# Book Review

## Statistical Methods for Annotation Analysis

**Silviu Paun, Ron Artstein, and Massimo Poesio**
(Queen Mary University of London, University of Southern California, & Queen Mary University of London and Turing Institue)

Springer Nature (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 54), 2022, xix+197pp; paperback, ISBN 978-3-031-03753-5; ebook, ISBN 978-3-031-03763-4; doi:10.1007/978-3-031-03763-4

*Reviewed by*
*Rodrigo Wilkens*
*Université catholique de Louvain*

A common task in Natural Language Processing (NLP) is the development of datasets/corpora. It is the crucial initial step for initiatives aiming to train and evaluate Machine Learning and AI systems, for example. Often, these resources must be annotated with additional information (e.g., part-of-speech and named entities), which leads to the question of how to obtain these values. One of the most natural and widely used approaches is to ask for people (e.g., from untrained annotators to domain experts) to identify this information in a given text or document and possibly for more than one annotator per item. However, this is an incomplete solution. It is still necessary to obtain a final annotation per item and to measure agreement among the different annotators (or coders). Presenting a survey on this topic, Ron Artstein and Massimo Poesio published an article ("Inter-coder Agreement for Computational Linguistics") in 2008 that addressed the mathematics and underlying assumptions of agreement coefficients (e.g., Krippendorff's $\alpha$, Scott's $\pi$, and Cohen's $\kappa$) and the use of coefficients in several annotation tasks. However, it left open questions, such as the interpretability of coefficients of agreement, and it did not cover topics that nowadays are important (e.g., the research within the field of statistical methods for annotation analysis, such as latent models of agreement or probabilistic annotation models). In 2022, Silviu Paun, Ron Artstein, and Massimo Poesio published a book addressing primarily the NLP community but also including other communities, such as Data Science. They intended to offer an introduction to latent models of agreement, probabilistic models of aggregation, and learning directly from multiple coders. They also reintroduced the topics presented in 2008, making an incremental contextualization. Although the reliability (agreement between coders) and the validity (the "correctness" of the annotations) are present in the entire book, it is divided into two parts. The first part covers the development of labeling scheme coefficients of agreement such as $\pi$, $\kappa$, and their variants used in NLP and AI. The second part includes methods developed to analyze the output of annotators (e.g., the most likely label for an item among those provided).

Chapter 2 recaps the content presented by Artstein and Poesio (2008), updating the discussion to include recent progress. It mainly targets the coefficients of agreement and their purpose of reliability, which is a prerequisite for demonstrating the validity of a coding scheme. Also, the term "reliability" can be used in different ways:

https://doi.org/10.1162/coli_r_00483

intercoder agreement (or test stability), measuring reproducibility, and accuracy. After presenting a brief context and the notation used in the book, the question "why are custom coefficients to measure agreement necessary?" is explored by reviewing chance-adjusted measures, the percentage agreement by chance, and the specific agreement coefficients. Then, the authors address some of the most common design questions in any annotation project. They start by addressing missing data caused by the coders failing to classify items (for whatever reason). In this discussion, they provide possible actions, raising pros and cons. Then, they discuss the identification of units in tasks where the coder is also required to identify the item boundaries (e.g., beginning and end of a named entity). Finally, they address severely skewed annotated data, discussing the bias problem and the prevalence problem. They finish Chapter 2 presenting proofs for the theorems presented.

Chapter 3 presents agreement measures for computational linguistic annotation tasks, dividing them into three main points: methodology, choice of coefficients, and interpretation of coefficients. They describe the challenges of different annotation tasks (e.g., part-of-speech tagging, dialogue act tagging, and named entities), as well as of labeling with and without a predefined set of categories. This description of methodological choices of various studies goes along with an observation that even if a work may report agreement, it may not necessarily follow a methodology as rigorous as that envisaged by Krippendorff (2004). Concerning the choice of coefficients, they discuss the most basic and common form of coding in computational linguistics (i.e., text segment labeling with a limited number of categories), then present coding schemes with hierarchical tagsets and coding schemes with set-valued interpretations (e.g., anaphora and summarization). The discussion about coefficient interpretation looks at the range of values and different authors' positions. They also discuss the use of weighted coefficients, arguably more appropriate for some annotation tasks, and the challenges in their interpretability.

In Chapter 4 the authors present studies on how to interpret the results of reliability by rephrasing the problem as one of confidence estimation of a particular label given the behavior of the coders. The chapter starts by raising an important topic concerning the annotation: The coders easily agree about some items while other items seem more difficult to agree on. This leads to the concept of item difficulty. The items might also be viewed as latent classes, which can be modeled as the likelihood of a coder assigning a given label to an item given that item's latent class. Considering this reformulation, the authors discuss how to measure and model the agreement (including different probability distributions) and the coders' stability. This chapter ends Part 1 by moving the reader from an annotation task carried out by experts, which can accurately identify the labels, to a richer formulation where the interaction between both the label and the annotator may be considered. Therefore, it moves away from the simple majority choice, which ignores the accuracy and biases of coders as well the characteristics of the items.

Chapter 5 focuses on the probabilistic models of annotation. It begins with a simple annotation model introducing the terminology and some key assumptions frequently made. Next, this model is extended to cover the annotation pattern of the coders. After this introduction, the authors address the issue of item difficulty and how it can affect coders' annotation. They also discuss hierarchical priors for the annotators (which can be used to estimate annotators' behavior when the data is scarce), how to model the characteristics of the items to discriminate between the labels, and how to have a richer model of annotator ability. Moving on, they present models where the items have inter-dependent labels (e.g., named entity recognition or information extraction tasks) and where the labels are not predefined classes (e.g., anaphoric annotations). Then, by the chapter's end, the authors shift from encoding assumptions about the annotation process when inferring the ground-truth labels to neural networks to aggregate the annotations, using a variational autoencoder. Afterwards, they present notes on modeling other types of annotation data.

Chapter 6 addresses a different source of disagreement from that presented in Chapter 5—namely the item's difficulty, which can come from ambiguity, for example. This chapter covers methods for learning from multi-annotated corpora starting from covering the use of soft labels and the coders' individual labels. Later, the chapter moves to distill the labels dealing with noise and pooling coder confusion. Finally, the authors finish the chapter and the book with recommendations about when to apply each method depending on the characteristics of the datasets the models are to be trained on. This chapter finishes with a summary of the lessons learned including also topics like (1) the decision aggregate or keep all the annotations, (2) crowdsourced labels versus gold labels, and (3) mixed results and what did not work for them.

In summary, this book provides a complete perspective of statistical methods for annotation analysis in NLP, covering meaningful references and contextualizing them critically and historically at the same time, while also putting forth the assumptions behind the different coefficients. Moreover, the book provides several practical examples of annotation designs and how to measure their agreement. Thus, it provides an insightful perspective on what the agreement measures can and cannot do, which is present throughout the entire book. The content is suitable for both those who want to carry out research on the subject and for those who are interested in assessing reliability. From the perspective of someone who has an annotated corpus, some sections may be less interesting (i.e., specialized in different tasks), but the coverage of the various NLP tasks makes this book also a good guide for assessing reliability and validity.

*Rodrigo Wilkens* is a postdoctoral researcher at CENTAL, the Center for Natural Language Processing of the Université catholique de Louvain in Belgium. He has worked on NLP topics such as text simplification, readability assessment, automated essay scoring, knowledge extraction, question answering, and information retrieval, and has developed resources and tools for different languages in both industry and academia. His e-mail address is rodrigo.wilkens@uclouvain.be.