Dialogue Club discussion note
May 23, 2005

# A kappa statistic for muldimensional dialogue act annotation

Harry Bunt
Tilburg University
harry.bunt@uvt.nl

## 1 Introduction

For measuring inter-annotator agreement, the standard kappa statistic is often used, which is defined as follows (Carletta, 1996)

$$\kappa \stackrel{def}{=} \frac{P(A) - P(E)}{1 - P(E)} \tag{1}$$

This statistic applies to sets of pairs of judgments, with P(A) the fraction of the judgements agreeing which are in agreement and P(E) the fraction of the judgements where agreement would be expected by chance. This formula applies to the comparison of judgements where only of two results is possible: agreement or disagreement.

When considering inter-annotator agreement for the use of multidimensional tags, this statistic is not an appropriate measure, because in the case of multidimensional annotation there can be *partial* agreement. Partial agreement can occur because one or more of the following situations may arise:

1. Annotators may assign different values within a hierarchical subsystem of communicative functions within a dimension (e.g. a task-oriented YN-Question or a Check). This corresponds to a relatively minor diagreement.

2. Annotators may disagree in whether or not they assign a tag in a dimension.

3. Utterances typically have a main function (possiblly more than one) and a number of secondary functions. Inter-annotator disagreement about secondary functions is less serious than disagreement about main functions.

4. Annotators may assign the same tags in the same dimensions, but differ in what they consider the main function (main dimension) and what they consider secondary functions.

5. Some secondary functions may be implied; for instance an answer to a question necessarily implies feedback. Whether or not an implied function is annotated or not does not really amount to an inter-annotator difference.

In this note we suggest a way to define a 'multidimensional kappa statistic' to deal with these complications.

# 2 A metric for partial inter-annotator agreement

## 2.1 Partial agreement within a dimension

The communicative functions in a dimension may be organised into one or more hierarchies. Different values within a dimension are mutually exclusive, except when they belong to the same hierarchy, in which case they express partial agreement. Two communicative function (CF) names that differ only one level in a hierarchy express a smaller disagreement then values that differ several levels in the hierarchy. If annotations are allowed to be highly underspecified, being permitted to use not only CF names but also names of dimensions and layers, then the assignment of the name of a dimension (or of a layer) represents partial disagreement compared with the assignment of a specific CF, since the dimension name (layer name) also covers CFs that would be in full disagreement with the CF name.

These considerations are taken into account in the following metric $\mu_d$ for a dimension $d$:

$$\mu_d(t_i, t_j) = \alpha^{\Lambda(t_i,t_j)} h(t_i, t_j) \Delta_{ij} \tag{2}$$

where:

- $\alpha$ is a constant, expressing the measure of agreement between two tags within a hierarchy that differ one level; a plausible value for $\alpha$ could be 0.9;
- $\Lambda(t_i, t_j)$ is the number of levels in which $t_i$ and $t_j$ differ within the same hierarchy;
- $h(t_i, t_j) = 1$ if $t_i$ and $t_j$ belong to the same hierarchy and 0 otherwise;
- for a dimension (or layer) with several subhierarchies, $\Delta_{ij} = 0.5$ if $i$ is the level of the dimension (or layer) name and $j$ is the top level of a hierarchy within that dimension (or layer); otherwise $\Delta_{ij} = 1$.[1]

## 2.2 Partial agreement in multiple dimensions

Moving from partial agreement in one dimension to that in multiple dimensions, we have to take into account the complications 2–5 mentioned above. We consider these in turn.

[1] $\Delta_{ij} = 1$ also for a dimension (or layer) with only one subhierarchy. Such dimensions (and layers) would be redundant, and their use should be avoided.

**Disagreement about whether to assign a tag in a certain dimension.** Complication 2, whether or not a tag is assigned in a certain dimension, may occur not only because annotators disagree but also because they have different assumptions about what it means not to tag an utterance in a dimension. One interpretation of the absence of a tag is that the annotator thinks the utterance does not have any function in that dimension; another is that the utterance is viewed as having a default value in that dimension. For instance, suppose that for the dimension of topic management there is a default value *'Topic Continuation'*. If nothing special happens in the dimension, that value may be considered the default, and default values do not have to be marked. We assume that *inherent to the notion of a default value is that, if no value is assigned in that dimension, then its value is meant to be the default value.* If a dimension has no default value, then absence of a tag in that dimension must be interpreted as "no value". An annotation system that forces the annotator to assign a tag in every dimension will need to have pseudovalues, meaning "no value", for those dimensions.

In order for the partial agreement metric to be independent of specific choices concerning default values and pseudovalues, we will assume that dimensions may have default values as well as pseudovalues, and we allow but do not force an annotation system to always assign a tag in each dimension.

In multidimensional tagging, a tag $t$ is a list $< t_1, t_2, ..., t_n >$.[2]. We assume a given multidimensional, layered DA assignment system as defined in Bunt (2005), with the dimensions $D = \{D_1, D_2, ..., D_n\}$. For each dimension $D_i$ we indicate its default value, if it has one, as $t_{id}$ and its pseudovalue ("no value"), if it has one, as $t_{i0}$.

For an annotation system that does not force the annotator to assign a tag in every dimension we have to consider inter-annotator agreement between two tags that may be of different lengths. To measure their agreement, we supplement both tags with components in those dimensions where they have no values as follows:

1. for each dimension $D_j$ which has a default value, we insert the tag component $t_{jd}$.

2. for each dimension $D_i$ which has no default value, we insert the tag component $t_{i0}$. (For those dimensions that have no 'no value' pseudovalue, this tag is added here just for the purpose of measuring agreement.)

**Disagreement about main vs. secondary functions.** We will propose a measure of multidimensional agreement that is a weighted sum of the agreement per dimension. To take into account that disagreements about main function are more serious than those about secondary functions, we use different weights for main and secondary functions, for instance assigning a weight 0.5 to secondary functions. In an annotation system that does not distinguish between

[2]We use the term 'tag' both to indicate such complex tags as well as to indicate component tags. Maybe we should introduce separate terminology to avoid any possible confusion...

main and secondary functions, all weights are simply 1.

**Disagreement about dimension of main function.** In an annotation system that does distinguish main and secondary functions, it may happen that annotators disagree about the dimension(s) in which an utterance has its main function(s). When using such an annotation system, we assume each tag component $t_i$ to be marked as either main or secondary.

**Disagreement about implied functions an certain dimensions.** We assume that differences in tagging an utterance for implied functions is only an apparent disagreement. They should not contribute the measure of disagreement.

We now define the multidimensional metric $\mu a$ as follows.[3] Let $\mu_i$ be defined as in (2) for a single dimension $D_i$. Then:

$$\mu a(t, t') \stackrel{def}{=} \frac{\sum_{i=1}^{n} w_i \mu_i(t_i, t'_i) + \sum_{i=1}^{n} \beta(t_i, t'_i)}{\sum_{i=1}^{n} \delta_i w_i} \tag{3}$$

where:

- with $w_i$ a weighting factor, used to differentiate between main functions, secondary functions and implied functions. A reasonable setting of values for $w_i$ would be as follows:
  - for those tag components marked as main functions in $t$ or in $t'$: $w_i = 1$
  - for secondary functions $w_i = 0.5$
  - for implied functions $w_i = 0$[4]
- $\beta(t_i, t'_i)$ expresses the seriousness of a disagreement about whether or not to assign a tag in a given dimension. Such a disagreement less serious than that of assigning conflicting tags, so a plausible value for it might be a 0.3 agreement. Therefore:
  $\beta(t_i, t'_i) = 0.3$ if $t_i = t_{i0}$ or $t'_i = t_{i0}$ and $\beta(t_i, t'_i) = 0$ otherwise.
- $\delta_i = 0$ if $t_i = t_{i0}$ or $t'_i = t_{i0}$ and $\delta_i = 1$ otherwise

## 2.3 Examples of partial agreement

We calculate the $\mu a$ values for some examples of single annotations, using the assumed values mentioned above.

[3] Since the Greek letter $\mu$ is transliterated as 'mu', the combination $\mu a$ may be pronounced as *'mua'* or *'mwha'* – which is what people often say when they do not fully agree or are not entirely happy with what someone else had said. The $\mu a$ metric can be seen as a formal metric for this *'mwha'* feeling.

[4] Taking into account that certain functions may be implied by other functions requires the annotation system to contain that information.

1. $\mu a$(YN-Question, Check) =
   $(0.9)^{\Lambda(QN-Question,Check}h$(YN-Question, Check)$\Delta$(YN-Question, Check)
   $= (0.9)^1 \times 1 \times 1 = 0.9$

2. $\mu a$(Inform + Pos.Feedback, Correction) = $\frac{0.925+0.5\times0.3}{1.5} = 0.78$

3. $\mu a$(Contact Man. + Turn-giving, Turn-giving + Contact Man.) =
   $\frac{0.5\times1+0.5\times0.5\times1}{1.5} = 0.5$

# 3 A multidimensional kappa statistic

We generalize the definition of the standard kappa statistic for a given set $J$ of pairs of judgements by replacing the fraction $P(A)$ of binary agreement judgements in $J$ by the average partial agreement over $J$:

$$\kappa_{\mu a}(J) = \frac{\overline{\mu a(\tau)} - P(E)}{1 - P(E)} \tag{4}$$

where $\overline{\mu a(\tau)}$ is $\frac{\sum_{\tau \in J} \mu a(\tau)}{|J|}$ and $P(E)$ is the value that $\overline{\mu a(\tau)}$ would be expected to have by chance.

For the DIT DA taxonomy, with 11 dimensions, 15 layers, and around 85 communicative functions, the expected chance value of $\overline{\mu a(\tau)}$ upon purely random assgnment of tags would be extremely low, less then 0.01. Against that baseline, $\overline{\mu a(\tau)}$ is a good approximation of $\kappa_{\mu a}(J)$.

A more sensible baseline is formed by taking a corpus of annotations, determining the relative frequencies of different tags, and consider random assignment of tags in accordance with those frequencies. For instance, one may find that annotators in general assign a main function in the Task/Domain dimension around 50% of the time, and within that dimension assign the Inform function 50secondary positive auto-feedback function and a turn management function, but rarely more then three secondary functions, etc. Against such a baseline, formula (4) cannot be reduced to the average agreement over the assigned tags.

# Bibliography


Bunt, H. (2005) A framework for dialogue act specification. Discussion paper for the 4th Joint ISO-SIGSEM Workshop on the Representation of Multimodal Semantic Information, Tilburg, January 10–11, 2005. Available at `http://let.uvt.nl/research/ti/sigsem/tdg`.

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