. We then study the bias introduced by prevalence by changing the way the contingency table is built. We finally propose an original way to synthesize the results by computing distances between categories, based on the produced annotations.

Keywords: manual annotation evaluation, inter-annotator agreement, prevalence

1. Introduction

Manual corpus annotation is often needed prior to Natural Language Processing (NLP) tasks, not only to train tools, but also to create a reference for evaluation. If it was demonstrated, among others by Alex et al. (2006) and Reidsma and Carletta (2008), that incoherent annotations lead to limited performance of the tools trained with them, the quality of the reference is seldom justified. Only few campaigns provide details on its creation and when inter-annotator agreement measures are given, they are in the form of a *de facto* standard, the "kappa", from Cohen (1960) or Carletta (1996), generally without any more precision.[1]

Di Eugenio and Glass (2004) showed the sensitivity of these coefficients to inter-annotator bias and to prevalence and the discussion remains open regarding the representativity of these coefficients and the necessity to present several (Passonneau, 2006). Artstein and Poesio (2008) produced a very interesting and complete review of the different computation modes for the inter-annotator agreement and discussed their usage in NLP tasks. However, it remains difficult to know which coefficient to use according to the characteristics of the data. We present in this article the evaluation we conducted on a manual annotation campaign, applying and comparing different methods. We then propose a new way to synthesize the results by computing similarities between categories based on the produced annotations.

[1]For more details, refer to the introduction of (Artstein and Poesio, 2008).

2. Overview of the Annotation Campaign

Within the framework of the Quæro program[2], experts were asked to manually annotate Bacillus Subtilis gene renaming couples in a 1,843 abstracts corpus. These abstracts were selected from Medline by a partner of the project, using gene names databases and a set of keywords denoting gene renaming relations. The resulting corpus includes more than 400,000 tokens.

This annotation aimed at, first, building a database of Bacillus Subtilis gene renaming couples, and second, training and evaluating automatic extraction tools. It was used for the BioNLP 2011 shared task[3] and is available for use for non-profit research purposes (see license).

This campaign allowed for the manual identification of approximately 200 renaming couples, such as:

> "*Inactivation of a previously unknown gene,* ***yqzB*** *(renamed* ***ccpN*** *for control catabolite protein of gluconeogenic genes [..]*".

We organized the campaign using the methodology proposed in (Bonneau-Maynard et al., 2005) and computed the inter-annotator agreement at the very beginning of the annotation process, in order to identify as early as possible the disagreements between annotators and modify the guidelines accordingly. To achieve this, we had two expert annotators (here A1 and A2) annotate the same sample of 93 files, i.e. more than 19,000 tokens, from which we then computed the inter-annotator agreement. It is important to note that no automatic pre-annotation was performed, as, to our knowledge, no present tool can recognize all the gene

[2]http://quaero.org/

[3]http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/BioNLP-ST/downloads/downloads.shtml

names and we did not want to risk missing some not pre-annotated renaming couples.

Those couples were manually annotated using Cadixe (Alphonse et al., 2004), an interface designed for named entity annotation, that does not allow for the direct annotation of relations. We therefore annotated the original name of the gene (*Former*), then its new name (*New*), with corresponding ids. The rest of the text is not annotated.It has to be noted that we are here in a very specific situation, where the relation is so simple that it can be reduced to tokens with ids and where the random baseline can be identified (which is often not the case, see for example (Alex et al., 2010) on named entities).

Obviously, some files (more than a third) do not contain any renaming, while others detail several. We obtained in average one renaming per file. Comparing two annotations with standard inter-annotator agreement measures also requires to define what the *markables* are, i.e. what is potentially "annotatable". In our case, this definition is reasonably simple: all the tokens are potentially marked. So, for each annotator, we associate the implicit category *Nothing* with all the unannotated tokens.

Table 1: Contingency table computed from all the tokens

		A1			
		Former	New	Nothing	Total
A2	Former	71	13	23	107
	New	8	69	15	92
	Nothing	7	8	18,840	18,855
	Total	86	90	18,878	19,054

The contingency table 1 is already quite informative, as it reveals the predominance of the baseline *Nothing* category, that represents more than 99% of the corpus, and thus shows that the annotated items are largely scattered. This is a situation of great prevalence of one category. It also shows that some renaming relations were incomplete, as, for both annotators, the number of *Former* and *New* is not equal (86 and 90 for A1 and 107 and 92 for A2). Furthermore, this imbalance is more important for A2 than for A1. The two annotators annotated almost the same number of tokens as *New* (90 for A1 vs 92 for A2), but A2 annotated a lot more tokens with the *Former* category than A1 did (86 for A1 vs 107 for A2). A1 is more likely to annotate *Nothing* than A2 and A2 is more likely to annotate *Former* than A1. Part of this can be explained by the fact that we considered that gene names are simple tokens, whereas in some rare cases, in particular operons (clusters of genes), one annotator chose to select more than one token.

3. Evaluating the Produced Annotations using Coefficients

We will use in the rest of the article the notations and the formulas from (Artstein and Poesio, 2008) concerning the inter-annotator agreement measures and formulas.

3.1. Using S, π and κ Coefficients

The most obvious measure for the inter-annotator agreement is the observed agreement (A_o), which corresponds to the proportion of items on which the annotators agree, i.e. the total number of items on which they agree divided by the total number of items, that is, in our case:

$$A_o = \frac{71 + 69 + 18,840}{19,054} = 0.996116$$

The result is extremely high, but cannot be used as such, as it does not take into account the possibility that the annotators select the same category for the same item by chance (A_e, expected agreement). To analyze our results we will use here coefficients described in (Artstein and Poesio, 2008) taking this expected agreement into account: S (Bennett et al., 1954), κ (Cohen, 1960) and π (Scott, 1955). The three of them are computed using the same formula:

$$S, \kappa, \pi = \frac{A_o - A_e}{1 - A_e}$$

These coefficients differ in the way the expected agreement (A_e) is computed, according to hypotheses on the behavior of the annotators, in case they annotate by chance. S assumes that the expected agreement follows a uniform distribution in the various categories (here 3). In our case, the expected agreement for S, A_e^S, is therefore computed in the following way:

$$A_e^S = \frac{1}{3} = 0.333333$$

$$S = 0.99417$$

The most important bias of this coefficient is that it is directly correlated to the number of categories and that, consequently, the higher the number of categories, the lower the expected agreement. It has to be noticed that it is generally low, as its maximum value is 1/2 (0.5) for two categories. We only present S here to show its proximity with Finn's R (see sub-section 3.2. below).

π (Scott, 1955), also known as K in (Siegel and Castellan, 1988) or $kappa$ in (Carletta, 1996) , also considers that the distributions made by the annotators by chance are equivalent, but it assumes that the items are not uniformly distributed into the categories and that this distribution can be estimated using the average category assignment realized by the annotators. In our case, the expected agreement for π, A_e^π, is therefore computed in the following way:

$$A_e^\pi = \frac{((\frac{86+107}{2})^2 + (\frac{90+92}{2})^2 + (\frac{18,878+18,855}{2})^2)}{19,054^2}$$

$$= 0.980464$$

$$\pi = 0.8012$$

As for κ (Cohen, 1960), it assumes in the way it models chance, that the distribution of items between categories may differ for each annotator. In this case, the probability for an item to be assigned to a category is the product of the probability that each annotator assigns it to this category. In our case, the expected agreement for κ, A_e^κ, is therefore computed in the following way:

$$A_e^\kappa = \frac{(86 \times 107) + (90 \times 92) + (18,878 \times 18,855)}{19,054^2}$$

$$= 0.980463$$

$$\kappa = 0.80121$$

If we compare the 3 coefficients, we observe that S is slightly lower than the observed agreement, and π and κ are similar, while being lower than A_o and S, which is coherent with the order $S \geqslant \pi$ and $\pi \leqslant \kappa$ described in (Artstein and Poesio, 2008). The high S value shows that the items are annotated according to a rationale that has nothing to do with chance. For a constant observed agreement, S only depends on the number of categories, it is therefore not sensitive to the items' distribution between categories, as opposed to π and κ (Di Eugenio and Glass, 2004). The authors of this article show that when categories are skewed, despite a high agreement on the dominant category, π and κ are sensitive to disagreements on small categories. According to the latest interpretations of inter-annotator reliability scales that state that "if a threshold needs to be set, 0.8 is a good value" (Artstein and Poesio, 2008), our κ and π can be considered as good, which is reassuring concerning the agreement reached on the two minority but meaningful categories.

3.2. Using Finn's R Coefficient

Faced with the same disproportion between categories in their annotation campaign, Laignelet and Rioult (2009) followed a suggestion from Hripcsak and Heitjan (2002) and used the R coefficient (Finn, 1970), that is proposed in the software environment for statistical computing R[4]. The R coefficient is computed in the following way:

$$R = 1 - \frac{Observed\,Variance}{Expected\,Variance}$$

the observed variance being the average variance on the annotated items and the expected variance being the variance of the uniform discrete distribution with n categories (hereafter $nb\ categories$), i.e.:[5]

$$Expected\,Variance = \frac{(nb\ categories)^2 - 1}{12}$$

In our case, we obtain $R = 0.9943713$. This value, very close to that of S (0.99417) can be explained by the fact that this coefficient models chance the same way S does, considering a uniform distribution of the categories. It is therefore no more affected than S by the distribution of items in the categories. We therefore claim that Finn's R is no more informative than S in cases of scattered annotations and asymmetry of categories.

3.3. Using Weighted Coefficients

According to (Artstein and Poesio, 2008), π and κ process all disagreements the same way and only weighted coefficients allow for giving more importance to some disagreements.

They describe two weighted coefficients: the weighted κ, κ_ω (Cohen, 1968) and α (Krippendorff, 1980; Krippendorff, 2004). The two coefficients are based on inter-annotator disagreements and use a distance between categories describing to which extent two categories are distinct. Artstein and Poesio (2008) discuss how to define this distance according to the annotation type, as it allows, among others, to process the annotation of complex structures by introducing several values of distances between annotations. The inconvenient of this method is that it makes the interpretation of the results more complex.

We have in our case, two meaningful categories, *Former* and *New*, and one less meaningful, *Nothing*. We consider that it is more important to identify the gene names couples than to determine the precedence of a name as compared to the other. Therefore, for us, the distance between *Former* and *New* should be less than that between these and *Nothing*. If we consider that it is twice as large, we will obtain the distances between categories described in table 2 (in the [0,1] interval):

Table 2: Example of distances between categories

	Former	New	Nothing
Former	0	0,5	1
New	0,5	0	1
Nothing	1	1	0

The weighted coefficients κ_ω and α are computed using the formula:

$$\kappa_\omega, \alpha = 1 - \frac{D_0}{D_e}$$

where D_0 stands for the observed disagreement between the annotators and D_e represents the expected disagreement, i.e. the disagreement appearing if the distribution is done by chance alone. The expected disagreement for κ_ω and α follows the same rationale as κ and π respectively, and includes the notion of distance between categories.[6] From the distances of the table 2, we obtain $\alpha = 0.8292$ and $\kappa_\omega = 0.8291$, values that are higher than π and κ. The weighted coefficients express the same disagreement but with lower values, hence raising the inter-annotator agreement.

The resulting coefficients are very high and show little bias. However, they seem to us somewhat uncertain, as they consider very heterogeneous categories in a similar way: two meaningful but minority categories (*Former* and *New*) and a less meaningful, majority one (*Nothing*). The problem here is therefore to ensure that these coefficients, computed on the three categories, reflect a significant agreement on the two meaningful categories, *Former* and *New*.

3.4. Using the F-measure and Slot Error Rate

The fact that we use all the tokens (or even only the gene names) as random baseline is an approximation: the *Nothing* category includes irrelevant tokens, the number of which is not precisely known. This situation is not unusual

[4] http://www.r-project.org/

[5] Finn (1970) does not detail the computation of the expected variance, but it can be found in the sources of the *irr* library of R. For a more thorough explanation, see: `http://mathworld.wolfram.com/DiscreteUniformDistribution.html`.

[6] It has to be noticed that if all the categories are perfectly distinct, $\alpha = \pi$ and $\kappa_\omega = \kappa$.

and is generally dealt with using other metrics, in particular the F-measure (see, for example (Alex et al., 2010)). The recall, precision and F-measure, as defined in the information retrieval field, are performance metrics that require only the annotated elements and no random baseline (as they do not take chance into account). According to Hripcsak and Heitjan (2002), the F-measure, i.e. the weighted harmonic mean of precision and recall, is equivalent to the average positive specific agreement among the annotators, here:

$$F = \frac{2C}{2C + 2S + \frac{2}{(1+\alpha)}D + \frac{2\alpha}{1+\alpha}I}$$

where C is the number of correct slots or agreement, S is the number of substitutions (incorrect slots), D is the number of deletions (missing slots), I is the number of insertions (spurious slots), with $\alpha = 1$, the most popular value, which is well adapted to our case as it allows us to use D and I in a symmetric way. In our case, this corresponds to:

$$F = \frac{2 \times (71 + 69)}{(2 \times (71 + 69)) + (2 \times (13 + 8)) + 23 + 15 + 7 + 8}$$

An interesting variant, that we call here F', implies that substitutions are considered half-correct when computing precision and recall, giving, for the balanced case:

$$F' = \frac{2C + S}{2C + 2S + D + I}$$

In our case, this corresponds to:

$$F' = \frac{(2 \times (71 + 69)) + 13 + 8}{(2 \times (71 + 69)) + (2 \times (13 + 8)) + 23 + 15 + 7 + 8}$$

An interesting characteristic of this variant is that it corresponds to the limit of the κ coefficient when the count of *Nothing* tends towards infinity (Hripcsak and Rothschild, 2005). Using these formulas, we end up, for Table 1, with $F = 0.747$ and $F' = 0.803$. The F' measure points out the interest of weighting different types of error differently, which has long been recognized in the systems evaluation side, giving birth to the Slot Error Rate (Makhoul et al., 1999; Galibert et al., 2010). This metric corresponds to an error enumeration methodology, where, for each error, a cost is given, and the total cost is divided by the number of annotations in the reference. We follow here the same rule as above, giving a half-point cost to substitution (considering there is something to annotate is half the work) and a full point for insertions and deletions.

$$SER = \frac{0.5S + D + I}{Reference\ entity\ count}$$

In our case, no annotation can be considered as a reference, so we propose to use the arithmetic mean of the number of annotations as the divider. We end up with a symmetric-SER of 0.339:

$$SER = \frac{0.5 \times (13 + 8) + 23 + 15 + 7 + 8}{0.5 \times (86 + 90 + 107 + 92))}$$

Mathematically, the symmetric-SER is the harmonic mean of the two oriented SERs, giving a structure similar to the F-measure. Instead, if we had chosen to give a full point cost to substitutions, the result would have been 0.395.
The SER allows for a much finer control on what is considered important in the annotation, which is very interesting from a system evaluation point-of-view, but on the other hand is hard to interpret, as there are no traditionally accepted limits above which the annotation is considered good enough.

4. Changing Points of View

4.1. Analyzing the Impact of Prevalence

4.1.1. Rebuilding the Contingency Table

To build the contingency table 1, we chose to take into account the total number of tokens (strings of characters separated by whitespace, annotation markers excluded), i.e. 19,054 (case 1).
Suppose now that we consider that gene names correspond to a specific subset of tokens in the texts, we could then use as total the number of gene names occurrences, that is 1,165 (case 2).[7] Note that this choice is questionable as, first, the reliability of the results depend on the exhaustivity of the dictionary, which, given the constant progress in the field, will never be sufficient and second, because it would mean neglecting the fact that the annotators often have to read the whole text to make decisions, the renaming being confirmed only at the end of the abstract. Table 3 shows the contingency table generated using the number of gene names occurrences to define the *Nothing* category. From this table, we obtain $S = 0.90472$, $\pi = 0.77557$, $\kappa = 0.77571$. F-measure and SER do not change from table 1 given that the cases "Nothing/Nothing" and "genes/genes" are not taken into account. Note that we were unable to compute Finn's R as we depend on partners for the gene names dictionary and that this intermediary result was not available. The three coefficients result in lower values and show the same differences between them. This demonstrates that, even if the role of items distribution and of the behavior of the annotators seems constant, the size of the category *Nothing* has an influence on the inter-annotator agreement. A second possible redesign of the contingency table is to consider only the meaningful categories, *Former* and *New* (case 3), as shown in table 4. Note that completely removing the *Nothing* category implies removing part of the results (the items annotated by only one annotator). We only use this redesign to eliminate the prevalence effect of

Table 3: Contingency table computed from identified gene names

		A1			
		Former	New	Nothing	Genes
A2	Former	71	13	23	107
	New	8	69	15	92
	Nothing	7	8	951	966
	Genes	86	90	989	1,165

[7] Results obtained by application of a gene names dictionary.

the *Nothing* category and to focus on the *Former*/*New* inversions. The obtained results should therefore be interpreted with caution.

Table 4: Contingency table without the *Nothing* category

		A1		
		Former	New	Total
A2	Former	71	13	84
	New	8	69	77
	Total	79	82	161

We obtain, in this case, $S = 0.73913$, $R = 0.73913$, $\pi = 0.73909$ and $\kappa = 0.73934$. These values are still high but lower than the previous one. The number of elements in each category is also rather small, which makes the disagreements more visible. In this case, the agreement on the 2 categories is important, F-measure then equals the observed agreement.

Table 5: Contingency table with grouped meaningful categories

		A1		
		Former/New	Nothing	Total
A2	Former/New	161	38	199
	Nothing	15	18,840	18,855
	Total	176	18,878	19,054

Finally, table 5 shows the results obtained by grouping together the two meaningful categories, *Former* and *New* (case 4). We then get $S = 0.99444$, $R = 0.99444$, $\pi = 0.85726$, $\kappa = 0.85727$ and $F = 0.85867$. These values are higher than the various coefficients obtained from the complete contingency table 1, which is not surprising as in this configuration, there is very little ambiguity left.
Note that κ is again very similar to the F-measure as the number of *Nothing* is still very high.

4.1.2. Analyzing the Obtained Results

All the results we obtained are summarized in table 6. Note that the SER, with its error typing, is relevant only for the complete localization and typing task.
This table shows that R and S are very close in all cases. This confirms our remarks in section 3.1.: R does not bring any more information than S.
This table also shows that the values of π and κ, computed from all the contingency tables are very close. The way chance is modeled in π implies that the distribution into categories is the same for both annotators, whereas in κ, chance is modeled in such a way that this distribution differs from one annotator to the other. Similar values of π and κ reflect that both annotators generate the same distribution into categories, which can be seen in the similar marginal distributions. This means that our data show little annotator bias (Artstein and Poesio, 2008).
If we consider, on the one hand S, and on the other hand κ and π, we can see that their values are quite different in all the three cases taking into account the *Nothing* category, whereas they are similar in the case taking only *Former* and *New* into account. This can be explained by the fact that the distribution of the annotations into the three categories (including *Nothing*) is not homogeneous. κ and π use this distribution in the way they model chance, which is not the case for S. This does not appear in case 3, where the annotations are homogeneously distributed in the two categories (but again, we removed some of the annotations in this case). Table 5 (case 4) can be used to check if the gene renaming couples are correctly identified in the texts. The values of the coefficients we obtained with this table are the highest, we can therefore conclude that this identification is done without problem. The coefficients computed from case 1 and 2, when compared, show the impact of the *Nothing* category. The fact that the values of the coefficients are higher for case 1 than for case 2, in which the *Nothing* category is much smaller, shows that these coefficients are influenced by the prevalence in the annotations.
In this case, F-measure and SER are more adapted to evaluate the inter-annotator agreement, even though κ and π are sensitive to disagreements on small categories.
The coefficients obtained in case 3 present the lower values. They show in a more precise way the difficulties encountered when distinguishing between *Former* and *New*. In this table, the four coefficients are almost identical, which shows that the (partial) inter-annotator agreement is not biased by the different models. Therefore, the inter-annotator agreement reaches higher values when annotators have to identify gene name couples involved in a renaming relationship than when they have to identify as *Former* and *New* these gene names within these couples.
Comparing various coefficients is therefore useful to estimate the biases induced by the distribution of the annotations and the behavior of the annotators. Table 6 also shows the influence of prevalence on the coefficients, which means that the way we choose to consider the categories in the contingency table has a significant impact on the results. We therefore claim that it is fundamental, when giving inter-annotator agreement results, not only to present the contingency table that was used to compute the coefficients, but also to justify the choices that were made.

4.2. Using Similarities Between Categories

We saw that in the computation of weighted coefficients, distances between categories are defined from prior knowledge of the annotation task. As tempting as it may seem, computing distances from the annotated data themselves would imply some kind of circularity. However, such distances could prove useful to get some information on the categories themselves, independently from the annotators. We showed that the *Former* and *New* categories tend to be more difficult to identify within gene names couples than these couples from the whole text. The role of coefficients is not to provide this type of interpretation, which corresponds more to similarities between categories. We therefore propose to directly evaluate these similarities according to the difficulty the annotators have to distinguish between categories. In order to do this, we use the contingency table 1. We consider that two categories are distinct

Table 6: A_o, S, R, π, κ, F-measure and SER using various contingency tables

Contingency tables	A_o	S	R	π	κ	F-measure	SER
Former/New/Nothing (case 1)	0.99611	0.99417	0.99437	0.8012	0.80121	0.74667	0.33867
Former/New/Nothing gene names (case 2)	0.93648	0.90472	n/a	0.77557	0.77571	0.74667	0.33867
Former/New (case 3)	0.86956	0.73913	0.73913	0.73909	0.73934	0.86957	-
Former+New/Nothing (case 4)	0.99722	0.99444	0.99444	0.85726	0.85727	0.85867	-

if there is little chance of distribution error between them. More precisely, let us consider two categories C_1 and C_2 from the considered categories, $P(C_2|C_1)$ represents the probability that an annotator assigned an item in the category C_2 while a second annotator assigned it in the category C_1. It is computed in the following way:

$$P(C_2|C_1) = \frac{n_{1C_1,2C_2} + n_{2C_1,1C_2}}{n_{C_1}}$$

with $n_{1C_1,2C_2}$ representing the number of items assigned by annotator 1 to the C_1 category while annotator 2 assigned them to the C_2 category; n_{C_1} represents the sum of the items assigned in the category C_1 by both annotators. When this probability is low, the C_2 category is highly dissimilar to C_1 and the risk of getting a different annotation is low. We obtain here:

$$P(New|Former) = \frac{13 + 8}{107 + 86} = 0.108808$$

Table 7 presents the values of the probabilities computed for our case. The diagonal results give an estimate of the agreement for each category. We can see that it is very important for *Nothing* and less so for the others (73% for *Former* and 75% for *New*). The other cells in the table can be used to estimate the disagreement between annotators, category by category. These probabilities are very low, which means that their are highly dissimilar. We can also notice that the probabilities are asymmetrical. The values $P(Former|Nothing)$ and $P(New|Nothing)$ are very low (<1‰), therefore, the chance of annotating an item with the *Former* or *New* category when it has already been annotated *Nothing* is close to zero. Conversely, the chance of annotating an item with the *Nothing* category when it has already been annotated *Former* or *New* is higher (15% and 12%).

Table 7: Table of Probabilities

↙	Former	New	Nothing
Former	0.735751	0.108808	0.155440
New	0.115385	0.758242	0.126374
Nothing	0.000795	0.000609	0.998595

The probabilities being asymmetrical, this formula cannot be used as such. We will assume that the annotators would produce similar distributions of items among categories. We therefore define the associated similarity as the average of the oriented probabilities (computed from table 7), using:

$$Sim(C_1, C_2) = \frac{P(C_2|C_1) + P(C_1|C_2)}{2}$$

Table 8: Similarities between categories

	Sim
Sim(Former,New)	0.112096
Sim(Former,Nothing)	0.078117
Sim(New,Nothing)	0.063491

In table 8, we notice that $Sim(Former, New)$ is higher than $Sim(Former, Nothing)$ and $Sim(New, Nothing)$, implying that *Former* and *New* are closer to each other than to *Nothing*. To our knowledge, this kind of table has never been used before, although it proves quite useful, in particular during the preliminary stages of the annotation campaign, when the categories are tested and questioned, as it allows for the identification of subsets of categories that might be ambiguous. Moreover, it allows for a synthetic view of the data, even with more than 2 annotators. From this point of view, it shows a higher usability than the table presented by Krippendorff (2004), and a solution to the impossibility to show a contingency table when more than 2 annotators are involved.

5. Conclusion

We used the results from a real annotation campaign to analyze several computation modes of the inter-annotator agreement. A characteristic of this campaign is the highly scattered annotations, inducing a bias due to the prevalence of the unannotated tokens. We confirmed in this article that whenever possible, the first result to present is the contingency table (Hripcsak and Heitjan, 2002) , with precise explanations about the choices that were made. Our results indicate a good agreement on the two minority but meaningful categories. Comparing the coefficients and studying their evolution according to the way the contingency table is built allowed us to check that there was no bias due to the annotators and to quantify the prevalence bias.

Finally, to obtain an analysis of the real chance of error between categories, we computed a table of similarities between them. This table, allowing for a synthetic view of the data, even with more than 2 annotators, constitutes a new tool for evaluation which, as a complement to coefficients like κ or F-measure, offers a different view on the data, more category-oriented. New annotation campaigns held within the same program should allow us to test the different coefficients and the reproducibility of our proposals. These campaigns concern various domains and applications, such as the patents in pharmacology (named entities, terms) or soccer matches comments (named entities, complex relations).

Acknowledgments

This work was realized as part of the Quæro Program[8], funded by OSEO, French State agency for innovation. We want to thank here all the participants in the campaign, in particular the INRA MIG team, as well as F. Tisserand and B. Taliercio, the annotators from INIST-CNRS.

6. References

Beatrice Alex, Malvina Nissim, and Claire Grover. 2006. The Impact of Annotation on the Performance of Protein Tagging in Biomedical Text. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 595–600, Genoa, Italy, 24-26 May.

Beatrice Alex, Claire Grover, Rongzhou Shen, and Mijail Kabadjov. 2010. Agile corpus annotation in practice: An overview of manual and automatic annotation of cvs. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW)*, pages 29–37, Uppsala, Sweden. Association for Computational Linguistics.

Erick Alphonse, Sophie Aubin, Philippe Bessières, Gilles Bisson, Thierry Hamon, Sandrine Laguarigue, Adeline Nazarenko, Alain-Pierre Manine, Claire Nédellec, Mohamed Ould Abdel Vetah, Thierry Poibeau, and Davy Weissenbacher. 2004. Event-based Information Extraction for the Biomedical the CADERIGE Project. In *Proceedings of the JNLPBA COLING 2004 Workshop*, Geneva, Switzerland.

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

Edward M. Bennett, R. Alpert, and A. C.Goldstein. 1954. Communications through Limited Questioning. *Public Opinion Quarterly*, 18(3):303–308.

Hélène Bonneau-Maynard, Sophie Rosset, Christelle Ayache, Anne Kuhn, and Djamel Mostefa. 2005. Semantic Annotation of the French Media Dialog Corpus. In *Proceedings of the InterSpeech*, Lisboa, Portugal.

Jean Carletta. 1996. Assessing Agreement on Classification Tasks: the Kappa Statistic. *Computational Linguistics*, 22:249–254.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Jacob Cohen. 1968. Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin*, 70(4):213–220.

Barbara Di Eugenio and Michael Glass. 2004. The Kappa Statistic: a Second Look. *Computational Linguistics*, 30(1):95–101.

R. H. Finn. 1970. A Note on Estimating the Reliability of Categorical Data. *Educational and Psychological Measurement*, 30:71–76.

Olivier Galibert, Ludovic Quintard, Sophie Rosset, Pierre Zweigenbaum, Claire Nédellec, Sophie Aubin, Laurent Gillard, Jean-Pierre Raysz, Delphine Pois, Xavier Tannier, Louise Deléger, and Dominique Laurent. 2010. Named and Specific Entity Detection in Varied Data: the Quaero Named Entity Baseline Evaluation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*.

George Hripcsak and Daniel F. Heitjan. 2002. Measuring Agreement in Medical Informatics Reliability Studies. *Journal of Biomedical Informatics*, 35(2):99–110.

George Hripcsak and Adam S Rothschild. 2005. Agreement, the f measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association (JAMIA)*, 12(3):2968.

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

Klaus Krippendorff, 2004. *Content Analysis: An Introduction to Its Methodology, second edition*, chapter 11. Sage, Thousand Oaks, CA., USA.

Marion Laignelet and François Rioult. 2009. Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In *Proceedings of the Traitement Automatique des Langues Naturelles (TALN 2009)*, Senlis, France.

John Makhoul, Francis Kubala, Richard Schwartz, and Ralph Weischedel. 1999. Performance measures for information extraction. In *In Proceedings of DARPA Broadcast News Workshop*, pages 249–252.

Rebecca Passonneau. 2006. Measuring Agreement on Set-Valued Items (MASI) for Semantic and Pragmatic Annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*.

Denis Reidsma and Jean Carletta. 2008. Reliability Measurement Without Limits. *Computational Linguistics*, 34(3):319–326.

William A Scott. 1955. Reliability of Content Analysis: The Case of Nominal Scale Coding. *Public Opinion Quaterly*, 19(3):321–325.

Sidney Siegel and N. John Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, USA, 2nd edition.

[8] http://quaero.org/