The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment

Yann Mathet*
Université de Caen Basse-Normandie,
GREYC, CNRS, UMR 6072

Antoine Widlöcher*
Université de Caen Basse-Normandie,
GREYC, CNRS, UMR 6072

Jean-Philippe Métivier*
Université de Caen Basse-Normandie,
GREYC, CNRS, UMR 6072

Agreement measures have been widely used in computational linguistics for more than 15 years to check the reliability of annotation processes. Although considerable effort has been made concerning categorization*, fewer studies address* unitizing*, and when both paradigms are combined even fewer methods are available and discussed. The aim of this article is threefold. First, we advocate that to deal with* unitizing*, alignment and agreement measures should be considered as a unified process, because a relevant measure should rely on an alignment of the units from different annotators, and this alignment should be computed according to the principles of the measure. Second, we propose the new versatile measure γ, which fulfills this requirement and copes with both paradigms, and we introduce its implementation. Third, we show that this new method performs as well as, or even better than, other more specialized methods devoted to categorization or segmentation, while combining the two paradigms at the same time.*

1. Introduction

A growing body of work in computational linguistics (CL hereafter) or natural language processing manifests an interest in corpus studies, and requires reference annotations for system evaluation or machine learning purposes. The question is how to ensure that an annotation can be considered, if not as the "truth," than at least as a suitable

* Normandie University, France; UNICAEN, GREYC, F-14032 Caen, France; CNRS, UMR 6072, F-14032 Caen, France. E-mail: {yann.mathet, antoine.widlocher, jean-philippe.metivier}@unicaen.fr

Submission received: 15 July 2013; revised version received: 5 October 2014; accepted for publication: 12 January 2015.

doi:10.1162/COLI_a_00227

reference. For some simple and systematic tasks, domain experts may be able to annotate texts with almost total confidence, but this is generally not the case when no expert is available, or when the tasks become harder. The very notion of "truth" may even be utopian when the annotation process includes a certain degree of interpretation, and we should in such cases look for a consensus, also called the "gold standard," rather than for the "truth."

For these reasons, a classic strategy for building annotated corpora with sufficient confidence is to give the same annotation task to several annotators, and to analyze to what extent they agree in order to assess the reliability of their annotations. This is the aim of inter-annotator agreement measures. It is important to point out that most of these measures do not evaluate the distance from annotations to the "truth," but rather the distance across annotators. Of course, the hope is that the annotators will agree as far as possible, and it is usually considered that a good inter-annotator agreement ensures the constancy and the reproducibility of the annotations: When agreement is high, then the task is consistent and correctly defined, and the annotators can be expected to agree on another part of the corpus, or at another time, and their annotations therefore constitute a consensual reference (even if, as shown for example by Reidsma and Carletta [2008], such an agreement is not necessarily informative for machine learning purposes). Moreover, once several annotators reach good agreement on a given part of a corpus, then each of them can annotate alone other parts of the corpus with great confidence in the reproducibility (see the preface to Gwet [2012, page 6] for illuminating considerations). Consequently, inter-annotator agreement measurement is an important point for all annotation efforts because it is often considered that a given agreement value provided by a given method validates or invalidates the consistency of an annotation effort.

How to measure agreement, and how we define a good measure, is another part of the problem. There is no universal answer, because how to measure depends on the nature of the task, hence on the kind of annotations.

Admittedly, much work has already been done for some kinds of annotation efforts, namely, when annotators have to choose a category for previously identified entities. This approach, which we will call **pure categorization**, has led to several well-known and widely discussed coefficients such as κ, π, or α, since the 1950s. Some more recent efforts have been made in the domain of **unitizing**, following Krippendorff's terminology (Krippendorff 2013), where annotators have to identify by themselves what the elements to be annotated in a text are, and where they are located. Studies are scarce, however, as Krippendorff pointed out: "Measuring the reliability of unitizing has been largely ignored in favor of coding predefined units" (Krippendorff 2013, page 310). This scarcity concerns either **segmentation**, where annotators simply have to mark boundaries in texts to separate contiguous segments, or more generally **unitizing**, where gaps may exist between units. Moreover, some even more complex configurations may occur (overlapping or embedding units), which are more rarely taken into account.

And **when categorization meets unitizing**, as is the case in CL in such fields as, for example, NAMED ENTITY RECOGNITION[1] or DISCOURSE FRAMING, very few methods are proposed and discussed. That is the main problem we focus on in this article and to which γ provides solutions.

1 Small caps are used to refer to the examples of annotation tasks introduced in Section 2.2.

The new coefficient γ that is introduced in this article is an agreement measure concerning the joint tasks of unit locating (**unitizing**) and unit labeling (**categorization**). It relies on an **alignment** of units between different annotators, with penalties associated with each positional and categorial discrepancy. The alignment itself is chosen to minimize the overall discrepancy in a **holistic** way, considering the full continuum to make choices, rather than making local choices. The proposed method is **unified** because the computation of γ and the selection of the best alignment are interdependent: The computed measure depends on the chosen alignment, whose selection depends on the measure.

This method and the principles proposed in this article have been built up since 2010, and were first presented to the French community in a very early version in Mathet and Widlöcher (2011). The initial motivation for their development was the lack of dedicated agreement measures for annotations at the discourse level, and more specifically for annotation tasks related to TOPIC TRANSITION phenomena.

The article is organized as follows. First, we fix the scope of this work by defining the important notions that are necessary to characterize annotation tasks and by introducing the examples of linguistic objects and annotation tasks used in this article to compare available metrics. Second, we analyze the state of the art and identify the weaknesses of current methods. Then, we introduce our method, called γ. As this method is new, we compare it to the ones already in use, even in their specialized fields (pure categorization, or pure segmentation), and show that it has better properties overall for CL purposes.

2. Motivations, Scope, and Illustrations

We focus in the present work on both categorizing and unitizing, and consider therefore annotation tasks where annotators are not provided with preselected units, but have to locate them and to categorize them at the same time. An example of a multi-annotated **continuum** (this continuum may be a text or, for example, an audio or a video recording) is provided in Figure 1, where each line represents the annotations of a given annotator, from left to right, respecting the continuum order.

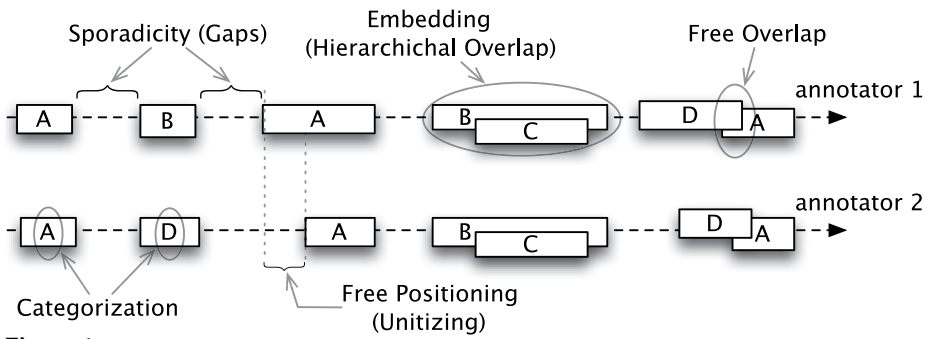


Figure 1
Multi-annotation including unitizing and categorizing.

2.1 Properties of Annotation Tasks and Annotated Items

In order to characterize the annotation efforts focusing on specific linguistic objects, we consider the following properties, illustrated in Figure 1.

Categorization occurs when the annotator is required to label (predefined or not) units.

Unitizing occurs when the annotator is asked to identify the units in the continuum: She has to determine each of them (and the number of units that she wants) and to locate them by positioning their boundaries.

Embedding (hierarchical overlap) may occur if units may be embedded in larger ones (of the same type, or not).

Free overlap may occur when guidelines tolerate the partial overlap of elements (mainly of different types). Embedding is a special case of overlapping. A segmentation without overlap (hierarchical or free) is said to be strictly **linear**.

Full-covering (vs. sporadicity) applies when all parts of the continuum are to be annotated. For other tasks, parts of the continuum are selected.

Aggregatable types or instances correspond to the fact that several adjacent elements having the same type may aggregate, without shift in meaning, in a larger span having the same type. This larger span is said to be **atomizable**: Labeling the whole span or labeling all of its atoms are considered as equivalent, as illustrated by Figure 2.

Two specific cases. We call hereafter **pure segmentation** (illustrated by Figure 3) the special case of unitizing with full-covering and without categorization, and we call **pure categorization** categorization without unitizing.

2.2 Examples of Annotation Tasks

To present the state of the art as well as our own propositions, and to make all of them more concrete, it is useful to mention examples of linguistic objects and annotation tasks for which agreement measures may be required. The following sections will then refer to these examples as often as possible, in order to illustrate discussions on abstract problems or configurations. Small caps are used to refer to the names of these tasks.

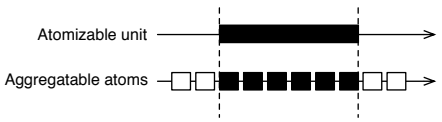


Figure 2
Aggregatable atoms and atomizable unit.

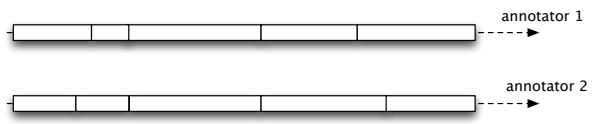


Figure 3
The particular case of pure segmentation.

Table 1
Properties associated with some examples of annotation tasks. ■: mandatory, □: possible.

Annotation tasks	Categorization	Unitizing	Embedding	Free Overlap	Sporadicity	Aggregatable
PART-OF-SPEECH	■	□				
GENE RENAMING	■				■	
WORD SENSE	■	□			■	
NAMED ENTITY	■	■	□	□	■	
ARGUMENTATIVE ZONING	■	□			■	■
DISCOURSE FRAMING	■	■	□	□	■	
COMMUNICATIVE BEHAVIOR	■	■	□	□	□	□
DIALOG ACT	■	■	□	□	□	
TOPIC SEGMENTATION		■				
HIERARCHICAL TOPIC SEGMENTATION		■	■			
TOPIC TRANSITION	■	■	■	■	□	
ENUMERATIVE STRUCTURES	■	■	■	□	■	

Table 1 summarizes the properties of the linguistic objects and annotation tasks mentioned in this article to illustrate and compare methods and metrics. These objects and tasks are briefly described for convenience in Appendix A. This table shows that annotation of TOPIC TRANSITIONS is the most demanding of the tasks regarding the number of necessary criteria a suitable agreement metric should assess, but most of the tasks listed here require assessment of both unitizing and categorization.

3. State of the Art

As we saw in the previous section, different studies in linguistics or CL involve quite different structures, which may lead to annotation guidelines having very different properties. They require suitable metrics in order to assess agreement among annotators. As we will see, some of the needs for which γ is suitable are not satisfied by other available metrics.

Note that this description of the state of the art mainly focuses on the questions which are of most importance for this work, in particular, chance correction and unitizing. For a thorough introduction to the most popular measures that concern categorizing, we refer the reader to the excellent survey by Artstein and Poesio (2008).

In this section, we first address the question of chance correction in agreement measures, then we give an overview of available measures in three domains: pure categorization, pure segmentation, and unitizing.

3.1 Agreement Measures and Chance Correction

We begin the state of the art with the question of chance correction, because it is a cross-cutting issue in all agreement measure domains, and because it influences the final value provided by most agreement measures.

It is important to distinguish between (1) measures to evaluate systems, where *the output of an annotating system is compared to a valid reference*, and (2) inter-annotator agreement measures, which try to quantify *the degree of similarity between what different*

annotators say about the same data, and which are the ones we are really concerned with in this article.

In case (1), the result is straightforward, providing for instance the percentage of valid answers of a system: We know exactly how far the evaluated system is from the gold standard, and we can compare this system to others just by comparing their results.

However, case (2) is more difficult. Here, measures do not compare annotations from one annotator to a valid reference (and, most of the time, no reference already exists), but they compare annotations from different annotators. As such, they are clearly not direct distances to the "truth." So, the question is: Above what amount of agreement can we reasonably trust the annotators? The answer is not straightforward, and this is where chance correction is involved.

For instance, consider a task where two annotators have to label items with 10 categories. If they annotate at random (with the 10 categories having equal prevalence), they will have an agreement of 10%. If we consider another task involving two categories only, still at random, the agreement expected by chance rises to 50%. Based on this observation, most agreement measures try to remove chance from the observed measure, that is to say, to provide the amount of agreement that is above chance. More precisely, most agreement measures (for about 60 years, with well-known measures κ, S, π) rely on the same formula: If we note A_o the **observed agreement** (i.e., the agreement directly observed between annotators) and A_e the so-called **expected agreement** (i.e. the agreement which should be obtained by chance), the final agreement A is defined by Equation (1).

To illustrate this formula, assume that the observed agreement is seemingly high, say $A_o = 0.9$. If $A_e = 0.5$, $A = 0.4/0.5 = 0.8$, which is still considered as good, but if $A_e = 0.7$, $A = 0.2/0.3 = 0.67$, which is not that good, and if $A_e = 0.9$, which means annotators did not perform better than chance, then $A = 0$.

Some other measures, namely, all α from Krippendorff, and the new γ introduced in this article, are computed from **observed** and **expected disagreements** (instead of agreements), denoted here respectively D_o and D_e, and they define the final agreement by Equation (2).

$$A = \frac{A_o - A_e}{1 - A_e} \tag{1}$$

$$A = 1 - \frac{D_o}{D_e} \tag{2}$$

However, the way the expected value is computed is the only difference between many coefficients (κ, S, π, and their generalizations), and is a controversial question. As precisely described in Artstein and Poesio (2008), there are three main ways to model chance in an annotation effort:

1. By considering a **uniform distribution**. For instance, in a categorization task, considering that each category (for each coder) has the same probability. The limitation of this approach is that it provides a poor model for chance annotation. Moreover, for a given task, the greater the number of categories, the lesser the expected value, hence the higher the final agreement.

2. By considering the *mean distribution* of the different annotators hence regarded as *interchangeable*. For instance, in a categorization task with two categories A and B, where the prevalences are respectively 90% for category A and 10% for category B, the expected value is computed as $0.9 \times 0.9 + 0.1 \times 0.1 = 0.82$, which is much higher than the 0.5 obtained by considering a uniform distribution.
3. By considering the *individual distributions* of annotators. Here, annotators are considered as *not interchangeable*; each of them is considered to have her own probability for each category (for a categorization task) based on her own observed distribution. It leads to the same results as with the mean distribution if annotators all have the same distribution, or to a lesser value (hence a higher final agreement) if not.

In the two cases of mean and individual distributions, expected agreement may be very high, depending on the *prevalence of categories*. In some annotation tasks, expected agreement becomes critically high, and any disagreements on the minor category have huge consequences on the chance-corrected agreement, as hotly debated by Berry (1992) and Goldman (1992), and criticized in CL by Di Eugenio and Glass (2004). However, we follow Krippendorff (2013, page 320), who argues that disagreements on rare categories are more serious than on frequent ones. For instance, let us consider the reliability of medical diagnostics concerning a rare disease that affects one person out of 1,000. There are 5,000 patients, 4,995 being healthy, 5 being affected. If doctors fail to agree on the 5 affected patients, their diagnostics cannot be trusted, even if they agree on the 4,995 healthy ones.

These principles have been mainly introduced and used for categorization tasks, because most coefficients address these tasks, but they are more general and may also concern segmentation and, as we will see further, unitizing.

3.2 Measures for Pure Categorization

The simplest measure of agreement for categorization is **percentage of agreement** (see for example Scott 1955, page 323). Because it does not feature chance correction, it should be used carefully for the reasons we have just seen.

Consequently, most popular measures are chance-corrected: S (Bennett, Alpert, and Goldstein 1954) relies on a uniform distribution model of chance, π (Scott 1955) and α (Krippendorff 1980) on the mean distribution, and κ (Cohen 1960) on individual distributions. Generalizations to three or more annotators have been provided, such as κ (Fleiss 1971), also known as K (Siegel and Castellan 1988). Moreover, **weighted coefficients** such as α and κ_w (Cohen 1968) are designed to take into account the fact that disagreements between two categories are not necessarily all of the same importance. For instance, for scaled categories from 1 to 10 (as opposed to so-called nominal categories), a mistake between categories 3 and 4 should be less penalized than a mistake between categories 1 and 10.

These metrics, widely used in CL, are suitable to assess agreement for pure categorization tasks—for example, in the domains of PART-OF-SPEECH TAGGING, GENE RENAMING, or WORD SENSE ANNOTATION.

From Carletta (1996) to Artstein and Poesio (2008), most of these methods have already been discussed and compared in the perspective of CL and we will not do so here.

3.3 Measures for Segmentation

In the domain of TOPIC SEGMENTATION, several measures have been proposed, especially to evaluate the quality of automatic segmentation systems. In most cases, this evaluation consists in comparing the output of these systems with a reference annotation. We mention them here because their use tends to be extended to inter-annotator agreement because of the lack of dedicated agreement measures, as illustrated by Artstein and Poesio (2008), who mention these metrics in a survey related to inter-annotator agreement, or by Kazantseva and Szpakowicz (2012).

In this domain, annotations consist of boundaries (between topic segments), and the penalty must depend on the distance from a true boundary. Thus, dedicated measures have been proposed, such as WindowDiff (WD hereafter; Pevzner and Hearst 2002), based on *Pk* (Beeferman, Berger, and Lafferty 1997). WD relies on the following principle: A fixed-sized window slides over the text and the numbers of boundaries in the system output and reference are compared. Several limitations of this method have been demonstrated and adjustments proposed, for example, by Lamprier et al. (2007) or by Bestgen (2009), who recommends the use of the Generalized Hamming Distance (GHD hereafter; Bookstein, Kulyukin, and Raita 2002), in order to improve the stability of the measure, especially when the variance of segment size increases.

Because these metrics are dedicated to the evaluation of automatic segmentation systems, their most serious weakness for assessing agreement is that they are not chance-corrected, but they present another limitation: They are dedicated to segmentation and assume a full-covering and linear tiling of the continuum and only one category of objects (topic segments). This strong constraint makes them unsuitable for unitizing tasks using several categories (ARGUMENTATIVE ZONING), targeting more sporadic phenomena (ANNOTATION OF COMMUNICATIVE BEHAVIOR), or involving more complex structures (NAMED ENTITY RECOGNITION, HIERARCHICAL TOPIC SEGMENTATION, TOPIC TRANSITION, DISCOURSE FRAMING, ENUMERATIVE STRUCTURES).

3.4 Measures for Unitizing

3.4.1 Using Measures for Categorization to Measure Agreement on Unitizing. Because of the lack of dedicated measures, some attempts have been made to transform the task of unitizing into a task of categorizing in order to use well-known coefficients such as κ.

They consist of atomizing the continuum by considering each segment as a sequence of atoms, thereby reducing a unitizing problem to a categorization problem. This is illustrated by Figure 4, where real unitizing annotations are on the left (with two annotators), and the transformed annotations are on the right. To do so, an atom granularity is chosen—for instance, in the case of texts, it may be character, word, sentence, or paragraph atoms. Then, each unit is transformed into a set of items labeled with the category of this unit, and a new "blank" category is added in order to emulate gaps between units.

In most cases, this method has severe limitations:

1. Two contiguous units seen as one. In zone (1) of the left part of Figure 4, one annotator has created two units (of the same category), and the other annotator has created only one unit covering the same space. However, once the continua are discretized, the two annotators seem to agree on this zone (with the four same atoms), as we can see in the right part of the figure.

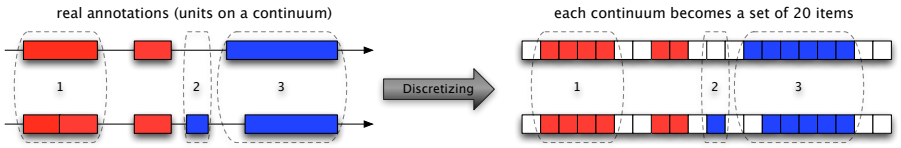


Figure 4
Discretizing the continua.

2. False positive/negative disagreement and slight positional disagreement considered as the same. Zone (2) of Figure 4 shows a case where annotators disagree on whether there is a unit or not, which is quite a severe disagreement, whereas zone (3) shows a case of a slight positional disagreement. Surprisingly, these two discrepancies are counted with the same severity, as we can see in the right side of the figure, because in each case, there is a difference of category for one item only (respectively, "blank" with "blue" in case (2), and "blue" with "blank" in case (3)).

3. Agreement on gaps. Because of the discretization, artificial blank items are created, with the result that annotators may agree on "blanks." The more gaps in real annotations, the more artificial "blank" agreement, and hence the greater the artificial increase in global agreement. Indeed, the expected agreement is less impacted by artificial "blanks," and it may even decrease.

4. Overlapping and embedding units are not possible. This results because of the discretizing process, which requires a given position to be assigned a single category (or it would require creating as many artificial categories as possible combinations of categories).

This kind of reduction is used to evaluate the annotation of COMMUNICATIVE BEHAVIOR in video recordings by Reidsma (2008), where unitizing is, on the contrary, clearly required: The time-line is discretized (*atomized*), then κ and α are computed using discretized time spans as units. It should be noted that Reidsman, Heylen, and Ordelman (2006, page 1119) and Reidsma (2008) claim that this "fairly standard" method (which we call **discretizing measure** henceforth) has certain drawbacks, such as the fact that it "does not compensate for differences in length of segments," whereas "short segments are as important as long segments" in their corpus (which is an additional limitation to the ones we have just mentioned). They propose a second approach relying on an alignment, as we mention in Section 4.2.1.

This reduction is also unacceptable for other annotation tasks. For instance, in the perspective of DISCOURSE FRAMING, two adjacent temporal frames should not be aggregated in a larger one. In the same manner, for TOPIC SEGMENTATION, it clearly makes no sense to aggregate two consecutive segments.

3.4.2 A Measure for Unitizing Without Chance Correction. Another approach, derived from Slot Error Rate (Makhoul et al. 1999), presented in Galibert et al. (2010), and called SER below, was more specifically used in the context of evaluation of NAMED ENTITY recognition systems.

Comparing a "hypothesis" to a reference, this metric counts the costs of different error types: error "T" on type (i.e., category) with cost 0.5, error "B" on boundaries

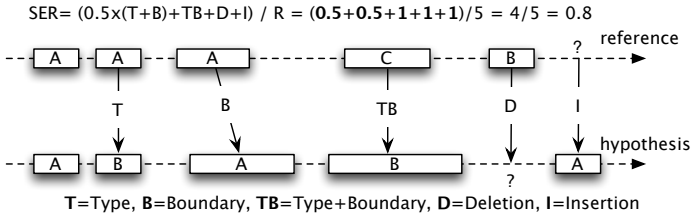


Figure 5
Slot Error Rate computation illustrated.

(i.e., position) with cost 0.5, error "TB" on both type and boundaries with cost 1, error "I" of insertion (i.e., false positive) with cost 1, and error "D" of deletion (false negative) with cost 1. The overall cost relies on an alignment of objects from reference and hypothesis, which is chosen to minimize this cost. The final value provided by SER is the average cost of the aligned pairs of units—0 meaning perfect agreement, 1 roughly meaning systematic disagreement. An example is given in Figure 5.

This attempt to extend Slot Error Rate to unitizing suffers from severe limitations. In particular, all positioning and categorizing errors have the same penalty, which may be a serious drawback for annotation tasks where some fuzziness in boundary positions is allowed, such as TOPIC SEGMENTATION, TOPIC TRANSITION, or DISCOURSE FRAMING. Moreover, it is difficult to interpret because its output is surprisingly not upper bounded by 1 (in the case where there are many false positives). Additionally, it was initially designed to compare an output to a reference, and so requires some adjustments to cope with more than 2 annotators. Last but not least, it is not chance corrected.

3.4.3 Specific Measures for Unitizing. To our knowledge, the family of α measures proposed by Krippendorff is by far the broadest attempt to provide suitable metrics for various annotation tasks, involving both categorization and unitizing.

In the survey by Artstein and Poesio (2008, page 581), some hope of finding an answer to unitizing is formulated as follows: "We suspect that the methods proposed by Krippendorff (1995) for measuring agreement on unitizing may be appropriate for the purpose of measuring agreement on discourse segmentation." Unfortunately, as far as we know, its usage in CL is rare, despite the fact that it is the first coefficient that copes both with unitizing and categorizing at the same time, while taking chance into account. The family of α measures would then be suitable for annotation tasks related, for example, to COMMUNICATIVE BEHAVIOR or DIALOG ACTS.

We will therefore pay special attention to Krippendorff's work in this article, because it constitutes a very interesting reference to compare with, both in terms of theoretical choices and of results. Let us briefly recap Krippendorff's studies on unitizing from 1995 to 2013 and introduce some of the α measures, which will be discussed in this article. The α coefficient (Krippendorff 1980, 2004, 2013), dedicated to agreement measures on categorization tasks, generalizes several other broadly used statistics and allows various categorization values (nominal, ordinal, ratio, etc.). Besides this well-known α measure, which copes with categorizing, a new coefficient called αU has been proposed since 1995 in Krippendorff (1995) and then Krippendorff (2004), which can apply to unitizing. Recently, Krippendorff (2013, pages 310, 315) proposed a new version of this coefficient, called ${}_u\alpha$, "with major simplifications and improvements over

previous proposals," and which is meant to "assess the reliability of distinctions within a continuum—how well units and gaps coincide and whether units are of the same or of a different kind." To supplement ${}_{u}\alpha$, which mainly focuses on positioning, Krippendorff has proposed ${}_{c|u}\alpha$ (Krippendorff 2013), which ignores positioning disagreement and focuses mainly on categories.

These measures will be discussed in the following sections. For now, it must be noted that ${}_{u}\alpha$ and ${}_{c|u}\alpha$ are not currently designed to cope with embedding or free overlapping between the units of the same annotator. These metrics are then unsuitable for annotation tasks such as, for instance, TOPIC TRANSITION, HIERARCHICAL TOPIC SEGMENTATION, DISCOURSE FRAMING, or ENUMERATIVE STRUCTURES.

3.5 Overview Table

To conclude the state of the art, we draw up a final overview of the coverage of the requirements by the different measures in Table 2. The γ measure, introduced in the next section, aims at satisfying all these needs.

4. The Proposed Method: Introducing γ

4.1 Our Proposal

The basic idea of this new coefficient is as follows: All local disagreements (called **disorders**) between units from different annotators are averaged to compute an overall disorder. However, these local disorders can be computed only if we know for each unit of a given annotator, which units, if any, from the other annotators it should be compared with (via what is called **unitary alignment**)—that is to say, if we can rely on a suitable alignment of the whole (called **alignment**). Because it is not possible to get a reliable preconceived alignment (as explained in Section 4.2.1), γ considers all

Table 2
Overview of available metrics. ■: satisfied constraint, □: partly satisfied constraint.

Measure	Categorizing	Segmentation	Unitizing	Unitizing with overlap	Weighted(category overlap)	Chance correction	Alignment	# Annotators ≥ 3
F-measure	■							
κ, π	■					■		
multi-π, multi-κ	■					■		■
$\kappa_{weigthed}$	■				■	■		
α	■				■	■		■
WindowDiff		■						
GHD		■					■	
modified SER	■	■	■	□			■	
${}_{u}\alpha$	□		■			■		■
${}_{c\|u}\alpha$	■		□		■	■		■
discretizing measure	■		□		■	■		■
γ	■	■	■	■	■	■	■	■

possible ones, and computes for each of them the associated overall disorder. Then, γ retains as the best alignment the one that minimizes the overall disorder, and the latter value is retained as the correct disorder. To obtain the final agreement, as with the familiar kappa and alpha coefficients, this disorder is then chance-corrected by a so-called expected disorder, which is calculated by randomly resampling existing annotations.

First of all, we introduce three main principles of γ in Section 4.2. We introduce in Section 4.3 the basic definitions. The comparison of two units (depending on their relative positions and categories) relies on the concept of **dissimilarity** (Section 4.4). A **unitary alignment** groups at most one unit of each annotator, and a set of unitary alignments covering all units of all annotators is called an **alignment** (Section 4.5). The **disorder** associated with a unitary alignment results from dissimilarities between all its pairs of units, and the disorder associated with an alignment depends on those of its unitary alignments (Section 4.6). The alignment having the minimal disorder (Section 4.7) is used to compute the **agreement** value, taking **chance correction** into account (Section 4.8).

4.2 Main Principles of γ

4.2.1 Measuring and Aligning at the Same Time: γ is Unified. For a given phenomenon identified by several annotators, it is necessary to provide an agreement measure permissive enough to cope with a double discrepancy concerning its position in the continuum, and the category attributed to the phenomenon.

Because of discrepancy in positioning, it is necessary to provide an agreement measure with an inter-annotator alignment, which shows which unit of a given annotator corresponds, if any, to which unit of another annotator. If such an alignment is provided, it becomes possible, for each phenomenon identified by annotators, to determine to what extent the annotators agree both on its categorization and its positioning. This quantification relies on a certain measure (called **dissimilarity** hereafter) between annotated units: The more the units are considered as similar, the lesser the dissimilarity.

But how can such an alignment be achieved? For instance, in Figure 6, aligning unit *A*1 of annotator A with unit *B*1 of annotator B consists in considering that their properties are similar enough to be associated: annotator A and annotator B have accounted for the same phenomenon, even if in a slightly different manner. Consequently, to operate, the alignment method should rely on a measure of distance (in location, in category assignment, or both) between units.

Therefore, agreement measure and aligning are interdependent: It is not possible to correctly measure without aligning, and it is not possible to align units without measuring their distances. In that respect, measuring and aligning cannot constitute two successive stages, but must be considered as a whole process. This interdependence

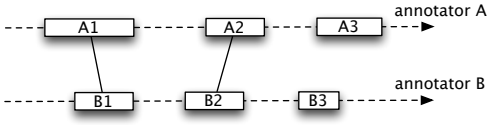


Figure 6
An example of alignment choices: Two pairs of units are aligned (*A*1 with *B*1, *A*2 with *B*2), one is not.

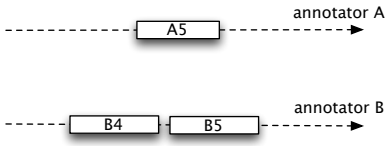


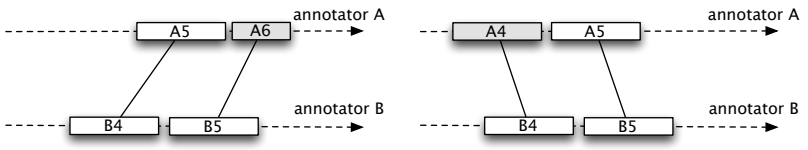
Figure 7
Inter-annotator configuration observed at local level.

reflects the unity of the objective: Establishing to what extent some elements, possibly different, may be considered as similar enough either to quantify their differences (when measuring agreement), or to associate them (when aligning).

Interestingly, Reidsma, Heylen, and Ordelman (2006, page 1119), not really satisfied by the use of the *discretizing measure* as already mentioned, "have developed an extra method of comparison in which [they] try to align the various segments." This attempt highlights the necessity to rely on an alignment. Unfortunately, the way the alignment is computed, adapted from Kuper et al. (2003), is disconnected from the measure itself, being an ad hoc procedure to which other measures are applied.

4.2.2 Aligning Globally: γ is Holistic. Let us consider two annotators A and B having respectively produced unit *A*5, and units *B*4 and *B*5, as shown in Figure 7. When considering this configuration at a local level, we may consider, based on the overlapping area for instance, that *A*5 fits *B*5 slightly better than *B*4. However, this local consideration may be misleading. Indeed, Figure 8 shows two larger configurations, where *A*5, *B*4, and *B*5 are unchanged from Figure 7. With a larger view, the choice of alignment of *A*5 may be driven by the whole configuration, possibly leading to an alignment with *B*4 in Figure 8a, and with *B*5 in Figure 8b: Alignment choices depend on the whole system and the method should consequently be **holistic**.

4.2.3 Accounting for Different Severity Rates of Errors: Positional and Categorial Permissiveness of γ. As far as positional discrepancies between annotators are concerned, it is important for a measure to rely on a progressive error count, not on a binary one: Two positions from two annotators may be more or less close to each other but still concern the same phenomenon (partial agreement), or may be too far to be considered as related to the same phenomenon (no possible alignment). For instance, for segmentation, specific measures such as GHD or WD rely on a progressive error count for positions, with an upper limit being half the average size of the segments. For unitizing, Krippendorff



(a) Context 1 (b) Context 2

Figure 8
Alignment choices depend on the whole system.

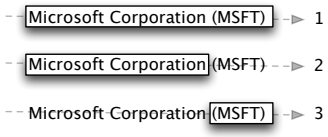


Figure 9
Alignments with no necessary intersection.

considers with ${}_u\alpha$ that units can be compared as long as they overlap. However, γ considers that in some cases, units by different annotators may correspond to the same phenomenon though they do not intersect. We base this claim on two grounds. First, if we observe the configuration given in Figure 9, annotators 2 and 3 have both annotated part of the NAMED ENTITY that has been annotated by annotator 1. Consequently, though they do not overlap, their units refer to the same phenomenon. In addition, we find a direct echo of this assumption in Reidsma (2008, pages 16–17) where, in a video corpus concerning COMMUNICATIVE BEHAVIOR, "different timing (non-overlapping) [of the same episode] was assigned by [...] two annotators." Regarding categorization, some available measures consider all disagreements between all pairs of categories as equal. Other coefficients, called *weighted* coefficients (see Artstein and Poesio 2008), as well as γ, consider on the contrary that mismatches may not all have the same weight, some pairs of categories being closer than others. This closeness is often referred to as *overlap*.

In our terminology, we call **category-overlapping** this closeness between categories, and **overlap** means **positional overlap**. For example, within annotation efforts related to WORD SENSE or DIALOG ACTS, it is clear that disagreements on labels are not all alike.

4.3 Definitions: Unit, Annotator, Annotation Set

Given a multi-annotated continuum t:

- let $\mathcal{A} = \{a_1, ..., a_n\}$ be the set of annotators
- let $n = |\mathcal{A}|$ be the number of annotators
- let $\mathcal{U}$ be the set of units from all annotators
- $\forall i \in [\![1, n]\!]$, let x_i be the number of units by annotator a_i for t
- let $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ be the average number of annotations per annotator
- for annotator $a = a_i$, $\forall j \in [\![1, x_i]\!]$, we note u_j^a unit from a of rank j

Annotation set: An annotation set s is a set of units attached to the same continuum and produced by a given set of annotators.

Corpus: A corpus c is defined with respect to a given annotation effort, and is composed of a set of continua, and of the set of annotations related to these continua.

Unit: A unit u bears a category denoted $cat(u)$, and a location given by its two boundaries, each of them corresponding to a position in the continuum, respectively denoted $start(u)$ and $end(u)$, *start* and *end* being functions from $\mathcal{U}$ to $\mathbb{N}^+$.

Equality between units is defined as follows:

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

4.4 Dissimilarity Between Two Units

We introduce here the first brick to build the notion of disorder, which works at a very local level, between two units. A dissimilarity tells to what degree two units should be considered as different, taking into account such features as their positions, their categories, or a combination of the two.

A dissimilarity is a function $d : \mathcal{U}^2 \rightarrow \mathbb{R}^+$, so that :

$$\forall (u,v) \in \mathcal{U}^2, \begin{cases} d(u,v) = d(v,u) \text{ (d is symmetric)} \\ u = v \Rightarrow d(u,v) = 0 \end{cases}$$

A dissimilarity is not necessarily a distance in the mathematical sense of the term, in particular because triangular inequality is not mandatory (for instance, in Figure 10, $d(A1,B2) > d(A1,C1) + d(C1,B2)$).

4.4.1 Empty Unit $u_\emptyset$, Empty Dissimilarity $\Delta_\emptyset$. As we will see, γ relies on an alignment of units by different annotators. In particular, this alignment indicates for unit $u_i^{a_1}$ of annotator a_1, to which unit $u_j^{a_2}$ of annotator a_2 it corresponds, in order to compute the associated dissimilarity. In some cases, though, the method will choose not to align $u_i^{a_1}$ with any unit of annotator a_2 (none corresponds sufficiently). We define the empty pseudo unit, denoted $u_\emptyset$, which corresponds to the realization of this phenomenon: ultimately, a pseudo unit $u_\emptyset$ is added to the annotations of a_2, and $u_i^{a_1}$ is aligned with it.

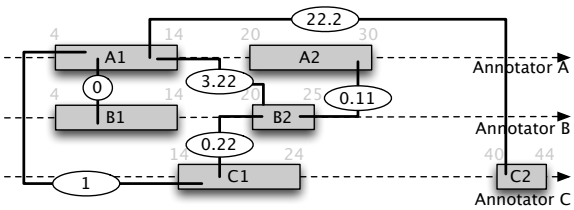


Figure 10
Real examples of $d_{pos-sporadic}$ values (divided by $\Delta_\emptyset$).

We also define the associated cost $\Delta_\emptyset$:

$$\forall u \in \mathcal{U}, d(u, u_\emptyset) = d(u_\emptyset, u) = \Delta_\emptyset$$
$$\text{and } d(u_\emptyset, u_\emptyset) = \Delta_\emptyset$$

Dissimilarities should be calibrated so that $\Delta_\emptyset$ is the value beyond which two compared units are considered critically different. Consequently, it constitutes a reference, and dissimilarities will be expressed in this article as multiples of $\Delta_\emptyset$ for better clarity. It is not a parameter of *gamma*, but a constant (which is set to 1 in our implementation).

4.4.2 Positional Dissimilarity d_{pos}. Different positional dissimilarities may be created, in order to deal with different annotation tasks. In this article, we use the dissimilarity shown in Equation 3, which is very versatile.

$$d_{pos-sporadic}(u, v) = \left(\frac{|start(u) - start(v)| + |end(u) - end(v)|}{(end(u) - start(u)) + (end(v) - start(v))} \right)^2 \cdot \Delta_\emptyset \tag{3}$$

Equation (3) sums the differences between the right and left boundaries of both units in its numerator. Its denominator sums the lengths of both units, so that this dissimilarity is not scale-dependent. Squaring the value is an option used here to accelerate dissimilarity when differences of positions increase. It is illustrated in Figure 10 with different configurations and their associated values, from 0 for the perfectly aligned pair of units $(A1, B1)$ to $22.2 \cdot \Delta_\emptyset$ for the worst pair $(A1, C2)$.

4.4.3 Categorial Dissimilarity d_{cat}. Let K be the set of categories. For a given annotation effort, $|K|$ different categories are defined. For more convenience, we first define **categorial distance** between categories $dist_{cat}$ via a square matrix of size $|K|$, with each category appearing both in row titles and column titles. Each cell gives the distance between two categories through a value in $[0, 1]$. Value 0 means perfect equality, whereas the maximum value 1 means that the categories are considered as totally different. As $dist_{cat}$ is symmetric, such a matrix is necessarily symmetric, and bears 0 in each diagonal cell. Table 3 gives an example for three categories, and shows that an association between a unit in category cat_1 with one in category cat_3 is the worst possible (distance = 1), whereas it is half as much between cat_1 with cat_2 (distance = 0.5). This makes it possible to take into account so-called **category-overlapping** (in our example, cat_1 and cat_2 are said to overlap, which means they are not completely different), as **weighted** coefficients such as κ_w or α already do. Note that in the case of so-called "nominal categories," the matrix will be full of 1 outside the diagonal, and full of 0 in the diagonal (different categories are considered as not matching at all).

This categorial distance matrix is then used to build the categorial dissimilarities, taking into account the $\Delta_\emptyset$ value. We define categorial dissimilarity between two units by:

$$d_{cat}(u, v) = f_{cat}(dist_{cat}\,(cat(u), cat(v))) \cdot \Delta_\emptyset \tag{4}$$

Table 3
Categorial matrix of $dist_{cat}$ for 3 categories.

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

Function f_{cat} can be used to adjust the way the dissimilarity grows with respect to the categorial distance values. The standard option[2] (used in this article) is to simply consider $f_{cat}(x) = x$, with which d_{cat} naturally increases gradually from zero when categories match, to $\Delta_\emptyset$ when categories are totally different ($dist_{cat}(cat(u), cat(v)) = 1 \implies d_{cat}(u, v) = \Delta_\emptyset$).

4.4.4 Combined Dissimilarity d_{combi}. Because in some annotation tasks units may differ both in position and in category, it is necessary to combine the associated dissimilarities so that all costs are cumulated. This is provided by a combined dissimilarity.

Let d_1 and d_2 be two dissimilarities. We define:

$$d^{\alpha,\beta}_{combi(d_1,d_2)}(u, v) = \alpha.d_1(u, v) + \beta.d_2(u, v) \tag{5}$$

It is easy to demonstrate that this linear combination of dissimilarities is itself a dissimilarity (if $(\alpha, \beta) \neq (0, 0)$). It enables the same weight to be assigned to positions and categories using $d^{1,1}_{combi(d_{pos},d_{cat})}$, which is currently used for γ.

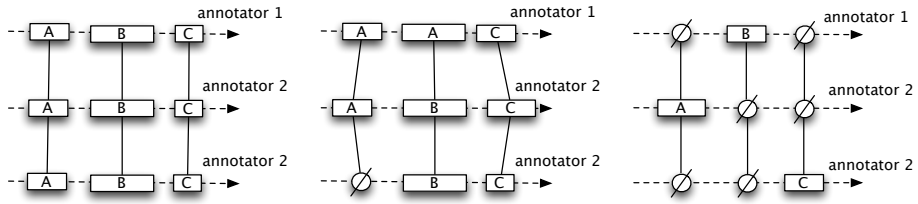
Then, we can note that it is the same cost $\Delta_\emptyset$ for a unit either not to be aligned with any other one, or to be aligned with a unit in the same configuration as $(A1, C1)$ of Figure 10 (if they have the same category), or to be aligned with a unit having an incompatible category (if they occupy the same position).

4.5 Unitary Alignment, Alignment

Unitary alignment $\breve{a}$. A unitary alignment $\breve{a}$ is an i-tuple, i belonging to $[\![1, n]\!]$ (n the number of annotators), containing at most one unit by each annotator: It represents the hypothesis that i annotators agree to some extent on a given phenomenon to be unitized. In order to make all unitary alignments homogenous, we eventually complete any unitary alignment that is an i-tuple with $n - i$ empty units $u_\emptyset$, so that all unitary alignments are ultimately n-tuples. Figure 11 illustrates unitary alignments with some $u_\emptyset$ units.

Alignment $\bar{a}$. For a given annotation set, an alignment $\bar{a}$ is defined as a set of unitary alignments such that each unit of each annotator belongs to one and only one of its

2 Another option is, for example, to use $f_{cat}(x) = -ln(1 - x)x^{30} + x$, which is a function almost similar to $f_{cat}(x) = x$ on the $[0, 0.9]$ range, and reaches ∞ near 1. Then, when the categorial distance is equal to 1, the categorial dissimilarity reaches infinity, which guarantees that the units cannot be aligned.



(a) best (b) intermediate (c) worst

Figure 11
Examples of best, intermediate, and worst possible disorders.

unitary alignments. Mathematically, it constitutes a **partition** of the set of units (if we do not take $u_\emptyset$ into account).

4.6 Alignment and Disorder

4.6.1 Disorder of a Unitary Alignment. The disorder of a unitary alignment $\breve{a}$, denoted $\breve{\delta}(\breve{a})$, is defined for a given dissimilarity d as the average of the one-to-one dissimilarities of its units:

$$\breve{\delta}(\breve{a}) = \frac{1}{C_n^2} \cdot \sum_{(u,v)\in\breve{a}^2} d(u,v) \tag{6}$$

Averaging dissimilarities rather than summing them makes the result independent of the number of annotators.

4.6.2 Disorder of an Alignment. The disorder of an alignment $\bar{a}$, denoted $\bar{\delta}(\bar{a})$, is the sum of the disorders of all its unitary alignments divided by the mean number of units per annotator:

$$\bar{\delta}(\bar{a}) = \frac{1}{\bar{x}} \cdot \sum_{i=1}^{|\bar{a}|} \breve{\delta}(\breve{a}_i) \tag{7}$$

We chose to consider the average value rather than the sum so that the disorder does not depend on the size of the continuum.

4.7 Best Alignment, Disorder of an Annotation Set

Best alignment $\hat{a}$. An alignment $\bar{a}$ of the annotation set s is considered as the best (with respect to a dissimilarity) if it minimizes its disorder among all possible alignments of s. It is denoted $\hat{a}$. The proposed method is holistic in that it is necessary to take into account the whole set of annotations in order to determine each unitary alignment.

Disorder of an annotation set $\delta(s)$. The disorder of the annotation set s, denoted $\delta(s)$, is defined as the disorder of its best alignment(s) $\bar{\delta}(\hat{a})$. Note that it may happen that several alignments produce the lowest disorder.

We have just presented the two crucial definitions of our new method, which make it "unified." Indeed, the best alignment is chosen with respect to the disorder,

therefore with respect to what computes the agreement measure; and, conversely and simultaneously, the resulting agreement value (see below) is given by the best alignment: agreement computation and alignment are fully intertwined, whereas in most agreement metrics, the alignment is fixed a priori or no alignment is used.

4.8 Expected Disorder for Chance Correction

4.8.1 The Model of Chance of γ. As we have already mentioned in the state of the art, it is necessary for an inter-annotator agreement measure to provide chance correction. We have also seen that there are several chance correction models, and that it is a controversial question. However, for γ, we follow Krippendorff, who claims that *annotators should be interchangeable*, because, as stressed by Krippendorff (2011) and Zwick (1988), Cohen's definition of expected agreement (using individual distributions) numerically rewards annotators for not agreeing on their use of values, that is to say when they have different prevalences of categories, and punishes those that do agree. Therefore, expected values of γ are computed on the basis of *the average distribution of observed annotations of the several annotators*.

More precisely, we define the expected (chance) disorder as the average disorder of a multi randomly annotated continuum where:

- The random annotations *fulfill the observed annotation distributions* for the following features:
  - the distribution of the number of units per annotator
  - the distribution of categories
  - the distribution of unit length per category
  - the distribution of gaps' length
  - the distribution of overlapping and/or covering between each pair of categories (for instance, units of categories A and B may never intersect, 7% of the units of category A may cover one unit of category C, and so on).
- The number of random annotators is the same as the number of annotators in the observed data

4.8.2 Two Possible Sources to Build Chance: Local Data versus Corpus Data. In addition, whereas other studies systematically compute the expected value on the data also used to compute the observed value (see Section 3.1), we consider that it should be computed, when possible (that is to say, when several continua have been annotated with the same set of categories and the same instructions), from the distribution observed in all continua of the annotation effort the evaluated continuum comes from: If distribution changes from one continuum to another one, it is more because of the content of each continuum than because of chance. Let us illustrate this by a simple example, where two annotators have to annotate several texts from a sentiment analysis point of view, using three available categories: positive, negative, and neutral. On average, on the whole corpus, we assume that the prevalence is $\frac{1}{3}$ for each category. The expected agreement on the whole corpus is thus 0.33. We also assume that for one particular text, there are only positive and neutral annotations, 50% of each, and no negative one. The expected agreement for this particular text is 0.5, which means that

this particular text is considered to facilitate agreement by chance, and which has the consequence that the final agreement will be more conservative than for the rest of the corpus. Why does the third category, "negative," not appear in this expected agreement computation? This conception of chance considers that when an annotator begins to annotate this particular text, which she does not already know, the third category no longer exists in her mind, and that it is the case for every other annotator, whereas they are not supposed to cooperate. It cannot be by chance that all annotators use one category in some texts, and not in another one, but because of the content, and of the interpretation, of a given text. For this reason, from our point of view, it is better to take into account the data observed on a whole annotation effort rather than on each individual continuum. The complete data tell more about the mean behavior of annotators, whereas data of a given continuum may depend more on the particularities of its content.

As a consequence, γ provides two ways to compute the expected values: one which considers only the data of the continuum being evaluated, as does every other coefficient; and a second one, which considers the data from all continua of the annotation effort the evaluated continuum comes from. When available, we recommend using the second one, for the reasons already expressed.

4.8.3 Using Sampling to Compute the Expected Value. Expected agreement (or disagreement) is the expected value of a random variable. But which random variable? For coefficients like kappa and alpha, observed agreement (or disagreement) is the mean agreement (or disagreement) on all pairs of instances, so the random variable can be as simple as a random pair of instances (however we interpret "random"). This value can be readily computed. For gamma, however, observed disagreement is determined on a whole annotation, so the random variable needs to be a whole random annotation. The expected value of such a complicated variable is much more difficult to determine analytically. Instead, gamma uses sampling, as introduced in Section 5.

4.9 Agreement Measure γ

Now that the disorder and the expected disorder have been introduced, we can define the agreement measure (of annotation set s belonging to corpus c, with $c = \{s\}$ if s is a sole annotation set) with Equation 8, which is derived from Equation 2:

$$\forall s \in c, \gamma = 1 - \frac{\delta(s)}{\delta_e(c)} \tag{8}$$

If all annotators perfectly agree (Figure 11a), $\gamma = 1$. Figure 11c corresponds to the worst case, where the annotators are worse than annotating at random, with $\gamma < 0$. Figure 11b shows an intermediate situation.

5. Implementation

In this section, we first propose an efficient solution to compute the disorder of an annotated continuum, which relies on linear programming. Second, we propose two ways to generate random annotated continua (with respect to the observed distributions) to compute the expected disorder, one relying on a single continuum, the other one relying on a corpus (i.e., several continua). Third, we determine the number of random data sets

that we must generate (and compute the disorder of) to obtain an accurate value of the expected disorder.

5.1 Computing the Disorder

In order to simplify the discussion and the demonstrations, we consider in this section that n annotators all made the same number of annotations p.

The proposed method has now been fully described on a theoretical level, but, being holistic, its software implementation leads to a major problem of complexity. One can demonstrate that there are theoretically $(p!)^{n-1}$ possible alignments. However, we will (1) show how to reduce the initial complexity, and (2) provide an efficient linear programming solution.

5.1.1 Reducing the Initial Complexity. The initial number of possible unitary alignments (which are used to build a possible alignment) is p^n. Fortunately, theorem provided as Equation (9) states that any unitary alignment with a cost beyond the value $n \cdot \Delta_\emptyset$ cannot belong to the best alignment, and so can be discarded. Indeed, any unitary alignment with a cost above $\Delta_\emptyset$ can be replaced by creating a separate unitary alignment for each unit (of cost $\Delta_\emptyset$ per unitary alignment, so of total cost $n \cdot \Delta_\emptyset$).

Demonstration. Consider the best alignment $\hat{a}$, of cardinality m. Let $\breve{a}$ be any of its unitary alignments. For convenience, we attribute to it the index 1 ($\breve{a} = \breve{a}_1$), while the others are indexed from 2 to m. This unitary alignment $\breve{a}$ contains n units (either real or $u_\emptyset$). For each of these units u_i ($1 \leq i \leq n$), we create the unitary alignment $\breve{a}_{m+i} = (u_i, u_\emptyset, ..., u_\emptyset)$ of cardinality n. It is possible to create an alignment $\bar{a}$ made up of the set of unitary alignments of $\hat{a} \setminus \{\breve{a}\}$, to which we add the unitary alignments $\breve{a}_{m+1}$ to $\breve{a}_{m+n}$ that we have just created.[3] It is of cardinality $m + n - 1$. Because $\hat{a}$ minimizes the disorder, we obtain:

$$\bar{\delta}(\hat{a}) \leq \bar{\delta}(\bar{a}) \Rightarrow \frac{1}{\bar{x}} \sum_{i=1}^{m} \breve{\delta}(\breve{a}_i) \leq \frac{1}{\bar{x}} \sum_{i=2}^{m+n} \breve{\delta}(\breve{a}_i)$$

$$\Rightarrow \sum_{i=1}^{m} \breve{\delta}(\breve{a}_i) \leq \sum_{i=2}^{m+n} \breve{\delta}(\breve{a}_i)$$

$$\Rightarrow \breve{\delta}(\breve{a}_1) \leq \sum_{i=m+1}^{m+n} \breve{\delta}(\breve{a}_i)$$

since $\forall i > m, \breve{\delta}(\breve{a}_i) = \frac{1}{C_n^2}(C_n^2 \Delta_\emptyset) = \Delta_\emptyset$, and since we have denoted $\breve{a} = \breve{a}_1$,

$$\Rightarrow \breve{\delta}(\breve{a}) \leq n \cdot \Delta_\emptyset \tag{9}$$

Experiments have shown that this theorem allows us to discard about 90% of the unitary alignments.

3 $\bar{a}$ is indeed an alignment, because each of its units appears in one and only one unitary alignment.

5.1.2 Finding the Best Alignment: A Linear Programming Solution. Finding the best alignment consists of minimizing the global disorder. Such a problem may be described as a linear programming problem, so that the solution can be computed by a linear programming solver. For convenience, we introduce two new definitions:

- Let $\mathcal{UA}$ be the set of all unitary alignments.
- Let $\mathcal{UA}_u$ be the set of the unitary alignments which contain unit u.

The description of the problem in linear programming terms is threefold.

First, for a given alignment $\bar{a}$, for each possible unitary alignment $\breve{a}_i$, we define the **Boolean variable** $X_{\breve{a}_i}$, which indicates if this unitary alignment belongs or not to the alignment:

$$\forall \breve{a}_i \in \mathcal{UA},\ X^{\bar{a}}_{\breve{a}_i} = \begin{cases} 0 \text{ iff } \breve{a}_i \notin \bar{a} \\ 1 \text{ iff } \breve{a}_i \in \bar{a} \end{cases}$$

Second, we have to express the fact that, by definition, each unit u (of each annotator) should *belong to one and only one unitary alignment* of the alignment $\bar{a}$, that is to say that among all unitary alignments containing u, exactly one $X_{\breve{a}_i}$ equals 1, and all the others equal 0:

$$\forall u \in \mathcal{U},\ \sum_{\breve{a}_i \in \mathcal{UA}_u} X^{\bar{a}}_{\breve{a}_i} = 1$$

Third, the *goal* is to minimize the global disorder $\bar{\delta}(\bar{a})$ associated with $\bar{a}$, among all possible alignments $\bar{a}$:

$$\textit{Minimize } \bar{\delta}(\bar{a}) = \sum_{\breve{a}_i \in \mathcal{UA}} \breve{\delta}(\breve{a}_i) \cdot X^{\bar{a}}_{\breve{a}_i}$$

The `LPSolve` solver[4] finds the best solution in less than one second with $n = 3$ annotators and $p = 100$ annotations per annotator on a current laptop (once the initial complexity has been reduced thanks to the previous theorem), which is fast enough to be practical.

5.2 Implementation of Expected Disorder

The next two subsections detail two strategies to generate randomly annotated continua with respect to the definition of the expected disorder of γ, and the third subsection explains how to choose the number of expected disorder samples to generate so that their average is an accurate enough value of the theoretical expected value. The two strategies correspond to the need expressed in Section 4.8.1 to compute the expected value on the largest set of available data, either a single continuum, or, when available, several continua from the same corpus.

4 `http://lpsolve.sourceforge.net`.

5.2.1 A Strategy to Compute the Expected Disorder Using a Single Continuum. When the annotation effort is limited to a single continuum, we can only rely on the annotated continuum itself to compute the expected value. To create random annotations that fulfill the observed distributions, the implemented strategy is as follows: We take the real annotated continuum of an annotator (such as the example shown on the left in Figure 12), choose at random a position on this continuum, split the continuum at this position, and permute the two parts of the split continuum. Three examples of split and permutation are shown in the right part of the figure, for split positions of, respectively, 15, 24, and 38, all coming from the same real continuum, with units that are no longer aligned (except by chance). However, we have to address the fact that some units may intersect with themselves, generating some part of agreement beyond chance. For instance, in Figure 12, unit 3 intersects with itself between #15 and #24, because the length of the unit, 12, is higher than the difference of shifts $24 - 15 = 9$. To limit this phenomenon, we do not allow the distance between two shifts to be less than the average length of units.

5.2.2 A Strategy to Compute the Expected Disorder Using Several Continua (from the Same Corpus). This strategy consists of mixing annotations coming from different continua, so that their units may align only by chance. To create a random annotation of n annotators, we randomly choose n different continua of the corpus, and pick the annotations of one annotator (randomly chosen) of each of these continua. When different texts are of different lengths, each of them is adjusted to the longest one by duplicating as many times as necessary (like a mosaic).

This is shown in Figure 13 for $n = 3$ annotators. We assume the corpus contains eight continua, each annotated by three annotators. To generate a random set of three annotations, we have randomly selected a combination of three values between 1 and 8, here $(2, 4, 7)$, to select three different continua among the eight available ones of the corpus. Then, for each of these selected continua, we choose one annotator, here annotator 2 for continuum 2, annotator 3 for continuum 4, and annotator 1 for continuum 7. We combine the associated annotations as shown in the right part of the figure, and obtain a set of random annotations that fulfill (on average) the observed distributions. The (very limited) extent of the resulting agreement we can see in this example (only two units have matching categories, but with discrepancies in position) is only by chance because the compared annotations come from different continua.

In addition, it is possible to create a great number of random sets of annotations with this strategy: With n annotators and m continua ($m \geq n$), it is possible to generate up to $C_m^n \cdot n^n$ different combinations. For instance, in our example, which assumes $n = 3$ and $m = 8$, there are $56 \times 3^3 = 1512$ combinations to create random annotations.

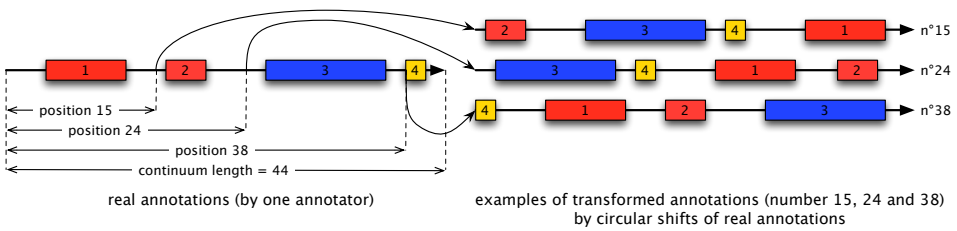


Figure 12
Principle of circular shift for creating random annotations.

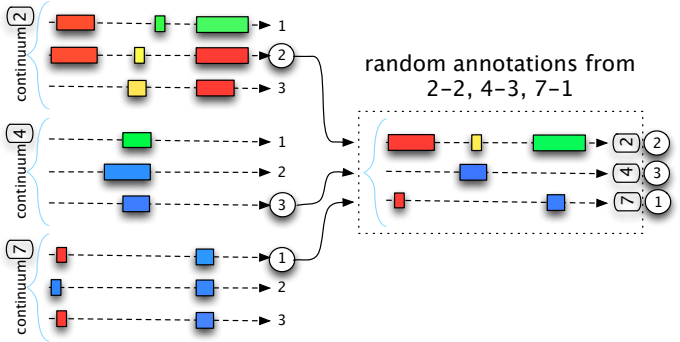


Figure 13
A corpus level strategy for creating random annotations (three annotators).

5.3 Computing an Average Value: A Sampling Question

Because the expected disorder is by definition randomly obtained **on average**, and because there is virtually an infinite number of possible random annotations (with a discrete and finite continuum, it is not really infinite, but still too big to be computed), we can only compute *a reduced but sufficient number of experiments* and obtain an approximate value of the expected disorder. This is a sampling problem as described, for example, in Israel (1992). What statistics provide is a way to determine the minimal number n_0 of experiments to do (and to average) so that we get an approximate result of *a given precision* with *a given confidence level*. It consists in: First, taking a small sample to estimate the mean and standard deviation; then, using these estimates to determine the sample size n_0 that is needed.

We follow the strategy provided in Olivero (2001) to compute a disorder value that differs less than $e = 2\%$ from the real value with a $(1 - \alpha) = 95\%$ confidence (the software distribution we provide is set by default with these values).

First, we consider a sample of chance disorder values of size $n = 30$. Let μ be the sample mean, and σ' be its standard deviation. μ is directly an unbiased estimator of the population mean, and $\sigma = \sqrt{(\frac{n}{n-1})} \cdot \sigma'$ is an unbiased estimator of the real standard deviation.

Let $C_v = \frac{\sigma}{\mu}$ be the coefficient of variation (i.e., the relative standard deviation).

Let $U_{1-\frac{\alpha}{2}}$ be the abscissa of the normal curve that cuts off an area α at the tails. This value is provided in statistical tables. We get n_0 by the following equation:

$$n_0 = \left(\frac{C_v \cdot U_{1-\frac{\alpha}{2}}}{e} \right)^2$$

Let us consider a real example. We generate a sample of random disorders of size $n = 30$. We compute its mean $\mu = 3.49$, its standard deviation $\sigma' = 0.1379$, hence $\sigma = 0.1403$, and $C_v = 0.040188$. We get $U_{1-\frac{0.05}{2}} = 1.96$ from the corresponding available table, hence we obtain $n_0 = 15.5$. This means that a sample of 16 disorder values gives 2% of precision with 95% confidence. The mean we have already computed with 30 values fulfills this condition, and is a good approximation of the real expected disorder. If we wish to obtain a high precision of 1%, we need $n_0 = 62$. It is beyond the

initial size of our sample (which is 30), and we will have to generate an additional set of 32 values in order to obtain the required number.

6. Comparing and Benchmarking γ

As γ is an entirely new agreement measure method, it is necessary to analyze how it compares with some well known and much studied methods. First, we carry out a thorough comparison between γ and the two dedicated alphas, ${}_{u}\alpha$ and ${}_{c|u}\alpha$, which are the most specific measures in the domain. Second, we benchmark γ by comparing it with other main measures, thanks to a special tool that is briefly introduced.

6.1 Krippendorff's Alphas: Introducing ${}_{u}\alpha$ and ${}_{c|u}\alpha$ and Comparing Them to γ

As already mentioned, Krippendorff's ${}_{u}\alpha$ and ${}_{c|u}\alpha$ are clearly the most suitable coefficients for combined unitizing and categorizing. To better understand the pros and cons as well as the behavior of these measures compared with γ, we first explain how they are designed in Section 6.1.1, and then make thorough comparisons with γ from Section 6.1.2 to 6.1.6 including: (1) how they react to slight categorial disagreements, (2) interlacement of positional and categorial disagreements, (3) the impact of the size of the units on positional disagreement, (4) split disagreements, and (5) the impact of scale (e.g., if the size of all units is multiplied by 2). We finish by showing a paradox of ${}_{u}\alpha$ in Section 6.1.7.

6.1.1 Introducing ${}_{u}\alpha$ and ${}_{c|u}\alpha$. To introduce how these two coefficients work, let us consider the example taken from Krippendorff (2013), shown in Figure 14. The length of the continuum is 76, there are two annotators, and there are four possible categories, numbered 1 to 4.

The ${}_{u}\alpha$ coefficient basically relies on the comparison of all pairs of sections among annotators, a section being either a categorized unit or a gap. To get the observed disagreement value ${}_{u}D_o$, squared lengths of the unmatching intersections are summed, and this sum is then divided by the product of the length of the continuum and $m(m-1)$, m being the number of annotators. In the example, mismatches occur around the second and third units of the two annotators. From left to right, there are the following intersections: cat 1 with gap (l = 10), cat 1 with cat 3 (l = 5), gap with cat 3 (l = 8), cat 2 with cat 1 (l = 5), and cat 2 with gap (l = 5). This leads twice (by symmetry) to the sum $10^2 + 5^2 + 8^2 + 5^2 + 5^2$, and so the observed disagreement ${}_{u}D_o = \frac{2(10^2+5^2+8^2+5^2+5^2)}{76 \cdot 2(2-1)} = 3.145$. The expected value ${}_{u}D_e$ is obtained by considering all the possible positional combinations of each pair, and not only the observed ones. This means that for a given pair, one of the two units is virtually slid in front of the other in all

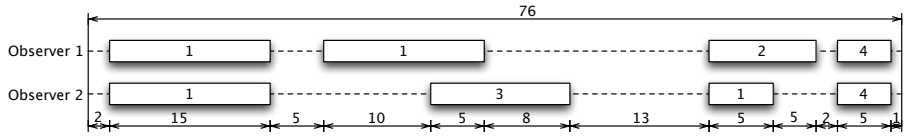


Figure 14
Example of a continuum with two observers and four categories.

possible ways, and the corresponding values are averaged. In this example, ${}_{u}D_{e} = 5.286$. Therefore, ${}_{u}\alpha = 1 - \frac{3.145}{5.286} = 0.405$.

Coefficient ${}_{c|u}\alpha$ relies on a coincidence matrix between categories, filled with the sums of the lengths of all intersections of units for each given pair of categories. For instance, in the example, the observed coincidence between category 1 and category 3 is 5, and so on. A metric matrix is chosen for these categories, for instance, an interval metric (for numerical categories), which says that the distance between category i and category j is $(i-j)^2$. Hence, the cost for a unitary intersection between categories 1 and 2 is $(1-2)^2 = 1$, but is $2^2 = 4$ between categories 1 and 3, and so on. Then, the observed disagreement is computed according to these two matrices. To finish, an expected matrix is filled (in a way which cannot be detailed here due to space constraints), and the expected value is computed the same way. In the example, ${}_{c|u}\alpha = 1 - \frac{0.833}{3.145} = 0.744$.

Hence, Krippendorff's alphas provide two clues to analyze the agreement between annotators. In the example, ${}_{u}\alpha = 0.405$ indicates that the unitizing is not so good, but also that the categorizing is much better, with ${}_{c|u}\alpha = 0.744$ (even though of course, these two values are not independent, since unitizing and categorizing coexist here by nature).

Now that these coefficients have been introduced in detail, let us analyze to what extent they differ from γ.

6.1.2 Slight Categorial Disagreements: Alphas Versus $\boldsymbol{\gamma}$. When annotators have slight categorial disagreements (with overlapping categories), ${}_{c|u}\alpha$ is slightly lowered. However, ${}_{u}\alpha$ does not take categorial overlapping into account, but has a binary response to such disagreements, and is lowered as much as if they were severe categorial disagreements. A consequence of this approach is illustrated in Figure 15, where two annotators perfectly agree both on positions and categories in the experiment on the left, and still perfectly agree on position but slightly diverge concerning categories in the experiment on the right (1/2, 6/7, and 8/9 are assumed to be close categories). However, ${}_{u}\alpha$ drops from 1 in the left experiment to –0.34 (a negative value means worse than random) in the right experiment, despite, in the latter, the positions being all correct, and the categories being quite good, since ${}_{c|u}\alpha = 0.85$. On such data, γ considers that there is no positional disagreement, and ${}_{c|u}\alpha$ and γ both consider that there are slight categorial disagreements.

6.1.3 Positional Disagreements Impacting Categorial Agreement: ${}_{c|u}\alpha$. Two different conceptions of how to account for categorial disagreement have, respectively, led to ${}_{c|u}\alpha$ and γ: ${}_{c|u}\alpha$ relies on **intersections** between the units of different annotators, which is basically equivalent to an observation at the atom level, whereas γ relies on **alignments between units** (any unit being finally attached and compared to, at most, only one other) based both on positional and categorial observation. Hence, in a configuration such as the one given in Figure 16, where two annotators annotated three units with the same categories

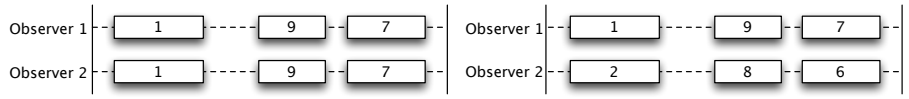


Figure 15
Consequences of no categorial disagreement (left) compared with slight categorial disagreements (right).

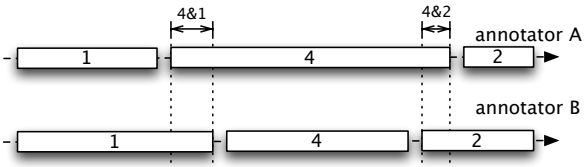
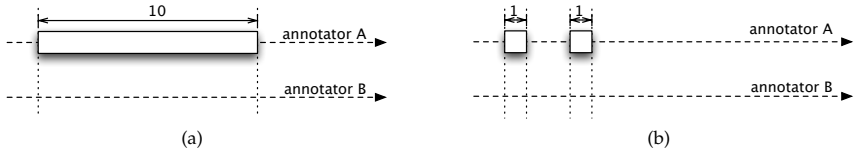


Figure 16
Positional discrepancies considered as category mistakes or not by ${}_{c|u}\alpha$ and γ.

1, 4, and 2, but not exactly at the same locations; ${}_{c|u}\alpha$ considers a certain account of categorial disagreements, whereas γ does not. According to the principles of ${}_{c|u}\alpha$, any part of the continuum (even at the atom level) with an intersection between different categories means some confusion between them, whereas γ considers here that the annotators fully agree on categories (they both observed a "1" then a "4" then a "2" with no confusion), and disagree only on where phenomena exactly start and finish. The crucial difference between the two methods is probably whether we consider units to be non-atomizable (and therefore consider alignments, as γ does), or atomizable (in which case two different parts of a given unit may be simultaneously and respectively compared to two different units from another annotator).

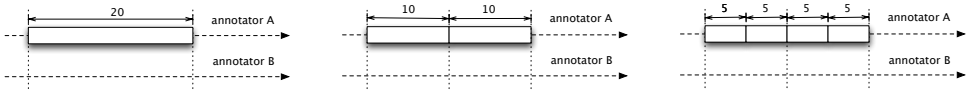
6.1.4 Disagreements on Long versus Short Spans of Texts. Here again, the way disagreements are accounted for may differ markedly between ${}_{u}\alpha$ and γ: when a unit does not match with any other, ${}_{u}\alpha$ takes into account the length of the corresponding span of text to assess a disagreement. As shown in Figure 17, an orphan unit of size 10 will cost 100 times as much as an orphan unit of size 1, whereas for γ, they will have the same cost. In the whole example in Figure 17, to compute the observed disagreements, ${}_{u}\alpha$ says the first case is 50 times worse than the second, whereas γ says on the contrary that the second case is twice as bad as the first. Here, γ fulfills the need (already mentioned) expressed by Reidsma, Heylen, and Ordelman (2006, page 3) to consider that "short segments are as important as long segments." This phenomenon is the same for categories between ${}_{c|u}\alpha$ and γ, the size of the units having consequences only for ${}_{c|u}\alpha$.

6.1.5 Split Disagreements. Sometimes, an annotator may divide a given span of text into several contiguous units of the same type, or may annotate the same span with one whole unit. In these cases, ${}_{c|u}\alpha$ computes its observed disagreement the same in both



${}_{u}\alpha_{observed} = 10^2 = 100$, $\gamma_{observed} = \Delta_{\emptyset}$ ${}_{u}\alpha_{observed} = 1^2 + 1^2 = 2$, $\gamma_{observed} = 2\Delta_{\emptyset}$

Figure 17
Observed disagreement on short versus long spans of texts.



(a) (b) (c)

$_u\alpha_{observed} = 400$ $_u\alpha_{observed} = 200$ $_u\alpha_{observed} = 100$
$\gamma_{observed} = \Delta_\emptyset$ $\gamma_{observed} = 2\Delta_\emptyset$ $\gamma_{observed} = 4\Delta_\emptyset$

Figure 18
Behavior of $_u\alpha$: splitting versus silence.

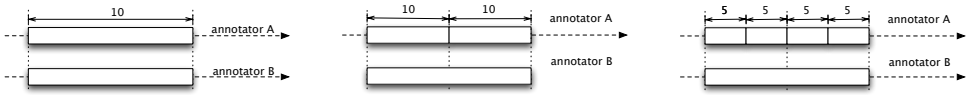
configurations, and $_u\alpha$ assigns decreasing disagreement when splitting increases, as shown in Figure 18, whereas γ assigns increasing disagreements.

Moreover, in Figure 19, the observed $_u\alpha$ is not responsive to splits at all, whereas γ is still responsive.

6.1.6 Scale Effects. The way $_u\alpha$ computes dissimilarities is directly proportional to squared lengths, as shown in Figure 20. On the other hand, γ may use any positional dissimilarity, and usually uses ones that are not scale-dependent for CL applications, such as $d_{pos-sporadic}$ (Equation (3)). For instance, if a text is annotated with two categories, one at word level, the other one at paragraph level, we may prefer to account for relative disagreements so that a missing word will be more heavily penalized in the first case than in the second. In Figure 20, the observed disagreement of $_u\alpha$ is $3^2 = 9$ times greater for B units than for A units, but would be the same for γ with $d_{pos-sporadic}$ since:

$$\left(\frac{0+3}{\left(\frac{7+10}{2}\right)}\right)^2 = \left(\frac{0+9}{\left(\frac{21+30}{2}\right)}\right)^2.$$

6.1.7 A Paradox: When Agreement on Positioning Reinforces Disagreement on Categorizing. In Figure 21a, the annotators disagree on categorization, and have a moderate agreement on unitizing. This configuration leads to $_u\alpha = 0.107$. In Figure 21b, the configuration is quite similar, but now annotators fully agree on unitizing: Each of them puts units in the same positions. Paradoxically, $_u\alpha$ drops to -0.287, which is less than in the first configuration. In brief, the reason for this behavior is that in the first case, the computed disagreement regarding a given pair of units is virtually distributed into shorter parts of the whole (an intersection of length 80 between them, and an intersection of length



(a) (b) (c)

$_u\alpha_{observed} = 0$ $_u\alpha_{observed} = 0$ $_u\alpha_{observed} = 0$
$\gamma_{observed} = 0$ $\gamma_{observed} \cong 1.5\Delta_\emptyset$ $\gamma_{observed} \cong 3.75\Delta_\emptyset$

Figure 19
Behavior of $_u\alpha$: splitting versus no splitting.

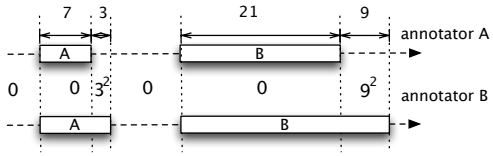
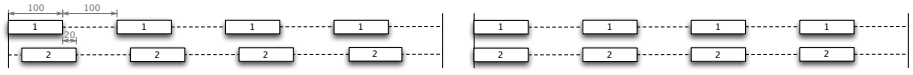


Figure 20
Scale effects on ${}_u\alpha$ versus γ.



(a) Moderate positioning agreement (b) Full positioning agreement

Figure 21
Categorizing disagreement combined with moderate versus full positioning agreement.

20 with a gap for each of them, which leads to $80^2 + 2 \times 20^2 = 7,200$) whereas the disagreement is maximum in the second case (an intersection of length 100 with a unit of another category, which leads to $100^2 = 10,000$). Contrarily, with similar data, γ provides a better agreement in the second case than in the first one. With its design, it considers that there is the same categorial agreement in both cases, but better positional agreement in the second case, which seems to better correspond to the CL tasks we have considered.

6.1.8 Overlapping Units (Embedding or Free Overlap). Both alpha coefficients are currently designed to cope only with non-overlapping units (the term overlapping also stands here for embedding), which is a limitation for several fields in CL. It is debatable whether they could be generalized to handle overlapping units. It seems that it would involve a major change in the strategy, which currently necessitates comparing the intersections of all pairs of units. In the example shown in Figure 22, even though annotators fully agree on their two units, the alphas will inherently compare A1 with B2 and A2 with B1 (in addition to the normal comparisons between A1 with B1 and A2 with B2), and will count the resulting intersections as disagreements. It is necessary here to choose once and for all what unit to compare to what other, rather than to perform all the comparisons. But making such a choice precisely consists in making an alignment, which is a fundamental feature of γ. Consequently, it seems that the alphas would need a structural modification to cope with overlapping.

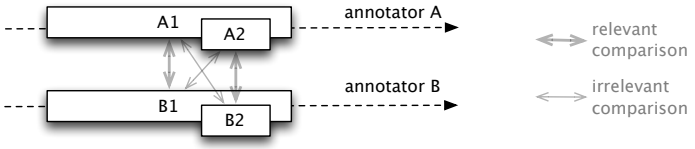


Figure 22
Full agreement with overlapping units (embedding in this example).

6.2 Discretizing Measure

As explained by Reidsma, Heylen, and Ordelman (2006), because of the lack of specialized coefficients coping with unitizing, a fairly standard practice is to use categorization coefficients on a discretized (i.e., atomized) version of the continuum: For instance, each character (or each word, or each paragraph) of a text is considered as an item; and a standard categorization coefficient such as κ is used to compute agreement. Such a measure is called κ_d (for discretized κ) hereafter. Several weaknesses of this approach have been already mentioned in the state-of-the-art section. It is interesting to compare such a measure to the specialized one ${}_{c|u}\alpha$: even if they both bear the *aggregatable hypothesis*, they have, however, significant differences (as confirmed by the experiments presented in the next section). The main one is that ${}_{c|u}\alpha$ does not use an artificial atomization of the continuum, and only compares units with units. In doing so, it is not prone to agreement on blanks, in contrast to κ_d. Another difference is that, for the same reason, ${}_{c|u}\alpha$ is not inherently limited to non-overlapping units: Even if it is not currently designed to cope with them, as we have already seen, it is possible to submit overlapping units to this measure (some results are shown in the next section).

6.3 Benchmarking Using the Corpus Shuffling Tool

In this section on benchmarking, we use the Corpus Shuffling Tool (CST) introduced by Mathet et al. (2012) to compare γ concretely and accurately to the other measures.

We first introduce the possible error types that it will provide: category (category mistakes may occur), position (the boundaries may be shifted), false positives (the annotators add units to the reference units), false negatives (the annotators miss some of the reference units), and splits (the annotators put two or more contiguous units instead of a reference unit, which occupy the same span of text).

This tool is used to simulate varying degrees of disagreement among different error types, and the metrics are compared with each other according to how they react to these disagreements. For a given error type, for each magnitude between 0 and 1 (with a step of 0.05), the tool creates 40 artificial, multi-annotator shuffled annotation sets, and computes the different measures for them. Hence, we obtain a full graph showing the behavior of each measure for this error type, with the magnitude on the x-axis, and the average agreement (over the 40 annotation sets) on the y-axis. This provides a sort of "X-ray" of the capabilities of the measures with respect to this error type, which should be evaluated against the following desiderata:

- A measure should provide a full response to the whole range of magnitudes, which means in particular that the curve should ideally start from 1 (at $m = 0$) and reach 0 (at $m = 1$), but never go below 0 (indeed, negative agreement values require a part of systematic disagreement, which is not simulated by the current version of the CST).
- The response should be strictly decreasing: A flat part would mean the measure does not differentiate between different magnitudes, and, even worse, an increasing part would mean that the measure is counter-effective at some magnitudes, where a worse error is penalized less severely.

We emphasize the fact that the whole graph is important, up to magnitude 1. Indeed, in most real annotated corpora, even when the overall agreement is high, errors

corresponding to all magnitudes may occur. For instance, an agreement of 0.8 does not necessarily correspond to the fact that all annotations are affected by slight errors (which correspond to magnitudes close to 0), but may for instance correspond to the fact that a few units are affected by severe errors (which may correspond to magnitudes close or equal to 1).

It is important to note that this tool was designed by the authors of γ, *for tasks where units cannot be considered as atomizable*. In particular, it was conceived so that disagreements concerning small units are as important as those concerning large ones. However, it is provided as open-source (see Conclusions section) so that anyone can test and modify it, and propose new experiments to test γ and other measures in the future.

6.3.1 Introducing the CST. The main principle of this tool is as follows. A reference corpus is built, with respect to a statistical model, which defines the number of categories, their prevalence, the minimum and maximum length for each category, and so forth. Then, this reference is used by the shuffling tool to generate a multi-annotator corpus, simulating the fact that each annotator makes mistakes of a certain type, and of a certain magnitude. It is important to remark that the generated corpus does not include the reference it is built from.

The magnitude m is the strength of the shuffling, that is to say the severity of mistakes annotators make compared to the reference. It can be set from 0, which means no damage is applied (and the annotators are perfect) to the extreme value 1, which means annotators are assumed to behave in the worst possible way (but still being independent of each other)—namely, at random.

Figure 23 illustrates the way such a corpus is built: From the reference containing some categorized units, three new sets of annotations are built, simulating three annotators who are assumed to have the same annotating skill level, which is set in this example at magnitude 0.1. The applied error type is position only, that is to say that each annotator makes mistakes only when positioning boundaries, but does not make any other mistake (the units are reproduced in the same order, with the correct category, and in the same number). At this low magnitude, the positions are still close to those of the reference, but often vary a little. Hence, we obtain here a slightly shuffled multi-annotator corpus. Let us sum up the way error types are currently designed in the CST.

Position. At magnitude m, for a given unit, we define a value $shift_{max}$ that is proportional to m and to the length of the unit, and each boundary of the unit is shifted by a

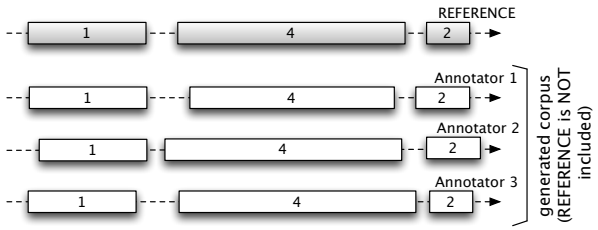


Figure 23
The CST generating three annotations with position errors at magnitude $m = 0.1$.

value randomly chosen between $-shift_{max}$ and $shift_{max}$ (note: at magnitude 0, because $shift_{max} = 0$, units are not shifted).

Category. This shuffling cannot be described in a few words (see Mathet et al. [2012] for details). It uses special matrices to simulate, using conditional probabilities, progressive confusion between categories, and can be configured to take into account overlapping of categories. The higher the magnitude, the more frequent and severe the confusion.

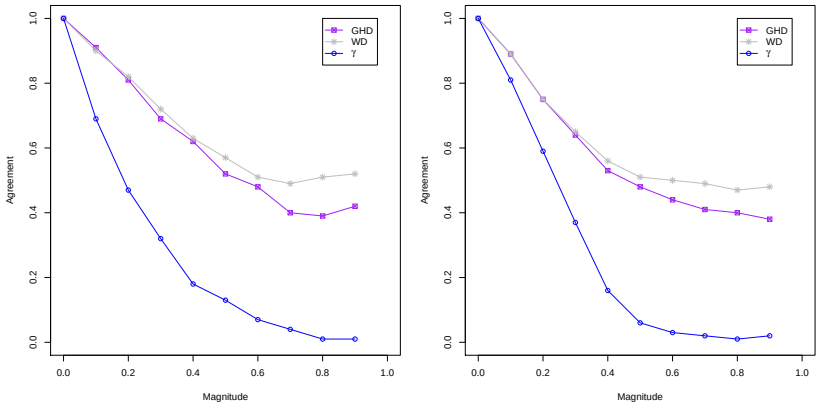
False negatives. At magnitude m, each unit has the probability m to be forgotten. For instance, at magnitude $m = 0.5$, each annotator misses (on average) half of the units from the reference (but not necessarily the same units as the other annotators).

False positives. At magnitude m, each annotator adds a certain number of units (proportional to m) to the ones of the reference.

Splits. At magnitude m, each annotator splits a certain number of units (proportional to m). A split unit may be re-split, and so on.

6.3.2 Pure Segmentation: γ, WD, GHD. Even if γ was created to cope with error types that are poorly or not at all dealt with by other methods, and, moreover, to cope simultaneously with all of them (unitizing of categorized and overlapping categories), it is illuminating to observe how it behaves in more specific error types, to which specialized and well known methods are dedicated. We start with pure segmentation. Figure 24 shows the behavior of WD, GHD, and γ for two error types.

For *false negatives*, WD and GHD are quite close, with an almost linear response until magnitude 0.6. Their drawback is that their responses are limited by an asymptote, because of the absence of chance correction, while γ shows a full range of agreements; for *shifts*, WD and GHD show an asymptote at about *agreement* = 0.4, while γ shows values from 1 to 0. This experiment confirms the advantage of using γ instead of these distances for inter-annotator agreement measure.



(a) False negatives (b) Shift

Figure 24
False negatives and shifts in pure segmentation.

6.3.3 Pure Categorization. In this experiment, the CST is set to three annotators, four categories with given prevalences. The units are all of the same size, positioned in fixed, predefined positions, so that the focus is on categorizing only. It should be noted that, with such a configuration, α *and* κ *behave exactly in the same way as* ${}_{c|u}\alpha$. It is particularly striking in Figure 25 that γ behaves in almost the same way as ${}_{c|u}\alpha$. In fact, the observed values of these measures are exactly the same, the only difference coming from a slight difference in the expected values, due to sampling. Other tests carried out with the pure categorizing coefficient κ yielded the same results on this particular error type, which means that γ performs as well as recognized measures as far as categorizing is concerned, with two or more annotators. The ${}_{u}\alpha$ curve goes below zero at magnitude 0.5 (probably for the reasons seen in Section 6.1.7). Moreover, its behavior depends on the size of the gaps: Indeed, with other settings of the shuffling, the curve may, on the contrary, be stuck over zero. κ_d fails to reach 0 because of the virtual agreement on gaps (but it would if there were no gaps). Lastly, SER (averaging the results of each pair of annotators) is bounded below by 0.6, which results from not taking chance into account.

6.3.4 Almost General Case: Unitizing + Categorizing. This section concerns the more general uses of γ, combining both unitizing and categorizing. However, in order to be compliant with ${}_{u}\alpha$, ${}_{c|u}\alpha$, and κ_d, we limit the configurations here so that the units do not overlap at all. In particular, the reference was built with no overlapping units, and we have used a modified version of the shifting shuffling procedure so that the non-overlapping constraint is fully satisfied, even at high magnitudes.

Positional errors (Figure 26a). An important point is that this shuffling error type, which is based only on moving positions, has a paradoxical consequence on category agreement, since units of different categories align when sufficient shifting is applied. Consequently, ${}_{c|u}\alpha$ is not blocked at 1, even though it is designed to focus on categories. Additionally, it starts to decrease from the very first shifts, as soon as units from different annotators start overlapping. This is a concrete consequence of what has been formally studied in Section 6.1.3. γ has a most progressive response, reaches 0.1 at magnitude 1,

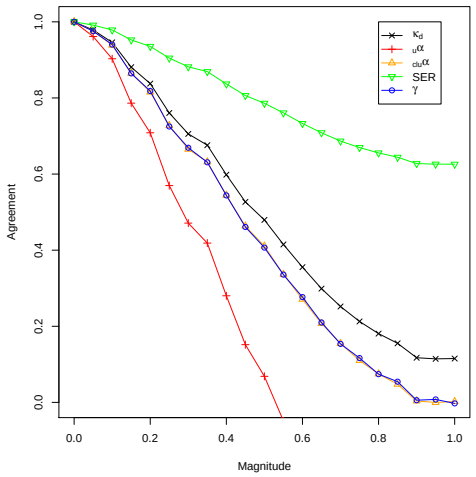
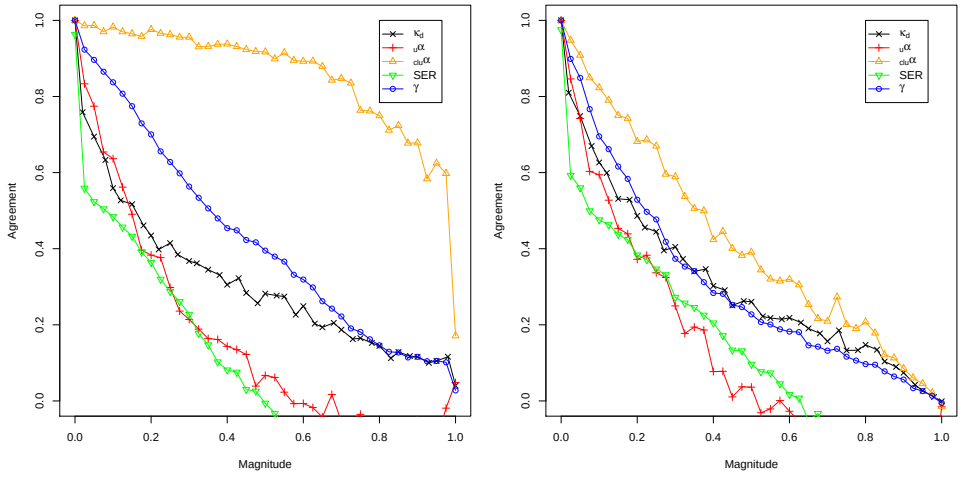


Figure 25
Agreement graphs for category errors, three annotators.



(a) Positional errors (b) Positional+Categorial errors

Figure 26
Agreement graphs for positional errors and positional + categorial errors.

and is the only measure to be strictly decreasing. *SER* immediately drops to agreement 0.5 at magnitude 0.05. As it relies on a binary positional distance, it fails to distinguish between small and large errors. This is a serious drawback of such a measure for most CL tasks. Then it goes below zero and is not strictly decreasing. ${}_u\alpha$ is mostly strictly decreasing, but has some increasing parts, and, even more problematic, negative values from 0.6 to 0.9, probably because of the reason explained in Section 6.1.7. κ_d is too responsive at the very first magnitudes, and is not strictly decreasing, probably because it "does not compensate for differences in length of segments" (Reidsma, Heylen, and Ordelman 2006, page 3).

Positional and categorial errors (Figure 26b). γ is strictly decreasing and reaches 0. The alphas are not strictly decreasing, and once again ${}_u\alpha$ drops below 0 from magnitude 0.6 onwards. κ_d is not strictly decreasing (again, probably because it "does not compensate for differences in length of segments"), but its general shape is not that far from γ.

Split errors (Figure 27). The split error type would need to create an infinite number of splits to mean pure chaos at magnitude 1. As this is computationally not possible, we restricted the number of splits to five times the number of units of the reference. We should therefore not expect measures to reach 0. In this context, γ shows a good range of responses, from 1 to 0.2, in an almost linear curve. *SER* is also quite linear, but gives very confusing values for this error type because it reaches negative values above magnitude 0.6. Finally, ${}_u\alpha$, ${}_{c|u}\alpha$, and κ_d are not responsive at all to this error type, as expected, and remain blocked at 1 (which is normal for ${}_{c|u}\alpha$, which focuses on categorizing).

False positives and false negatives (Figure 28). In the current version of the CST, the false positive error type creates some overlapping (new units may overlap), and it is the reason why ${}_u\alpha$ and κ_d were discarded from this experiment. However, we have kept ${}_{c|u}\alpha$ because it behaves quite well despite overlapping units.

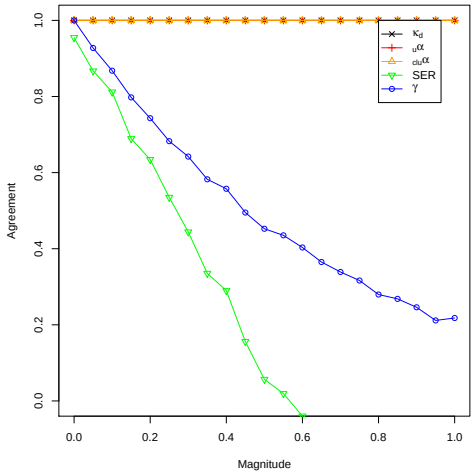
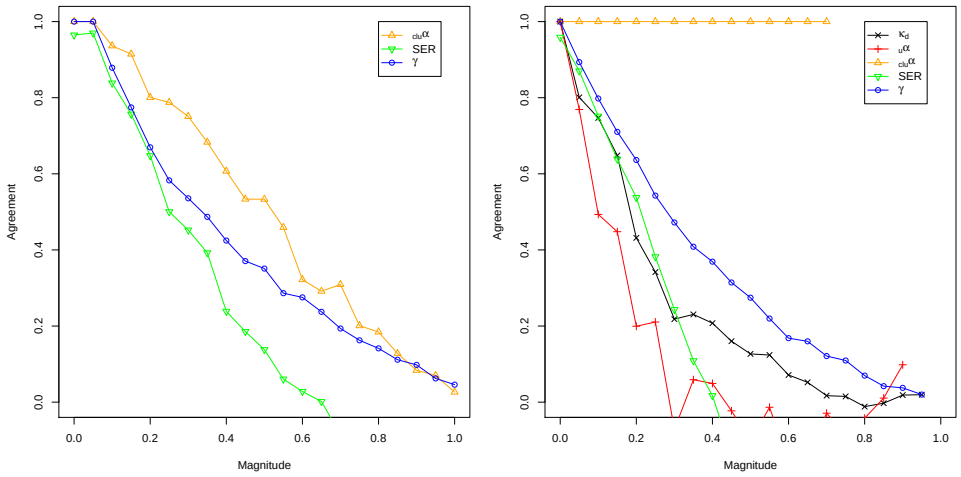


Figure 27
Agreement graph for splits.

All the measures have overall a good response to the false positives error type, as shown in Figure 28a, even if the shape of $_{c|u}\alpha$ is delayed compared with the others, but it should be pointed out that *SER* has a curious and unfortunate final increasing section (not visible in the figure because this section is below 0).

On the other hand, bigger differences appear with false negatives (Figure 28b). γ is still strictly decreasing and almost reaches 0 (0.025), but $_{u}\alpha$ is not strictly decreasing, and is at 0 or below from $m = 0.3$; *SER* quickly drops below 0 from $m = 0.4$, κ_d is not strictly decreasing, and $_{c|u}\alpha$, as for splits, does not react at all but remains stuck at 1, which is desired for this coefficient focused on categories (values of $_{c|u}\alpha$ over $m = 0.7$



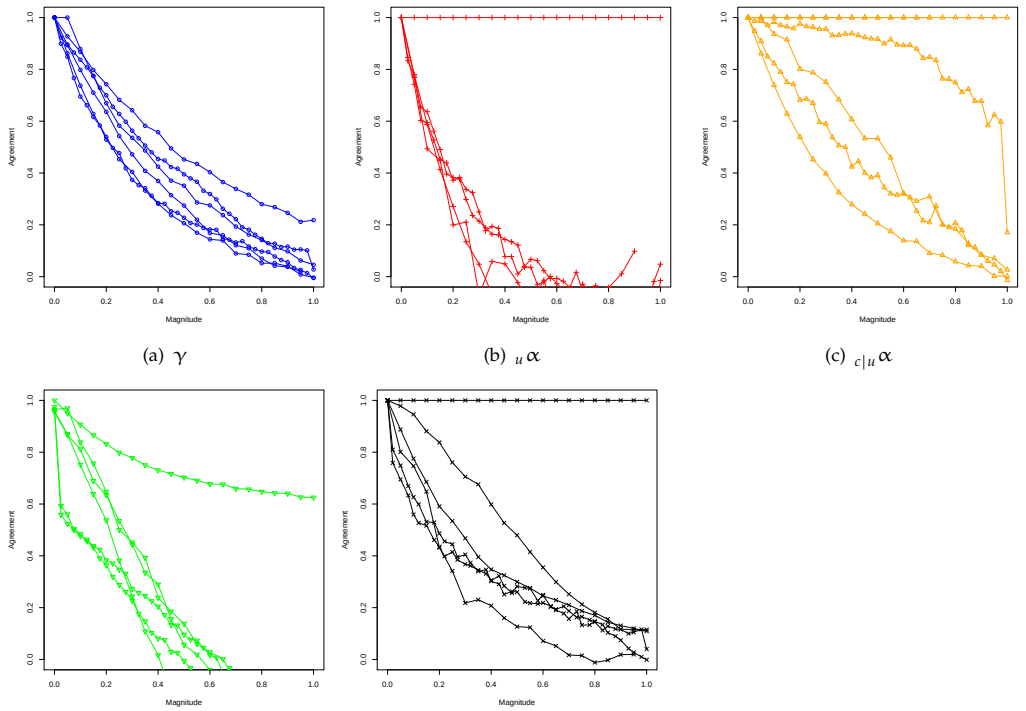
(a) False Positives (b) False Negatives

Figure 28
Agreement graphs for false positives and false negatives.

are missing since there are not enough intersections between units for this measure to work).

Overview of each measure for the almost general case. In order to summarize the behavior of each measure in response to the different error types for the almost general case (without overlap), we pick all curves relative to a given measure out of the previous plots and draw them in the same graph, as shown in Figure 29. Briefly, γ shows a steady behavior for all error types, almost strictly decreasing from 1 to 0. ${}_{u}\alpha$ has some increasing parts and negative values and is sometimes not responsive. ${}_{c|u}\alpha$ is very responsive for some error types, is less responsive for some other types, and is sometimes not responsive at all (which is desired, as already said). *SER* has unreliable responses, being either too responsive (reaching negative values) or not responsive enough. Finally, κ_d is not always responsive, is most of the time not strictly decreasing, but is sometimes quite progressive.

6.3.5 Fully General Case: Unitizing + Categorizing. This last section considers the fully general case, where overlapping of units within an annotator is allowed. In this experiment, we took a reference corpus with no overlap, but the errors applied (combination of positioning and false positives) progressively lead to overlapping units. The results are shown in Figure 30. As expected, γ behaves quite the same as it does with non-overlapping configurations. Admittedly, ${}_{c|u}\alpha$ was not designed to handle these configurations (and so should not be included in this experiment), but surprisingly it



(a) γ (b) ${}_{u}\alpha$ (c) ${}_{c|u}\alpha$

(d) *SER* (e) κ_d

Figure 29
Overviews: γ, ${}_{u}\alpha$, ${}_{c|u}\alpha$, *SER*, and κ_d.

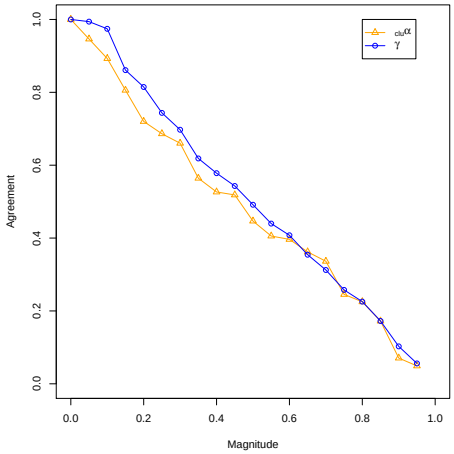


Figure 30
An example of graph for positioning + false positive errors leading to some positional overlapping.

seems to perform in rather the same way as it does with no overlapping; this must be investigated further, but judging from this preliminary observation, it seems this coefficient could still be operational and useful in such cases. On the contrary, ${}_{u}\alpha$ does not handle correctly this experiment and so was not included in the graph.

7. Conclusion

The present work addresses an aspect of inter-annotator agreement that is rarely studied in other approaches: the combination of unitizing and categorizing. Nevertheless, the use of methods that have been transposed from other domains (such as κ, which was originally dedicated to pure categorizing) in CL, for example at the discourse level, leads to severe biases, and manifests the need for specialized coefficients, fair and meaningful, suitable for annotation tasks focusing on complex objects.

In the end, only Krippendorff's coefficients ${}_{u}\alpha$ and ${}_{c|u}\alpha$ come close to the needs we expressed in the introduction, with the restriction that they are natively limited to non-overlapping units.

The main reason why research on this topic is sparse, and why it may be difficult to enlarge Krippendorff's coefficients to overlapping units, probably results from the fact that we are facing here a major difficulty: the simultaneous double discrepancy between annotators, with annotations possibly differing both in positioning relevant units anywhere on a continuum, and in categorizing each of these free units. Consequently, it is difficult for a method to choose precisely which features to compare between different annotators (unlike pure categorizing, where we know exactly what each annotator says for each predefined element to be categorized); and this problem is exacerbated when overlapping units (within an annotator) occur.

To cope with this critical point, we advocate the use of an alignment that ultimately expresses which unit from one annotator should be compared to which unit, if any, from another one, and consequently makes it natural and easier to compute the agreement. Moreover, we have shown that this alignment cannot be done in an independent way, but is part of the measure method itself. This is the "unified" aspect of our approach.

We have also shown that in order to be relevant, this alignment cannot be done at a local level (unit by unit), but should consider the whole set of annotations at the same time, which is the "holistic" aspect.

This is how the new method γ presented here was designed. Moreover, this method is highly configurable to cope with different annotation tasks (in particular, boundary errors should not necessarily be considered the same for all kinds of annotations), and it provides the alignment that emerges from an agreement measurement. Not only is this alignment a result in itself, which can be used to build a gold standard very quickly from a multi-annotated corpus (by listing all unitary alignments, and for each of them showing the corresponding frontiers and category proposed by each annotator), but it also behaves as a kind of a "flight recorder" of the measure: Observing these alignments gives crucial information on the choices the measure makes and whether it needs to be adjusted, unlike other methods which only provide a sole "out of the box" value.

Finally, we have compared γ to several other popular coefficients, even in their specific domains (pure categorization, pure segmentation), through a specific benchmark tool (namely, CST) which scans the responsivity of the measures to different kinds of errors and at all degrees of severity. Overall, γ provides broader and more progressive responsivity than the others in the experiments shown here. Concerning pure categorizing, γ does not have an edge over the well-known coefficients, such as α, but it is interesting to see that it behaves in much the same way as others in this specific field. Concerning segmentation, γ outperforms WD and GHD, by taking chance into account, but also by not depending on the heterogeneity of the segment sizes. Concerning unitizing with categorizing, as theoretically expected and confirmed by the benchmarking, SER shows severe limitations, such as a binary response to various (small or severe) positional or categorial errors, the fact that it does not make chance correction, or its limitation to two annotators only. Krippendorff's coefficients ${}_{u}\alpha$ and ${}_{c|u}\alpha$ present very interesting properties, such as chance correction. However, as we have shown with thorough comparisons, they rely on quite different hypotheses to ours, since they consider intersections between units whereas we advocate considering aligned units. We have identified several situations in CL where considering alignments is advantageous, for instance, when contiguous segments of the same type may occur, or when errors on several short units should be considered as more serious than one error on a long unit, but we do not posit these situtations as a universal rule. In conclusion, when unitizing and categorizing involve internal overlapping of units, only γ is currently available, and, even if it cannot be compared to any other method at the moment for this reason, benchmarking reveals very similar responses to overlapping configurations and to non-overlapping ones, which already demonstrates its consistency and its relevance.

We can summarize the features of γ as follows: It takes into account all varieties of unitizing, combines unitizing and categorizing simultaneously, enables any number of annotators, provides chance correction, processes an alignment while it measures agreement, and provides progressive responsivity to errors both for unitizing and for categorizing. This makes γ suitable for annotation tasks such as relative to NAMED ENTITY, DISCOURSE FRAMING, TOPIC TRANSITION, or ENUMERATIVE STRUCTURES.

The full implementation of γ is provided as Java open-source packages on the `http://gamma.greyc.fr` Web site. It is already compatible with annotations created with the Glozz Annotation Platform (Widlöcher and Mathet 2012), and with annotations generated by the Corpus Shuffling Tool.

Appendix A. Examples of Linguistic Objects and Possible Annotation Tasks

Terms emphasized hereafter refer to the terminology defined in Section 2.

PART-OF-SPEECH. Part-of-speech (POS) tagging (see, for example, Güngör [2010] for a recent state of the art) gives a well-known illustration of a pure *categorization without unitizing* task: for all the words in a text (*predefined units*, *full-covering*, *no overlap*), annotators have to select a category, belonging to quite a small set of exclusive elements. POS units (words) having the same label are obviously not *aggregatable*.

GENE RENAMING. In a study on gene renaming presented in Fort et al. (2012), all the tokens (*predefined units*, *no overlap*) are markable (*categorization*) with "Nothing" (the default value), "Former" (the original name of a gene) or "New" (its new name). This work at word level considers sparser (*sporadic*) phenomena than POS tagging. However, the annotation is defined as *full-covering*, with "Nothing" as a default tag. Note that the presence of the "Nothing" category also reveals here the reduction of a *unitizing* problem (detection of renaming) to a pure coding system (*categorization*). These units are not *aggregatable*.

WORD SENSE. For the annotation task described in Passonneau et al. (2012), annotators were asked to assign sense labels (*categorization without unitizing*) to preselected moderately polysemous words (*sporadicity*, *predefined units*, *no overlap*) in preselected sentences where they occur. Adjacent words are *not aggregatable* with sense preservation.

NAMED ENTITIY. Well-established named entity (NE) recognition tasks (see, for example, Nadeau and Sekine [2007]) led to many annotation efforts. In such tasks, the annotator is often asked to identify the units in the text's continuum (*unitizing*, *sporadicity*) and to select a NE type from an inventory (*categorization*). It is well known that some difficulties of NE annotation relate to the delimitation of NE boundaries. For example, for a phrase such as "Mr X, the President of Y," it makes sense to annotate subparts ("X," "Mr X," "the President of Y") and/or the whole. "Y" is also a NE of another type. This may result in *hierarchical* or *free overlapping* structures. Adjacent NE are *not aggregatable*.

ARGUMENTATIVE ZONING. Studies concerned by argumentative zoning (Teufel 1999; Teufel, Carletta, and Moens 1999; Teufel and Moens 2002) consider the argumenative structure of texts, and identify text spans having specific roles. For each sentence (*full-covering*, *predefined units*, *no overlap*), a category (*categorization*) is selected. Adjacent sentences of the same type are aggregated into larger spans (argumentative zones). This reveals an underlying question of *unitizing*. However, it has to be noted that the *categorization* mainly concerns predefined sentences: argumentative types are *aggregatable*.

DISCOURSE FRAMING. In Charolles et al.'s discourse framing hypothesis (Charolles et al. 2005, page 121), "a discourse frame is described as the grouping together of a number of propositions which are linked by the fact that they must be interpreted with reference to a specific criterion, realized in a frame-initial introducing expression." Thus, temporal or spatial introducing expressions lead, for example, to temporal or spatial discourse frames in the text continuum (*unitizing*, *sporadicity*, *categorization*). Discourse frames are *not aggregatable*. Subordination is possible, leading to possibly *hierarchical overlap*, where frames (of the same type or of different types) are embedded.

COMMUNICATIVE BEHAVIOR. The multimodal AMI Meeting corpus (Carletta 2007) covers a wide range of phenomena, and contains many different layers of annotation describing the communicative behavior of the participants of meetings. For example, in Reidsma (2008), annotators are required to identify fragments in a video recording (*unitizing*, *sporadicity*) and to categorize them (*categorization*). For such an annotation task, one can easily imagine instruction manuals allowing annotators to use multiple labels and to identify embedded (*hierarchical overlap*) or *free overlapping* units, even if the example provided by Reidsma (2008) does not.

DIALOG ACT. Annotating dialog act conforming to a standard as defined, for example, in Bunt et al. (2010), leads annotators to assign communicative function labels and types of semantic content (*categorization*) to stretches of dialogue called functional segments. The possible mulitfunctionality of segments (one functional segment is related to one or more dialog acts), and the fact that annotations may be attached directly to the primary data such as stretches of speech defined by begin and end points, or attached to structures at other levels of analysis, seems to allow different kinds of configurations and annotation instructions: *unitizing* or pure *categorization* of pre-existing structures, *sporadicity* or *full-covering*, *hierarchical*, *overlapping* or *linear segmentation*.

TOPIC SEGMENTATION. Topic segmentation (see, for example, the seminal work by Hearst [1997] or Bestgen [2006] for a more recent state of the art), which aims at detecting the most important thematic breaks in the text's continuum, gives an illuminating example of pure segmentation. This *unitizing* problem of *linear segmentation* is *full-covering* and restricted to the detection of breaks (the right boundary of a unit corresponds to the left boundary of the following segment) (*no overlap*). If we consider the resulting segments, there is just *one category* (topic segment) and then *no categorization*. Adjacent topic segments are obviously *not aggregatable* without a shift in meaning.

HIERARCHICAL TOPIC SEGMENTATION. In order to take better into account the fact that lexical cohesion is a multiscale phenomenon and that discourse displays a hierarchical structure, hierarchical topic segmentation proposed, for example, by Eisenstein (2009) preserves the main goal and properties of text-tiling (*unitizing*, *no categorization*, *full-covering*, *not aggregatable segments*), but allows a topic segment to be subsegmented into sub-topic segments (*hierarchical [but not free] overlap*).

TOPIC TRANSITION. The topic zoning annotation model presented in Labadié et al. (2010) is based on the hypothesis that, in a well constructed text, abrupt topic boundaries are more the exception than the rule. This model introduces *transition zones* (*unitizing*) between topics, zones that help the reader to move from one topic to another. The annotator is asked to identify and categorize (*categorization*) topic segments, introduction, conclusion, and transition zones. *Hierarchical overlap* is possible (embedded elements of the same type or of different types are allowed). *Free overlapping* structures are frequent, by virtue of the nature of transitions. Adjacent topic zones and adjacent transition zones are *not aggregatable*.

ENUMERATIVE STRUCTURES. A study on complex discourse objects such as enumerative structures (Afantenos et al. 2012) illustrates both the need for *sporadic unitizing* and the need for *categorization*. The enumerative structures have a complex internal organization, which is composed of various types of subelements (*hierarchical overlap*) (a trigger of the enumeration, items composing its body, etc.) which are *not aggregatable*.

Acknowledgments

We wish to thank three anonymous reviewers for helpful comments and discussion. The authors would also like to warmly thank Klaus Krippendorff for his support when they implemented his coefficients in order to test them. This work was carried out in the GREYC Laboratory, Caen, France, with the strong support of the French Contrat de Projet État-Région (CPER) and the Région Basse-Normandie, which have provided this research with two engineers, Jérôme Chauveau and Stéphane Bouvry.

References

Afantenos, Stergos, Nicholas Asher, Farah Benamara, Myriam Bras, Cécile Fabre, Mai Ho-Dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Pery-Woodley, Laurent Prévot, Josette Rebeyrolles, Ludovic Tanguy, Marianne Vergez-Couret, and Laure Vieu. 2012. An empirical resource for discovering cognitive principles of discourse organisation: The annodis corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2727–2734, Istanbul.

Artstein, Ron and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Beeferman, Douglas, Adam Berger, and John Lafferty. 1997. Text segmentation using exponential models. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing*, pages 35–46, Providence, RI.

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

Berry, Charles C. 1992. The K statistic [letter]. *Journal of the American Medical Association*, 268(18):2513–2514.

Bestgen, Yves. 2006. Improving text segmentation using latent semantic analysis: A reanalysis of Choi, Wiemer-Hastings, and Moore (2001). *Computational Linguistics*, 32(1):5–12.

Bestgen, Yves. 2009. Quel indice pour mesurer l'efficacité en segmentation de textes ? In *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis.

Bookstein, A., V. A. Kulyukin, and T. Raita. 2002. Generalized Hamming Distance. *Information Retrieval*, (5):353–375.

Bunt, Harry, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex C. Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. 2010. Towards an ISO standard for dialogue act annotation. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, pages 2548–2555, Valletta.

Carletta, J. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Carletta, Jean. 2007. Unleashing the killer corpus: Experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation*, 41(2):181–190.

Charolles, Michel, Anne Le Draoulec, Marie-Paule Pery-Woodley, and Laure Sarda. 2005. Temporal and spatial dimensions of discourse organisation. *Journal of French Language Studies*, 15:115–130, 7.

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

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

Di Eugenio, Barbara and Michael Glass. 2004. The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101.

Eisenstein, Jacob. 2009. Hierarchical text segmentation from multi-scale lexical cohesion. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 353–361, Stroudsburg, PA.

Fleiss, Joseph. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, (5):378–382.

Fort, Karën, Claire François, Olivier Galibert, and Maha Ghribi. 2012. Analyzing the impact of prevalence on the evaluation of a manual annotation campaign. In *Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 1474–1480, Istanbul.

Galibert, Olivier, 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 *Seventh International Conference on Language*

Resources and Evaluation (LREC 2010), pages 3453–3458, Valetta.
Goldman, Ronald L. 1992. The K statistic [reply]. *Journal of the American Medical Association*, 268(18):2513–2514.
Güngör, Tunga. 2010. Part-of-speech tagging. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, pages 205–236, Boca Raton, FL.
Gwet, Kilem Li. 2012. *Handbook of Inter-rater Reliability*. Advanced Analytics, LLC, third edition.
Hearst, Marti. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.
Israel, Gleen D. 1992. Determining sample size. Agricultural Education and Communication Department, University of Florida, IFAS Extension, PEOD6 (Reviewed 2013). Fact sheet.
Kazantseva, Anna and Stan Szpakowicz. 2012. Topical segmentation: A study of human performance and a new measure of quality. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 211–220, Stroudsburg, PA.
Krippendorff, Klaus, 1980. *Content Analysis: An Introduction to Its Methodology*, chapter 12. pages 129–154, Sage, Beverly Hills, CA.
Krippendorff, Klaus. 1995. On the reliability of unitizing contiguous data. *Sociological Methodology*, (25):47–76.
Krippendorff, Klaus. 2004. *Content Analysis: An Introduction to Its Methodology*, chapter 11. pages 211–256, Sage, Thousand Oaks, CA, 2nd edition.
Krippendorff, Klaus. 2011. Agreement and Information in the Reliability of Coding. *Communication Methods and Measures*, (5.2):1–20.
Krippendorff, Klaus, 2013. *Content Analysis: An Introduction to Its Methodology*, chapter 12, pages 267–328, Sage, Thousand Oaks, CA, 3rd edition.
Kuper, Jan, Horacio Saggion, Hamish Cunningham, Thierry Declerck, Franciska de Jong, Dennis Reidsma, Yorick Wilks, and Peter Wittenburg. 2003. Intelligent multimedia indexing and retrieval through multi-source information extraction and merging. In *IJCAI*, pages 409–414. Acapulco.
Labadié, Alexandre, Patrice Enjalbert, Yann Mathet, and Antoine Widlöcher. 2010. Discourse structure annotation: Creating reference corpora. In *Workshop on Language Resource and Language Technology Standards - State of the Art, Emerging Needs, and Future Developments*, La Valetta.
Lamprier, S., T. Amghar, B. Levrat, and F. Saubion. 2007. On evaluation methodologies for text segmentation algorithms. In *Proceedings of ICTAI 2007*, pages 19–26. Patras.
Makhoul, John, Francis Kubala, Richard Schwartz, and Ralph Weischedel. 1999. Performance measures for information extraction. In *Proceedings of DARPA Broadcast News Workshop*, pages 249–252, Herndon, VA.
Mathet, Yann and Antoine Widlöcher. 2011. Une approche holiste et unifiée de l'alignement et de la mesure d'accord inter-annotateurs. In *Traitement Automatique des Langues Naturelles 2011 (TALN 2011)*, Montpellier.
Mathet, Yann, Antoine Widlöcher, Karën Fort, Claire Francois, Olivier Galibert, Cyril Grouin, Juliette Kahn, Sophie Rosset, and Pierre Zweigenbaum. 2012. Manual corpus annotation: Giving meaning to the evaluation metrics. In *COLING 2012*, pages 809–817, Mumbai.
Nadeau, David and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26.
Olivero, Patrick. 2001. Calcul de la taille des Échantillons. CETE du Sud-Ouest / DAT / ZELT. Technical report.
Passonneau, Rebecca J., Vikas Bhardwaj, Ansaf Salleb-Aouissi, and Nancy Ide. 2012. Multiplicity and word sense: Evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46(2):219–252.
Pevzner, L. and M. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.
Reidsma, D. 2008. *Annotations and Subjective Machines of Annotators, Embodied Agents, Users, and Other Humans*. Ph.D. thesis, University of Twente.
Reidsma, D., D. K. J. Heylen, and R. J. F. Ordelman. 2006. Annotating emotions in meetings. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006*, pages 1117–1122, Paris. ELRA.

Reidsma, Denis and Jean Carletta. 2008. Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326.

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

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

Teufel, Simone. 1999. *Argumentative Zoning: Information Extraction from Scientific Articles*. Ph.D. thesis, University of Edinburgh.

Teufel, Simone, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117, Bergen.

Teufel, Simone and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Widlöcher, Antoine and Yann Mathet. 2012. The glozz platform: A corpus annotation and mining tool. In *ACM Symposium on Document Engineering (DocEng'12)*, pages 171–180, Paris.

Zwick, Rebecca. 1988. Another look at interrater agreement. *Psychological Bulletin*, (103):347–387.