CHAPTER 9

Analysis of Nominal-Scale Inter-Rater Reliability Data

OBJECTIVE

This chapter presents statistical techniques for analyzing the extent of agreement among raters in special situations that do not fit well into the general framework that was developed in the past few chapters. For example, evaluating inter-annotator agreement in computational linguistics or in Natural Language Processing (NLP) entail a host of new challenges not encountered in most practical settings. Testing the difference of agreement coefficients for statistical significance is often done in practice and will be discussed in this chapter. A few other special situations will be addressed in this chapter as well.

Contents

9.1 Overview

This chapter addresses several important and seemingly unrelated topics regarding the design and analysis of inter-rater reliability experiments. These topics do not fit well into the general framework of inter-rater reliability assessment presented in the past few chapters and require a separate treatment given their importance in practice.

In section 9.2, I will present methods for evaluating inter-annotator agreement in computational linguistics or in natural language processing. The methods presented in the previous chapters all assume that the subjects to be rated are identified before the beginning of the experiment, and are the same for all raters. In content analysis however, each annotator must first identify segments of text, which are then classified into one of several possible categories. However, the text segments to be categorized may differ from one annotator to another one. Therefore, categorization is no longer the sole determinant of raters' agreement. Both segmentation and categorization must be accounted for when quantifying agreement. The need to combine both concepts for defining agreement is what makes inter-annotator agreement a special problem, which requires a special treatment.

In section 9.3, I address the problem of testing the difference between 2 agreement coefficients for statistical significance. It is common in practice to evaluate the extent of agreement among raters on 2 occasions and to want to know whether the observed difference is statistically significant. When the subject and rater samples in both experiments are independent, then standard a standard t-test could be used. However, samples of subjects or raters may overlap, in which case the 2 associated agreement coefficients will be dependent and the standard t-test is no longer applicable. The well-known paired t-test could have been used, except that the most widely-used agreement coefficients are nonlinear, making the standard paired t-test not applicable. The approach presented in this chapter was recommended by Gwet (2016).

Section 9.4 is devoted to a special problem related to the design of inter-rater reliability experiments. How can one design an inter-rater reliability experiment, which makes it possible to quantify the extent of agreement among 3 raters or more, when a maximum of 2 raters can rate the same subject. The situation will occur if the data collection costs are prohibitive, or the data collection procedures are too demanding on human subjects. In section 9.4, I discuss an approach that resolves this problem.

In section 9.5, I briefly discuss how one would detect problem raters who may need further training, when the estimated inter-rater reliability coefficient is deemed low. I will also explore the situation where you want to identify specific subjects on

which raters experience difficulties reaching an agreement.

While previous chapters focussed primarily on the concept of inter-rater reliability, this chapter will discuss a different concept of *Intra-Rater Reliability* in section 9.6. In that section, I will show how by properly organizing your rating data, you can adapt the inter-rater reliability techniques of the previous chapters to compute intra-rater reliability coefficients. That is, no new technique will be needed to accomplish this task.

Although agreement coefficients in the literature are often associated with a group of raters scoring the same subjects, that is not always the case. It is common practice for researchers in the field of psychometrics for example, to evaluate the extent to which a set of questions contribute towards measuring the same concept. This internal consistency of the questionnaire is evaluated based on the responses provided by a group of respondents, and is often seen as an agreement among respondents. The most commonly-used coefficient in this evaluation is Cronbach's alpha that I discuss in section 9.7.

9.2 Inter-Annotator Reliability in Natural Language Processing (NLP)

In the fields of Computational Linguistics (CL) and Natural Language Processing (NLP), text annotation is a critical activity for content analysis and machine learning. *Text annotation* is essentially the process of reading a text and adding notes, comments or tags to specific text segments in order to highlight a content of interest. In general, annotation allows humans to ensure they have captured the essence of what is happening in a document with a thorough evaluation of many segments of text. Moreover, a database of annotated text data has become essential for training machine algorithms.

In the introduction section 9.2.1, I present an overview of several aspects of the problem of quantifying the inter-annotator agreement (IAA). In sections 9.2.2 and 9.2.3 I will formulate an approach for quantifying the IAA, which to some extent differs from the approaches presented by Krippendorff (2004) and more recently Mathet et al. (2015). These 2 references along with Krippendorff (2018) provide most what is known today in the field of inter-annotator agreement. Other approaches to the IAA problem, which are based on the traditional information retrieval metrics[1] and used in the field of Medical Informatics, are discussed by Hripcsak and Rothschild (2005). The IAA field is as important as it is complex. In spite of all advances made in recent years, much research is still needed in that area to reduce the complexity of IAA assessment, improve their efficiency and promote their use among researchers.

[1]These methods are not discussed in this section. They rely on each annotation being categorized as "Positive," and "Negative," or as "Correct" and "Incorrect". However, different incorrect discourse units can substantially differ in their incorrectness, without these differences being accounted for.

9.2.1 *Introduction*

The real problem with text annotation is that different annotation projects may present challenges that can be so different that strategies developed for one project are likely to be revised before they can be applied to another project. How do you then know whether a given document was properly annotated? In small projects, one can count on an expert to review the work done. However, in larger projects or in the absence of expert annotators, the strategy often used is to assign the same annotation task to several annotators and to evaluate the extent to which they agree. The ultimate goal is to quantify the ability of different annotators to achieve a consensus. This evaluation is done with a measure known as the *Inter-Annotator Agreement*. . A high Inter-Annotator Agreement (IAA) proves the existence of consensus and the reliability of the annotation process.

As indicated by Mathet et al. (2015), annotation can be complex and when it includes personal interpretation to some degree, "The very notion of 'truth' may even be utopian ⋯" In this case, inter-annotator agreement becomes an even more important measure. However, no matter which way you define it, the IAA is not your regular chance-corrected agreement coefficient such as those discussed in the past few chapters. Annotation projects often involve highly complex multi-stage processes, where even defining what constitutes an agreement between 2 annotators can be challenging.

Let us consider once again the practical example of section 2.4 taken from the medical field where an annotator must annotate a clinical note such a patient chart prepared by a physician, a nurse or a lab technician. The goal is to review the clinical note and to perform the following 2 tasks:

(*i*) Identify specific segments of text with a clinical finding in the form of a medical condition.

(*ii*) Mark the name of the clinical finding, and whether or not it was present or absent.

The medical note may contain the following sentence:

The patient has had a mild sore throat without fever

Figure 9.1: A sentence to be annotated

Now, suppose that after analyzing this same sentence, 2 annotators produce the

following 2 annotated texts:

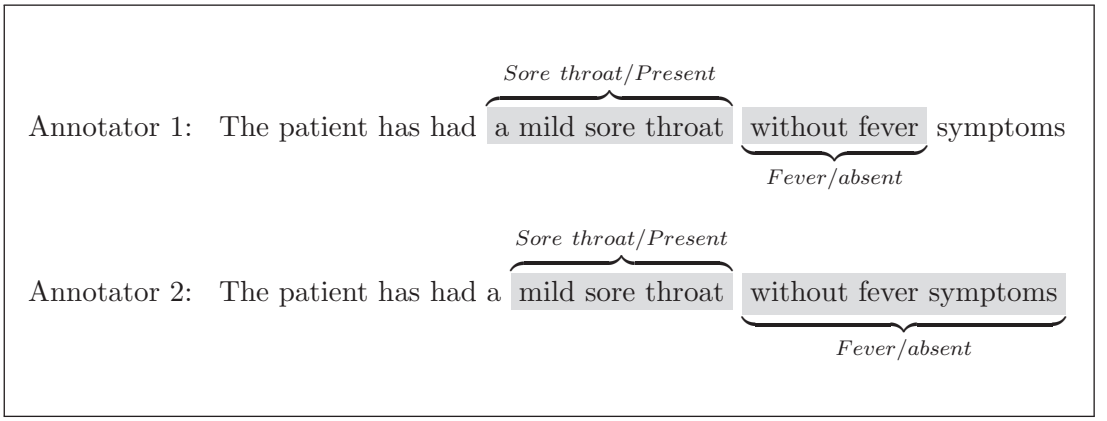

Figure 9.2: Annotation of a continuum by 2 annotators

How do you determine the extent to which these 2 annotators agree? Annotation here is a three-step process:

- The annotator must identify the text segments that will be annotated. This phase in the annotation process is known in the literature as *Unitizing*.
- The annotator must mark the correct name of the clinical finding (e.g. "Fever" or "Sore Throat"). This phase is the first step of what is known in the literature as *Categorization.*
- The annotator must label the clinical finding as "Present" or as "Absent." This phase is the second step of the categorization process.

Note that in simple annotation projects, the units or text segments can be pre-identified, in which case the only task assigned to annotators will be categorization. Under this scenario, one may use the standard chance-corrected agreement coefficients discussed in the past few chapters.

It is common that the unitizing task will be left to annotators. In this scenario, there is no way one can know the exact number of text segments that will be annotated nor their description. This information is expected to change from one annotator to another one. Moreover, if there are for example 79 possible clinical findings that must be labelled as present or absent, then categorization will involve a total of 158 possible categories. Therefore, for 2 annotators agree, they will need to agree on the unit, on the clinical finding and finally on the presence or absence of that finding. In each of the 3 levels of agreement, one may also expect partial agreement to occur at each level.

Mathet et al. (2015) provide a good overview of the current methods that have proposed in the literature. But, there are a few comments I like to make regarding this challenging problem of quantifying the IAA.

- Regarding the annotation problem I just described, I consider the likelihood that 2 annotators would reach agreement on unitization by pure chance very slim, given the level of effort often required in this process. As indicated by Savkov et al. (2016, page 536), agreement or disagreement by chance would be *"... negligible due to the relatively unrestricted position and length of each annotation."* Unitizing is not tinkering. Why would agreement occur here by pure chance? The notion of chance agreement was initially introduced in the context of simple problems where raters must classify subjects into a small number of categories (e.g. 2 , 3, 4, ...). Moreover, in these simple studies classification is often a one-step process that often does not require a close examination of hard facts. Annotation on the other hand, can reach an entirely different level of complexity where there is hardly any room for chance. At least during the segmentation phase. However, chance agreement may occur during the categorization phase of the annotation task. Therefore, categorization chance agreement should be accounted for when formulating Inter-Annotator Agreement coefficients.

- The unit of analysis to which a rating is assigned should not be the text segment (i.e. the annotation unit). This would make any statistical analysis a difficult task, since the number of text segments identified for annotation cannot be anticipated prior to the beginning of the annotation task, and changes from one annotator to another one. Instead, the unit of analysis would be a document, a paragraph, a page or chapter of a manuscript. The unit should always be predetermined before the annotation task begins.

Note that different annotation units created by an annotator could be distinct, can overlap or can be embedded in other annotation units. For any annotator however, as suggested by Krippendorff (2004) "... units of the same category may not overlap.". Consequently, overlapping units that are assigned to the same category should be collapsed into a single unit. The techniques discussed here are based on a pairwise comparison of annotation units from 2 annotators, which have a portion of the continuum in common. Two units that do not have any portion of the continuum in common are considered to be totally dissimilar. An alternative and more complex approach based on the notion of alignment of units is discussed by Mathet et al. (2015).

9.2.2 *Inter-Annotator Agreement: Generalities*

Let us introduce a few definitions, some of which are inspired by Mathet et al. (2015) although there is some discrepancy in symbolism.

- Let $\mathcal{A} = \{1, \cdots, i, \cdots, r\}$ be the set of r annotators who participated in the inter-annotator reliability experiment.

- *Continuum:* An annotation project is assumed to have been broken down in smaller and more manageable chunks that are named *Continuum* (or *Continua* when there are many of them). A continuum could be a chapter, a section, a paragraph or a sentence in a manuscript to be annotated. It may also be a portion of an audio or video recording and represents the fundamental unit of analysis, or the *Statistical Unit*, which *must be defined upfront before the beginning of any annotation project*. It is not an annotation unit, which will be defined next. Letter g will often be used to refer to a particular continuum.

- *Corpus:* The corpus is defined for a given annotation project as the set of all continua (plural of continuum). It is the set of all statistical units included in a given annotation effort. In statistical lingo, this would be equivalent to the statistical sample.

- Let n be the number of continua in a corpus. In statistical lingo, you would refer to n as the sample size.

- *Annotation Unit:* Within each continuum, the annotator will identify specific segments of text that will be annotated. These segments of text will be referred to as annotation units. While statistical units are predefined and fixed for all annotators, annotation units are defined by each annotator and are expected to vary from one annotator to another one. Annotation units are sometimes referred to in the literature as units of analysis (Krippendorff, 2004) and simply as units (Mathet et al., 2015). Letter j will be often used to designate a particular annotation unit.

  Figure 9.3 shows an example of annotation units defined based on annotation data from Figure 9.2. In this example, each annotator created 2 annotation units. Sometimes different annotators will create a different number of annotation units.

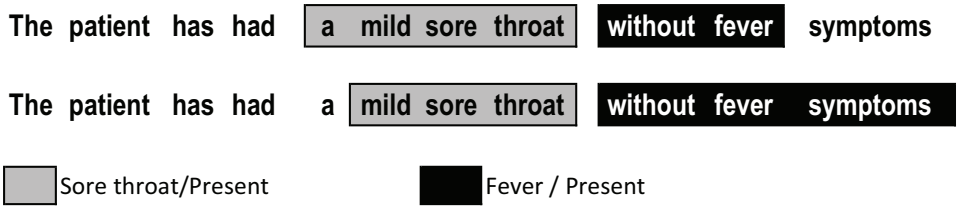

Figure 9.3: A continuum with 2 annotators and 3 response categories

- *Unitization Analysis Zone (UAZ):* The Unitization Analysis Zone (or the zone), is defined for a given pair of raters as a continuous portion of the continuum that is annotated by either annotator and on which both annotators agree, or disagree with respect to its categorization. It follows from Figure 9.4 that 4 zones were identified on the continuum of Figure 9.3. The first zone is made up of a single character "a", which is included in the first annotation unit of annotator 1 and categorized as "sore throat present", but excluded from the first unit of annotator 2. For annotator 1, zone 1 is associated with the annotated unit 1, while for annotator 2, zone 1 represents a gap. The second zone is defined by the text segment "mild sore throat," which both annotators categorized as "sore throat/present." The remaining 2 zones are defined in a similar manner.

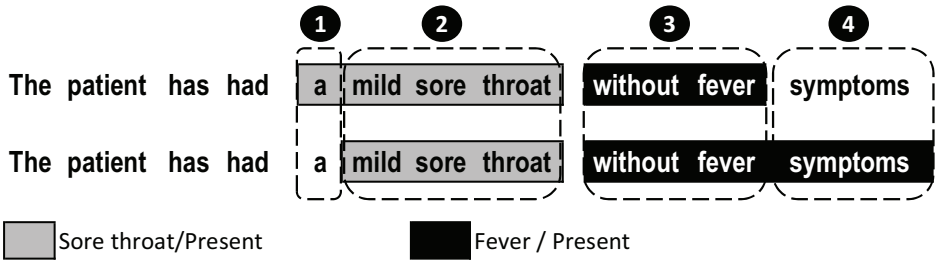

Figure 9.4: Definition of 4 Unitization Analysis Zones

- k_{ij} is a variable that defines the number of overlapping annotation units created by annotator i and which intersect with zone j. This also matches the number of different categories into which annotator i classified zone j, since overlapping units cannot be assigned to the same category. In Figure 9.4 for example, $k_{11} = 1$ (i.e. only one unit created by annotator $i = 1$ intersects zone $j = 1$) and $k_{21} = 0$ (i.e. no unit created by annotator 2 intersects zone 1). In most applications, k_{ij} will be either 0 or 1.

- $Cat_g(j,i)$ is defined for a given continuum g as the category or categories into which annotator i classified the unit(s) that intersect(s) zone j. More formally,

$$Cat_g(j,i) = \begin{cases} \cdot(missing) & \text{if } k_{ij} = 0, \\ \{1, \cdots, k_{ij}\} & \text{if } k_{ij} >= 1. \end{cases} \tag{9.2.1}$$

  Again, in most situations, $Cat_g(j,i)$ will either take a single value or be missing if zone j is not included in any unit created by annotator i.

- $l_{gj|i_1 i_2}$ refers to length of the j^{th} zone associated with continuum g and the pair of annotators (i_1, i_2).

- $Z^{(g)}_{i_1 i_2}$ is the number of zones produced by the annotator pair (i_1, i_2) from continuum g.

Figure 9.5 shows what a continuum annotated by 2 annotators would look like. The shaded rectangles represent the annotation units both annotators identified. The numbers inside these shaded rectangles are the categories into which the units must be classified. The length of this continuum is 22, and each unit within the continuum is defined by its beginning and by its length. As indicated by Krippendorff (2004), "*The unit for measuring these lengths is the smallest distinguishable length, duration, or number, for example the characters in text, frames of film, or smallest division on a ruler. Lengths are expressed in full integers, not in decimal points, not in units of varying size (like fractions of inches for small length and feet or miles for larger lengths.*"

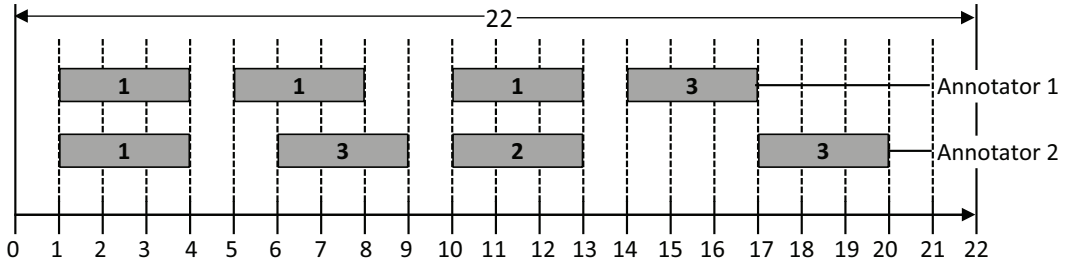

Figure 9.5: A continuum with 2 annotators and 3 response categories

For the sake of calculating the IAA coefficient, it is necessary to organize the different annotation units of Figure 9.5 into 6 section as shown in Figure 9.6. A zone may include 2 full units from 2 annotators (e.g. zones 1 and 4), one full unit and a gap (e.g. zones 5 and 6), one partial unit and a gap (e.g. zone 2) or 1 partial unit

and 1 full unit (e.g. 3). In what follows, a typical zone will generally by denoted by j.

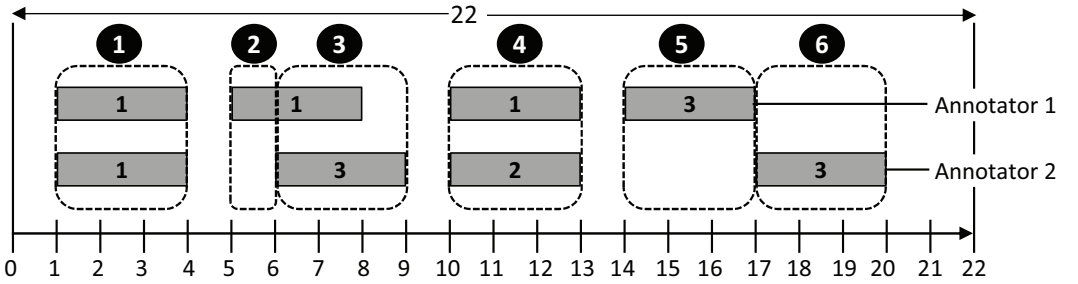

Figure 9.6: Defining annotation zones on a continuum

The inter-annotator agreement will now be calculated in the same way other chance-corrected agreement coefficients were calculated in the past few chapters. It is defined as (Cohen, 1960, page 40) put it "... *the proportion of agreement after chance agreement is removed from consideration.*" If A and C be respectively the 2 events "agreement between 2 annotators" and "chance agreement between 2 annotators", then the Cohen's definition translates into the conditional probability $P(\mathrm{A}/\overline{\mathrm{C}})$, that 2 raters agree given that chance agreement did not occur ($\overline{\mathrm{C}}$ being the complement of event C). This conditional probability can be rewritten as follows:

$$\begin{aligned} P(\mathrm{A}/\overline{\mathrm{C}}) &= \frac{P(\mathrm{A} \cap \overline{\mathrm{C}})}{P(\overline{\mathrm{C}})} \\ &= \frac{P(\mathrm{A}) - P(\mathrm{C})}{P(\overline{\mathrm{C}})}, \text{ (since } C \subset A)^{2} \\ &= \frac{P(\mathrm{A}) - P(\mathrm{C})}{1 - P(\mathrm{C})}. \end{aligned} \tag{9.2.2}$$

This is how one can see why the traditional form of chance-corrected agreement coefficients is correct. Unfortunately, from this point on, Cohen's brilliant idea was poorly executed first by himself and by many others afterwards. The problem lies on the formulation of the propensity of chance agreement $P(\mathrm{C})$. Many authors ended up interpreting $P(\mathrm{C})$ as the expected propensity for agreement under the assumption of random rating. Why would you want to make such an assumption? How does that assumption fit into the inter-rater reliability experiment that was conducted?

[2]Note that $C \subset A$, a notation commonly used in the field of probability, is another way of saying that any chance agreement is also an agreement. Therefore, C is a subset of A.

One may assume that chance agreement occurs when there is agreement between 2 raters and one of them at least performs a random rating (one can also consider a variant of this assumption). Therefore, the propensity for chance agreement $P(\mathrm{C})$ would be expressed as follows:

$$P(\mathrm{C}) = P(\mathrm{R})P(\mathrm{C}/\mathrm{R}) + \big[1 - P(\mathrm{R})\big]P(\mathrm{C}/\overline{\mathrm{R}}), \tag{9.2.3}$$

where $P(\mathrm{R})$ is the propensity for a rater to perform a random rating. Note that in equation 9.2.3, there is no general assumption of random rating that underlies the calculation of $P(\mathrm{C})$. While the conditional probabilities of equation 9.2.3 can be evaluated with no major problem, the situation is very different regarding the probability $P(\mathrm{R})$ for a rater or an annotator to perform a random rating. For studies that are conducted on an ongoing basis, practitioners may know often raters are unsure about their ratings. Alternatively, one may conduct a pilot study with a few raters and a few subjects, with the sole purpose of quantifying this probability. As a last resort, raters may be asked after at the end of the study, how often they were unsure about their ratings, and use the average of these numbers. One would hardly expect $P(\mathrm{R})$ to exceed 0.5. Otherwise, the whole rating exercise becomes meaningless.

Note from equation 9.2.3 that $P(\mathrm{C}/\overline{\mathrm{R}}) = 0$ (always). That is, whenever there is no random rating one cannot achieve agreement by pure chance. Consequently, the only formulation you need to worry about is given by,

$$P(\mathrm{C}) = P(\mathrm{R})P(\mathrm{C}/\mathrm{R}). \tag{9.2.4}$$

The main reason why the initial idea of chance agreement was poorly implemented first by Cohen himself, then by others is two-fold:

- It was assumed that $P(\mathrm{C}) = P(\mathrm{C}/\mathrm{R})$ without justification, overlooking the need to downweight the conditional probability $P(\mathrm{C}/\mathrm{R})$ by the propensity for random rating $P(\mathrm{R})$.
- The conditional probability $P(\mathrm{C}/\mathrm{R})$ itself was evaluated assuming the random assignment of ratings is governed by the observed probabilities of classification into the various categories. Consequently, if agreement among raters is highly concentrated in a few categories, then the chance-agreement probability will be close to 1, reducing the magnitude of the agreement coefficient to near 0. This is the primary reason why Cohen's Kappa, Fleiss' generalized kappa or Krippendorff's alpha could sometimes lead to spurious results.

In some applications, it is easier to evaluate the propensity for disagreement rather than the propensity for agreement. In this case, $P(\mathrm{A}/\overline{\mathrm{C}})$ can be rewritten as follows:

$$P(\mathrm{A}/\overline{\mathrm{C}}) = 1 - \frac{1 - P(\mathrm{A})}{1 - P(\mathrm{C})} = 1 - \frac{P(\overline{\mathrm{A}})}{P(\overline{\mathrm{C}})}, \tag{9.2.5}$$

where $P(\overline{\mathrm{C}})$ can be reformulated as follows:

$$P(\overline{\mathrm{C}}) = 1 - P(\mathrm{R})\Big[1 - P(\overline{\mathrm{C}}/\mathrm{R})\Big]. \tag{9.2.6}$$

This equation takes into consideration the fact that $P(\overline{\mathrm{C}}/\overline{\mathrm{R}}) = 1$. That is if there is no random rating, then no agreement by pure chance can occur. Equations 9.2.5 and 9.2.6 will be used in formulating the Inter-Annotator Agreement (IAA). Let θ_{G} be that IAA. It follows that,

$$\boxed{\theta_{\mathrm{G}} = 1 - \frac{P(\overline{\mathrm{A}})}{1 - P(\mathrm{R})\Big[1 - P(\overline{\mathrm{C}}/\mathrm{R})\Big]}.} \tag{9.2.7}$$

Now, what is needed are an expression for $P(\overline{\mathrm{A}})$ and another one for $P(\overline{\mathrm{C}}/\mathrm{R})$. As for the propensity for random rating $P(\mathrm{R})$, I would suggest to use $P(\mathrm{R}) = 0.5$, unless the researcher possesses a better value to use here.

9.2.3 *Calculating Inter-Annotator Agreement*

Let us first derive the probability $P(\overline{\mathrm{A}})$, which can be seen as a measure of dissimilarity. It is the propensity for 2 annotators to disagree. Let $d_{g|i_1i_2}$ be a measure of dissimilarity between 2 annotators i_1 and i_2 with respect to a particular continuum g. It is defined by,

$$d_{g|i_1i_2} = \sum_{j=1}^{Z^{(g)}_{i_1i_2}} \Big(l_{gj|i_1i_2}/L_{g|i_1i_2}\Big)^2 \Delta_{gj|i_1i_2}, \tag{9.2.8}$$

where $\Delta_{gj|i_1i_2}$ is the category dissimilarity[3], which is defined differently according as the categories are of nominal type or not, and $L_{g|i_1i_2}$ is the total length of all sections defined by,

$$L_{g|i_1i_2} = \sum_{j=1}^{Z^{(g)}_{i_1i_2}} l_{gj|i_1i_2},$$

for a given continuum g and pair of annotators i_1 and i_2.

If the categories are nominal then,

$$\Delta_{gj|i_1i_2} = \begin{cases} 1, & \text{if } Cat_g(j,i_1) \neq Cat_g(j,i_2) \text{ or a category is missing,} \\ 0, & \text{otherwise.} \end{cases} \tag{9.2.9}$$

[3]Zone-j category dissimilarity aims at quantifying the extent to which 2 units differ at their intersection with zone j, with respect to the categories they are assigned to.

If the categories are ordinal, interval or ratio then,

$$\Delta_{gj|i_1i_2} = \begin{cases} \dfrac{\left(Cat_g(j,i_1) - Cat_g(j,i_1)\right)^2}{\underset{c_1,c_2 \in \mathcal{K}}{Max} \left(c_2 - c_1\right)^2}, & \text{if no category is missing,} \\ 1, & \text{if a category is missing.} \end{cases} \tag{9.2.10}$$

where $\mathcal{K}$ is the list of categories used in the annotation project.

> Note that equations 9.2.9 and 9.2.10 assume that each zone j is classified into a single category. That will not always be the case. If units overlap then a zone j could intersect many of them and be classified into multiple categories. In this case, $Cat_g(j,i_1)$ will be a set of k_{i_1j} categories and $Cat_g(j,i_2)$ a set of k_{i_2j} categories. Each of these sets of categories must be sorted in ascending order. Then, the 2 series of categories will be paired sequentially. If one series has more categories than the other, then the unpaired categories will be paired with missing categories.
>
> For example, suppose $Cat_g(j,i_1) = \{1,2\}$ and $Cat_g(j,i_2) = \{1,3,4,5\}$. Then these categories will be paired as follows $\{(1,1),(2,3),(\cdot,4),(\cdot,5)\}$. Therefore, $\Delta_{gj|i_1i_2}$ will be will be evaluated for each of the 4 pairs according to equation 9.2.9 or 9.2.10, and its value divided by 4. That is, $\Delta_{gj|i_1i_2}/4$ will be used for each of the 4 pairs of categories retained.

The measure of dissimilarity with respect to continuum g is defined as follows:

$$d_g = \sum_{i_1<i_2}^{r} d_{g|i_1i_2}, \tag{9.2.11}$$

where $d_{g|i_1i_2}$ is given by equation 9.2.8.

To summarize, $P(\overline{\mathrm{A}})$ is given by,

$$P(\overline{\mathrm{A}}) = \frac{2}{r(r-1)} \frac{1}{n} \sum_{g=1}^{n} d_g, \tag{9.2.12}$$

or,

$$\boxed{P(\overline{\mathrm{A}}) = \frac{2}{r(r-1)} \frac{1}{n} \sum_{g=1}^{n} \sum_{i_1<i_2}^{r} \sum_{j=1}^{Z_{i_1i_2}^{(g)}} \left(l_{gj|i_1i_2}/L_{g|i_1i_2}\right)^2 \Delta_{gj|i_1i_2},} \tag{9.2.13}$$

The next step now, is to compute $P(\overline{\mathrm{C}}/\mathrm{R})$ (see equation 9.2.7). I will derive 2 expressions, one to be used with nominal ratings (with no ordering), and another one to be used with ordinal, interval or ratio ratings.

Since chance agreement is assumed to occur during the categorization phase only, the conditional probability $P(\overline{\mathrm{C}}/\mathrm{R})$ represents a measure of dissimilarity under the assumption of random rating. It is based on equation 9.2.13, and defined as follows:

$$P(\overline{\mathrm{C}}/\mathrm{R}) = \frac{2}{r(r-1)} \frac{1}{n} \sum_{g=1}^{n} \sum_{i_1<i_2}^{r} \sum_{j=1}^{Z_{i_1 i_2}^{(g)}} \left(l_{gj|i_1 i_2} / L_{g|i_1 i_2} \right)^2 \Delta_{gj}, \tag{9.2.14}$$

where Δ_{gj} is the zone-level dissimilarity, which is independent of a specific pair of annotators since categorization is assumed to be a random process. Δ_{gj} is defined differently according as the categories are nominal or not. Assume that the annotators must classify the units into 1 of K possible categories.

- *Nominal Categories*
  Let the list of nominal categories be $\mathcal{A} = \{1, \cdots, k, \cdots, K\}$.

$$\Delta_{gj} = \begin{cases} 1, & \text{if zone } j \text{ contains a gap,} \\ (K-1)/K, & \text{if zone } j \text{ contains no gap.} \end{cases} \tag{9.2.15}$$

- *Ordinal, Interval or Ratio Categories*
  Suppose that the list of ordinal, interval or ratio categories is given by $\mathcal{A} = \{x_1, \cdots, x_k, \cdots, x_K\}$.

$$\Delta_{gj} = \begin{cases} 1, & \text{if zone } j \text{ contains a gap,} \\ \dfrac{2}{K(K-1)} \displaystyle\sum_{k<l}^{K} \sum^{K} \left(\frac{x_k - x_l}{x_{Max} - x_{Min}} \right)^2, & \text{if zone } j \text{ contains no gap.} \end{cases} \tag{9.2.16}$$

Example 9.1

In this example, I will analyze annotation data previously reported in the appendix of Krippendorff (2004). This annotation task involves a single continuum and 2 annotators identified as "Annotator 1" and "Annotator 2." Each must identify text segments before classifying them into one of 2 categories labelled as "Category 1" and "Category 2." This completed annotation task is shown in Figure 9.7. It appears the total length of the continuum is L=300 units of measurements. Without loss of generality, the unit of measurement is assumed to be a character. The first 2 units of each annotator overlap. These are units that share a portion of the continuum.

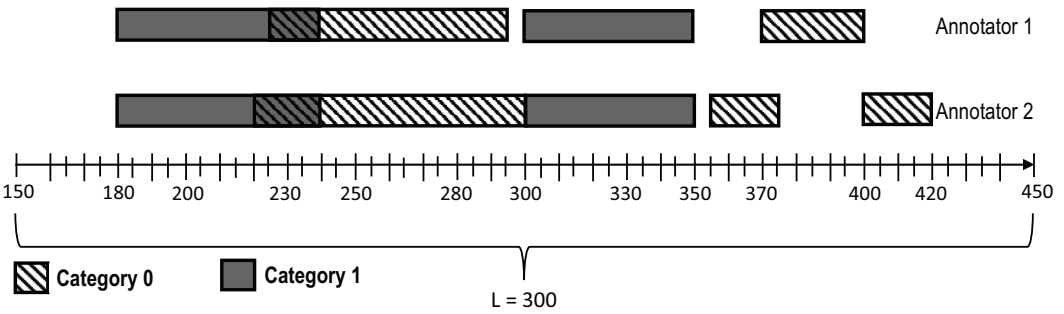

Figure 9.7: A continuum with 2 annotators and 3 response categories

Before annotated corpora can be analyzed, the data must be logically organized as shown in Table 9.1. The first column would be the continuum identifier, which in this case takes a single value as the current annotation task only involves one continuum. The second column identifies the annotator and contains the 2 values associated with the 2 annotators that performed the annotation task. In the third column, a number is assigned to each annotation unit identified by both annotators. In this example, I decided to number these units sequentially from 1 to the number of units identified. This numbering is also done separately for each annotator. The Start variable determines the start point of the unit on the continuum, while the Length variable represents the associated number of measurement units. Finally, the Category variable contains the category into which the associated unit was classified.

To compute the extent of agreement between these 2 annotators, you need to first define the *Unitization Analysis Zones* (UAZ) as shown in Figure 9.8. For this example, 9 UAZs were identified. I will first compute the propensity $P(\overline{A})$ for disagreement using equation 9.2.13. Then I will compute $P(\overline{\mathrm{C}}/\mathrm{R})$, the propensity for chance agreement given a random assignment of categories, using equation 9.2.14. Finally, the inter-annotator agreement θ_{G} of equation 9.2.7 with the assumption of a high propensity for