Assessing Reliability on Annotations (1): Theoretical Considerations

Jens Stegmann Andy Lücking
{jens.stegmann,andy.luecking}@uni-bielefeld.de

Collaborative Research Centre
SFB 360 "Situated Artificial Communicators"
B3 "DeiKon"

Abstract

This is the first part of a two-report mini-series focussing on issues in the evaluation of annotations. In this theoretically-oriented report we lay out the relevant statistical background for reliability studies, evaluate some pertaining approaches and also sketch some arguments that may lend themselves to the development of an original statistic. A description of the project background, including the documentation of the annotation scheme at stake and the empirical data collected, as well as results from the practical application of the relevant statistics and the discussion of our respective results are contained in the second, more empirically-oriented report [Lücking and Stegmann, 2005]. The following points are dealt with in detail here: we summarize and contribute to an argument by Gwet [2001] which indicates that the popular *pi* and *kappa* statistics [Carletta, 1996] are generally not appropriate for assessing the degree of agreement between raters on categorical type-ii data. We propose the use of AC_1 [Gwet, 2001] instead, since it has desirable mathematical properties that make it more appropriate for assessing the results of expert raters in general. As far as type-i data are concerned, we make use of conventional correlation statistics which, unlike their AC_1 and *kappa* cousins, do not deliver results that are adjusted with respect to agreements due to chance. Furthermore, we discuss issues in the interpretation of the results of the different statistics. Finally, we take up some loose ends from the previous chapters and sketch some advanced ideas pertaining to inter-rater agreement statistics. Therein, some differences as well as common ground concerning Gwet's perspective and our own stance will be highlighted. We conclude with some preliminary suggestions regarding the development of the original statistic *omega* that will be different in nature from those discussed before.

Acknowledgements

We shall start this report by saying "thank you" to some very important persons. That is, we joyfully express our gratitude towards the following people:

Peter Kühnlein and Manja Nimke, who are former project members that have participated in the conduct of the original empirical studies, as well as in the discussions and decisions that helped to shape the very first version of the present annotation scheme.

Hannes Rieser, our project leader, who, besides having escorted earlier versions in the past, has pushed and carried the continuation and broadening of the annotation work with an emphasis on dialogue phenomena. Furthermore, he was our co-author on a workshop contribution that proved to be an early milestone on our way to this report (he humbly refused from being mentioned as an author of the present work). Last but not least, he was an invaluable aid with regards to corrections and proof-reading all parts of this document. This notwithstanding, of course, all the remaining errors and possible sources of misconceptions—it deems to us, there may be some—are truely and utterly ours.

We also wish to thank our DEIKON project partners in computer science/artificial intelligence, namely Ipke Wachsmuth, Stefan Kopp, Marc Latoschik, Timo Sowa, Alfred Kranstedt, and Thies Pfeiffer. Furthermore, we would like thank all members of the Collaborative Research Centre SFB 360 "Situierte Künstliche Kommunikatoren" at Bielefeld University, wherein our project is located and which has been a great place to work in.

Finally, we wish to express our deepest gratitude towards our families, and especially to Christiane and Karola. They patiently endured all inconveniences that arose during our work on this document. Without their support, this report could not have been completed.

Bielefeld, December 2005, Andy Lücking and Jens Stegmann.

[*Rosencrantz and Guildenstern pass their time by betting on the toss of a coin in the following manner. Guildenstern takes a coin out of his bag, spins it, lets it fall. Rosencrantz studies it, announces it as "heads" (as it happens) and puts it into his own bag. They have been doing this for some time and are witnesses of a highly improbably run of "heads" for ninety-two times in a row. Guildenstern, who is losing all the time, is well alive to the oddity of it. He is worried about the implications, not so much about the money he loses.*]

ROSENCRANTZ: Heads. — Heads. — Heads. — Heads. — Heads.

GUILDENSTERN: There is an art to be building up of suspense.

ROSENCRANTZ: Heads.

GUILDENSTERN: Though it can be done by luck alone.

ROSENCRANTZ: Heads.

GUILDENSTERN: If that's the word I'm after.

ROSENCRANTZ: Heads.

GUILDENSTERN: A weaker man might be moved to re-examine his faith, if in nothing else at least in the law of probability.

ROSENCRANTZ: Heads.

GUILDENSTERN: The law of averages, if I have got this right, means that if six monkeys were thrown up in the air for long enough they would land on their tails about as often as they would land on their—

ROSENCRANTZ: Heads.

GUILDENSTERN: Which even at first glance does not strike one as a particularly rewarding speculation, in either sense, even without the monkeys. I mean you wouldn't *bet* on it. I mean *I* would, but *you* wouldn't.

ROSENCRANTZ: Heads. — Heads. Getting a bit of a bore, isn't it?

GUILDENSTERN: A bore? What about the suspense?

ROSENCRANTZ: What suspense?

GUILDENSTERN: Well, it was an even chance ... if my calculations are correct.

ROSENCRANTZ: Well ...

GUILDENSTERN: No questions? Not even a pause?

ROSENCRANTZ: You spun them yourself.

GUILDENSTERN: Not a flicker of doubt?

ROSENCRANTZ: Well, I won—didn't I?

GUILDENSTERN: And if you'd lost? If they'd come down against you, one after another, eighty-five times, one after another, just like that?

ROSENCRANTZ: Eighty-five times in a row? *Tails*?

GUILDENSTERN: Yes! What would you think?

ROSENCRANTZ: Well ... Well, I'd have a good look at your coins for a start!

GUILDENSTERN: I'm relieved. At least we can still count on self-interest as a predictable factor. We have been spinning coins together since—this is not the first time we have spun coins!

ROSENCRANTZ: It'll take some beating, I imagine.

GUILDENSTERN: Is *that* what you imagine? Is that it? No *fear*?

ROSENCRANTZ: Fear?

GUILDENSTERN: *Fear!* The crack that might flood your brain with light!
ROSENCRANTZ: Heads. I'm afraid—
GUILDENSTERN: So am I.
ROSENCRANTZ: I'm afraid it isn't your day.
GUILDENSTERN: I'm afraid it is.

[to be continued . . .]

(from *Rosencrantz And Guildenstern Are Dead*, Act One, by Tom Stoppard, 1967, Faber and Faber, London—printed here with considerable omissions and slight modifications (not indicated individually) by the authors of the present report)

Contents

1 Introduction

This is the first part of a two-report mini-series focussing on issues in the evaluation of annotations. In this first, theoretically-oriented part, we lay out the relevant statistical background, evaluate some pertaining approaches and also sketch some arguments that may lend themselves to the development of an original statistic. The second, more empirically-oriented part [Lücking and Stegmann, 2005], comprises a description of the project background including the annotation scheme at stake and the data collected, results from the practical application of the relevant statistics and, of course, the discussion of our respective results. Thus, in fact, the two reports come as a couple, like two sides of a coin. They have been separated in order to allow for a more linear discussion of the general theoretical issues (in this report) and the setting-specific application issues (in the other report), respectively. We believe that the points that are made in this part have significance on their own, reaching beyond the boundaries of the "DEIKON"[1] project. Generally, this document is the second one in a series of technical reports authored by linguists of the B3 "DEIKON" project. Taken in conjunction with its sister report, it continues the direction set forth by the first report in the series [Kühnlein and Stegmann, 2003] in its aim of discussing empirical issues with an eye towards the theoretical underpinnings, as well as the practical consequences of the findings obtained.

These publications are complemented by those of our project partners in computer science. Hence, both realms—and the interface between them—are covered: empirical, theoretical and computational aspects of the linguistic integration of speech and deictic gestures, as well as its (re-)synthesis and recognition by means of artificial intelligence methodology in a virtual reality setting [Kühnlein et al., 2003, Rieser, 2004, Kopp and Wachsmuth, 2004, Kranstedt et al., 2002, Kühnlein et al., 2004]. Last but not least, it should be noted that some of the statistical results that will be reported in the empirical counterpart to this report have been touched upon (among other statistical results) in a workshop contribution [Lücking et al., 2004].

Linguists working on dialogue data have to overcome at least two characteristics of human language in it's primary form: firstly, spoken language is an ephemeral phenomenon which exhales barely nascent, and secondly, natural language is, for the most part at least, not endowed with explicit structural and content markers (e. g. tags that indicate discourse segment boundaries or labels that name the performed dialogue moves).[2] To dispose of the first problem, natural language data are often conserved by using recording

[1] DEIKON is an acronym for the German project title "DEIxis in KONstruktionsdialogen".

[2] Though it is possible to include signals that carry meta-communicative content such as "Now I start a new discourse segment", such language use can be disregarded not only as being marginal, but also as highly artificial.

techniques, such as audio taping, video filming, or applying systems of manual transcription. This stage of empirical work comes with its own difficulties (potential problems include—but are not limited to—ensuring ecological validity, issues in the protection of the private spheres of the individuals involved, and the ever-present danger of overly theory-ladenness of observation), but we will not be concerned with these matters to a greater degree of detail here. Rather, our main topic is bound up with the second problem that has been mentioned above: Since linguists are often interested in language properties that are covert, respective data can't be merely recorded, they rather have to be "produced". A typical *data-generating source*[3] is the linguist involved, who acts as a *rater* of the bare recorded data. Thus, empirical work in linguistics often builds on the subjective judgements of the researchers themselves. One of the vehicles raters use to furnish their data, i. e. part of the *data-collection method*, is the *annotation scheme* (plus the instructions on how to make use of it). Such a scheme provides a "classification blueprint" which regulates how to classify the phenomena under observation with respect to predefined categories and rules of allocation. In order to indicate that data augmented in this way fulfill the usual scientific requirements, e. g. reproducibility, one has to take steps in order to assure the quality of the scheme. As Carletta [1996, p. 249] puts it from the perspective of dialogue research in computational linguistics:

> Now researchers are beginning to require evidence that people besides the authors themselves can understand and make the judgements underlying the research reliably. This is a reasonable requirement because if researchers can't even show that different people can agree about the judgements on which their research is based, then there is no chance of replicating the research results.

What this report addresses, then, is to introduce and to discuss some methods proposed for assessing reliability. The focus is on three agreement coefficients, namely *kappa*, *pi*, and AC_1. Their formalæ are presented and they are discussed in the light of their underlying heuristic and theoretic assumptions.

This report is organized as follows. Firstly, chapter 2 addresses the main theoretical considerations. We will be concerned with some prelimary background that proved to be helpful for a better understanding of what follows. Kinds of data are presented and types of agreement are distinguished. In what follows, appropriate statistics for determining reliability are segregated along the lines of *agreement* and *correlation* on a fundamental level. While the latter can be handled by "classical" correlation techniques, the former are settled in the field of inter-rater agreement statistics in a narrow sense.[4] To conduct such evaluation, we take up cudgels for the rather unknown and just recently introduced AC_1 coefficient [Gwet, 2001], since it has desirable mathematical properties. In opposition, the theoretical foundation of the popular *kappa* and *pi* statistics is put to scrutiny here. Finally, in chapter 3 we take down the essentials of the previous discussion in a more systematic way and try to approach an alternative foundation for an original agreement coefficient ω.

[3] By using the general terms "data-generating source" and "data-collecting method" we adopt the terminology of [Gwet, 2001].

[4] Here, "agreement" has to be read in a non-naive way, since respective statistics, unlike their correlation counterparts, are adjusted with respect to agreement-by-chance.

2 Theoretical Considerations

In this chapter we describe the statistical background of reliability studies. The chapter is cut into three sections: in the first one, we will focus on conceptual foundations and theoretical preliminaries which have to be presupposed in order to investigate the internal working of the different statistics in the second section. Then, in section number three, we deal with the issue of how to interpret our prospective results (to be reported in the next chapter). Note that the discussion of the agreement coefficients will be split where appropriate, since different kinds of statistics are appropriate for the different types of data measured.

2.1 Conceptual Foundations and Theoretical Preliminaries

There are certain conceptual and theoretical distinctions to be made with respect to our aims. These pertain to the specific criteria to be met, the types of measurement involved, and, finally, the types of statistical inferences applied.

First, as is very well known, there are certain standards that should be met in the context of scientific observation with *reliability* and *validity* figuring prominently among them. Loosely speaking, validity refers to the quality of scientific findings to represent phenomena that are real, i. e. findings that conform to facts. Krippendorff [1980] distinguishes several aspects of validity: as he notes, on the most fundamental level a distinction is sometimes made between *external* and *internal validity*. However, the latter is just another label for reliability, which will be covered in detail below. Further differentiation is possible with respect to the former aspect, that is validity proper: among such, Krippendorff distinguishes between data-oriented, pragmatic (or: product-oriented), and process-oriented kinds.[1] Even finer grained divisions are possible—compare Krippendorff's typology for the whole story. Since we will be more concerned with reliability here, we decide to leave out those details and start to elaborate on our central topic:

> "Fundamentally, reliability concerns the extent to which an experiment, test, or any measuring procedure yields the same result on repeated trials."
>
> (Carmines and Zeller [1979, p. 11])

In general, reliability deals with questions concerning the consistency of one's results. Opting for a more fine-grained approach, however, the different aspects of this issue are

[1] "Data-related validity assesses how well a method of analysis represents the information inherent in or associated with available data. [...] Pragmatic or product-oriented validity assesses how well a method 'works' under a variety of circumstances. [...] Process-oriented validity assesses the degree to which an analytical procedure models, mimics or functionally represents relations in the context of data." [Krippendorff, 1980, p. 157-158]

exhausted by the following list: stability, reproducibility, and accuracy. Again we follow Krippendorff [1980] in expounding what is involved respectively:

> "*Stability* is the degree to which a process is invariant or unchanging over time. [...] *Reproducibility* is the degree to which a process can be recreated under varying circumstances, at different locations, using different coders. [...] *Accuracy* is the degree to which a process functionally conforms to a known standard, or yields what it is designed to yield." (Krippendorff [1980, p. 130-131])

Table 2.1 summarizes the relevant information concerning reliability: as is obvious, the different perspectives are associated with specific test-designs for means of empirical investigations and each design focusses on a different kind of possible inconsistency. Furthermore, a look into the pertinent literature reveals that most reliability studies constrain themselves to one of the mentioned foci: reproducibility seems to get the most attention in this respect.

Table 2.1: [taken from (Krippendorff, 1980, p. 131) with modifications and omissions]

Type of Reliability	*Test Design*	*Focus of Error*
stability	test *vs.* retest	intra-observer
reproducibility	test *vs.* test	inter-observer
accuracy	test *vs.* standard	deviation from norm

To apply the above considerations to our main topic, i. e. the assessment of inter-rater agreement: since we do not have a god's eye viewpoint to compare our ratings to, there is no intention from our side nor a possibility of measuring validity. What we will be concerned with is estimating measures of reliability. Loosely speaking, however, such results may be seen to indicate some sort of an abductive measure for validity (but compare the more cautious remarks made below!). Concerning the different aspects of reliability, we will be mostly fiddling with reproducibility here, that is, measuring *inter-observer reliability* by using a test *vs.* test design. However, at least part of our annotation data have also been tested for *intra-observer reliability*, i. e. stability in Krippendorff's terms. Therefore, both raters had to re-do their respective classification tasks, which allows for comparison between the old and the new results with a test *vs.* retest design. If we had arbitrarily declared one of the ratings as a "gold standard" (all other ratings would have to live up to it) we could have tested for accuracy also. However, in the absence of compelling reasons that justify such an extra-ordinary position for one of our expert raters, we refrained from doing so.

In day-to-day scientific practice, the distinction between reliability and validity is blurred at times. This regularly leads to conclusions being drawn which are not justified in the light of the facts alone. In a similar vein, as mentioned above, we may feel tempted to speculate about the validity of our findings on grounds of abductive inference from reliability results. Strictly speaking, such a move is unwarranted, of course. There can

be no guarantees pertaining validity, even in the face of very good reliability results, since they may be due to various reasons, possibly unconnected to validity.[2] On the other hand, it seems to us that researchers who are unable to come up with reliable findings will be in a rather weak position, when it comes to the point of discussing the potential validity of their results. To summarize on the connection between reliability and validity: usually, reliability should be regarded as a necessary, but nevertheless insufficient condition for validity.

We shall turn to considerations concerning data scales and types of measurement now. As is very well-known, compare any introductory textbook on statistics, there are four types of metrics to be distinguished. We can have data on *nominal*, *ordinal*, *interval*, and/or *rational* scales (the latter two will collectively be referred to as *magnitude* scale levels below). The different scales are defined, among others, according to the meaningful intra-class relations that apply.[3]

Pertaining to the quality of measurements, a less well-known distinction can be made between *type-i* and *type-ii* measurements [Gwet, 2001]. Type-i measurements are those, where the process leading to the measurement is comparably well-understood and, hence, the result easily interpretable, whereas this is not the case for measurements being of type-ii. This distinction is probably best illustrated with an example. Take a doctor measuring a patient's blood pressure. The outcome displayed on the blood pressure gauge reflects the true level of the patient's blood pressure (or at least approximates it in a sufficient way). All other doctors (or persons familiar with hemodynamometry) will come, *ceteris paribus*, to the same result. Furthermore, there is a clear interpretation for the measured blood pressure value, namely in terms of the frequency of systolic against diastolic pressure. Such a measurement will be classified as being of type-i. Now contrast this kind of measurement with a classification task where, on the basis of, say, data from a psychological questionnaire, raters have to determine the satisfaction level of various subjects assigning them to the categories "happy", "stoic", or "sad". Such a measurement would be said to be of type-ii. The distinction between type-i and type-ii measurements manifests the degree of transparency concerning the applied methods of measurement, how clear-cut the conditions are in applying them, and whether there is an indisputable interpretation of the results obtained. Note, that there is some degree of affinity between type-ii measurements and data on nominal scales, as opposed to type-i measurements and data on magnitude scales. However, the correlation is not perfect.[4]

[2] As long as the raters act consistently, reliability results are guaranteed to be good. Of course, such observed consistency might be due to a valid scheme and an according rating process. However, consistency alone is not enough to guarantee those qualities. For example, the used categories might be inappropriate for the task at hand or the underlying theory might be badly wrong, but still allow for consistent application.

[3] The different scales are characterized by the following relations being defined among their members: nominal scale: $=$, $\neq$; ordinal scale: $=$, $\neq$, $<$, $>$; interval scale: $=$, $\neq$, $<$, $>$, $+$, $-$; and, finally, ratio scale: $=$, $\neq$, $<$, $>$, $+$, $-$, $\times$, $\div$.

[4] For example, think of type-i interval scale data that get transposed to a nominal scale representation. Of course, such a transformation would run at a loss of information, but is nevertheless possible to conduct. In a similar vein, although data resulting from a type-i measurement will usually be collected on magnitude scales, they may of course also be measured against a coarser scale.

Note also that type-i measurements can be handled within the framework of classical reliability theory [see e. g. Lord and Novick, 1968], whereas type-ii measurements cannot [Gwet, 2001].

A last dimension, which we shall discuss here, concerns statistical inference. *Analytical statistics*, unlike it's *descriptive* counterpart, is concerned with probabilistic-based generalizations, i. e. hypotheses-testing with respect to target populations of interest. Here we have to distinguish between *parametric* approaches and those which are based on a *non-parametric* foundation. Whereas the former rely on precise mathematical models of the target populations, e. g. a certain property is assumed to be found *binominally-* or *normally distributed* according to certain parameters, the latter methods are exclusively based on observable data. To overstate the facts somewhat: non-parametric methods are applicable in a distribution free manner [Sprent, 1989].[5] An example for a parametric test is the well-known *t-test* which assumes a variant of the normal distribution. Given a respective sample, the test may be exploited to decide the issue of consistency with respect to a *population mean* hypothesized in advance, since the related *t-distribution* allows the appointment of an appropriate *confidence interval* within which the sample value should be supposed to lie within a certain probability. However, one's data may be inappropriate with respect to presumed distributions, or one may want to make inferences that have nothing to do with the respective parameters.[6] In such situations, non-parametric methods are most welcome. Furthermore, non-parametric tests often come out as valid under weaker assumptions than parametric ones. Most of the statistical tests that we will come across below will be of the non-parametric kind (if not stated otherwise, the reader shall so assume).

The point that lies at the heart of introducing these distinctions is that different data types demand different statistical treatment. The applicability of a certain statistical procedure is constrained by both the scale niveau and the type of the subject data and may require specific parameters to be met. This is of concern to us in the evaluation of our coding scheme for multi-modal dialogue [see Lücking and Stegmann, 2005]: the classifications of dialogue move types and gesture functions set up type-ii data on nominal scale niveaus. In contrast, appointing the boundaries of words and gesture phases results in ratio-scaled data of type-i. Assessing reliability on these kinds of data has to be put into practice using different techniques: the partition runs between *correlation* (type-i, magnitude scale) and *agreement* (type-ii, nominal scale) statistics.[7] The for-

[5] "Many tests that are universally regarded as being non-parametric or distribution-free do involve parameters and distributions (often the familiar normal or binominal distributions). This is because the tags 'non-parametric' and 'distribution-free' apply *not* to the distribution of the test statistics, but to the fact that the methods can be applied to samples that come from populations having any of a wide class of distributions." [Sprent, 1989, p. 3]

[6] Examples of parameters are: *1.* mean and variance for normal distributions, and *2.* number of observances and probability of outcome for binominal distributions.

[7] A similar distinction is made by Rietveld and van Hout [1993]. Theirs runs between reliability and agreement statistics. Indeed, we think that our nomenclature is more to the point and less irritating, since we embed both correlation and agreement under the heading of "reliability", thereby reflecting the conceptual bifurcation which arises in the present context. Furthermore, our terminology seems to be in better accordance with the established language use among researchers.

mer implement traditional correlation coefficients which give a gauge of the degree of coherence between two measurement series; the latter express the extent of accordance between two ratings in terms of chance-corrected agreement coefficients. Those concepts and some of their appendant statistical tools will be dealt with in the following sections.

2.2 Measures of Inter-Rater Agreement on Nominal Scales

Linguistic data are often set on nominal scales, i. e. sets of independent categories with no other relations apart from equality and inequality holding among their members. Examples are grammatical categories, say case, person, number, part-of-speech, as well as judgements concerning acceptability or well-formedness. Pertaining to our annotation scheme [cf. Lücking and Stegmann, 2005, section 2.1] the ratings of gesture function, conversational move types, and categorical placement of the gestural stroke (among the linguistic tokens of the utterance) are of this type.[8] This section is split into two subsections. In the first one, we will approach the intuitions and the definitions concerning three agreement coefficients for data set on nominal scale niveau. This point will be discussed to some degree of detail, since our decision in favor of a certain statistic sets us apart from the current mainstream in the dialogue camp, cf. [Carletta, 1996, DiEugenio and Glass, 2004]. However, we think that there are good reasons for our choice and we will try to render them as transparent as possible. After all, we think that many researchers in the field may be struggling with similar difficulties, which makes this an important issue to address. It has to be noted that many of the points that will be made are indebted to the eye-opening discussions of Gwet [2001], whose influence on our presentation simply cannot be overestimated. However, we will also have some arguments and thought-experiments to add on our own. In the second subsection, we will be concerned with the somewhat delicate question of how to interpret the respective results.

2.2.1 The Agreement Coefficients kappa, pi, and AC_1

As has been noted by various researchers, a naïve agreement measure for data set on nominal scales is the bare proportion of agreement, i. e. the ratio of the number of identically categorized tokens against the total number of cases rated. However, apart from being vulnerable to the number of categories used, this measure does not take into account that some cases of agreement may only be due to chance, while others mark cases of "serious agreement". Observed agreement may, in fact, be due to different reasons which do not have to be connected to the goodness of the annotation scheme or the seriousness of the rating itself. We should therefore strive for a measure of agreement that is "corrected" with respect to such spurious agreements. In this vein, a definitive

[8] Whereas the former two are rather of type-ii, the latter category seems to mark data resulting from a type-i measurement. Therefore those data will, construed as collected on a different measurement scale, i. e. a magnitudal one, also be covered in the section on correlation statistics below. Indeed, the original data result from a magnitudal scheme and have been transposed to nominal scale niveau for reasons that will become obvious during the course of our presentation.

property of all the coefficients[9] that we will discuss in this section is that they incorporate a respective correction term, i. e. an estimate of that proportion of agreement that we assume to be merely due to chance. A common skeleton for calculating the different statistics can be recognized in terms of what we call *the golden formula of inter-rater agreement*. It looks as follows, compare [Carletta, 1996] on *kappa*:

$$P(A|\overline{C}) = \frac{P(A) - P(C)}{1 - P(C)} \tag{2.1}$$

In order to render this formula transparent: what we are interested in is the proportion[10] of agreement that is not conditioned by chance, i. e. the value that corresponds to the $P(A|\overline{C})$ term (here "$A|\overline{C}$" stands for the event of an agreement that is not condition to the event of an agreement by chance). In order to determine that amount, the $P(A)$ term on the right hand side of equation (2.1) introduces the proportion of actually observed agreement (hence, label "A" for agreement events). In opposition, the function of the $P(C)$ term which is the other main player in the above formula, is that of acting as a chance corrective (appropriately, we have label "C" representing events of agreement by chance) in introducing the share of chance-correlated agreement.[11] Therefore, the subtraction $P(A) - P(C)$ in the numerator gives a corrected measure of agreement which, however, is difficult to interpret on its own, due to being sensitive to the relative proportions of agreement and chance agreement. It has to be set into perspective against the overall proportion of cases (agreements and disagreements alike) where no chance agreement is involved, given by $1 - P(C)$ in the denominator.

Indeed, equation (2.1) displays the correct scheme, as can be shown by interpreting the symbols in terms of probabilities. From Bayes' rule it can be inferred that:

$$P(A) = P(A|C)P(C) + P(A|\overline{C})P(\overline{C}) \tag{2.2}$$

What equation (2.2) means is that the unconditional probability of agreement ($P(A)$) *equals* the sum of the conditional probability of agreement given agreement by chance ($P(A|C)$) times the probability of agreement by chance ($P(C)$) *plus* the conditional probability of agreement given no agreement by chance ($P(A|\overline{C})$) times the probability

[9] In principle, a fundamental distinction can be made between a *coefficient of agreement* and its *statistic*. While the former is an idealized theoretical construct, the latter can be used to calculate an estimate of the former against the background of actual empirical data. For the sake of easiness of our presentation, however, this distinction will be blurred in what follows.

[10] Since the probability of a possible experimental outcome e, denoted $P(e)$, is aproximated by its proportion of occurence in multiple trials of that experiment [cf. Gwet, 2001, p. 18-19], we feel free to switch from probability to proportion when this fosters accessibility of the exposure given here. There may be some affinity between probabilities and coefficients *vs.* proportions and statistics, compare the previous footnote.

[11] In many presentations a different nomenclature is chosen: $P(E)$ (with "E" for error-correlated agreement events) instead of $P(C)$. We decided in favor of the latter in order to stress the fact, that the fundamental equation of classical reliability theory, that is $x_i = t_i + e_i$ (measurement equals true score plus measurement error), is not applicable for data set on nominal scales. Furthermore, as Gwet notes, the concept of a "true score" may be ill-defined when dealing with type-ii data. We will have to say more about this at the end of the report in section 3.1.

of no agreement by chance ($P(\overline{C})$). Now, of course, event C implies event A, hence $P(A|C) = 1$. Furthermore, we know that $P(\overline{C})$ equals $1 - P(C)$. Therefore, equation (2.2) can be easily transformed and we end up with the form of equation (2.1) above.[12]

The resulting coefficient can be understood to form a chance-corrected and therefore truly useful measure of inter-rater agreement. However—since the "true" value of $P(C)$ is unknown to us—it should be obvious that the goodness of the whole procedure depends on the quality of the estimate of the $P(C)$ term. As we will see, the differences between the approaches that we discuss here reside exactly in the heuristics exploited in order to estimate that term.

In detail, we shall be concerned with the following statistics here: *pi* [Scott, 1955], *kappa* [Cohen, 1960], and AC_1 [Gwet, 2001]. For better or worse, those statistics that adhere to the general scheme of the golden formula (2.1) often get subsumed under the heading of kappa statistics [Carletta, 1996, DiEugenio and Glass, 2004]. This can be irritating at times: for example, the main presentation in [Siegel and Castellan, Jr., 1988], which has been cited as a canonical kappa reference more than once, e. g., [Carletta, 1996], is indeed based on a derivative of *pi*. Therefore, we feel obliged to provide some terminological clarity: it is necessary to distinguish between a narrow and a wide sense of the kappa label. While the former kappa should be taken to indicate usage of Cohen's *kappa* exclusively, the latter "kappa" shall be taken to encompass not only Cohen's statistics but also Scott's *pi* and generalizations of both. The reader may safely presume that when we make use of the word set in slanted letters (*kappa*), what we intend is the exclusive, the narrow sense just mentioned. On the other hand, when we aim for the wider, the encompassing sense, we will set the word in normal font surrounded by quotes ("kappa").

In order to provide the shortest possible history of the pertinent literature: the papers of Scott [1955] and Cohen [1960] introduce the original *pi* and *kappa* coefficients. They deal with the basic case of comparing two raters' classification of complete (= no missing) data. Later on, Fleiss [1971] developed a generalization of Scott's *pi* to account for several raters: that is, *generalized kappa* (irritating nomenclature, once again). Further, Cohen [1968] proposed a modified variant of his original *kappa*, i. e. *weighted kappa*, which is a second-order statistic, aiming to include the relative seriousness of certain disagreements.[13] The statistics of Krippendorff [1980], i. e. Krippendorff's *alpha*, has desirable properties concerning applicability (with respect to scale niveaus, number of raters, and missing data). However, it has been shown to be reducible to Cohen's *pi* at least for the special case of two raters and no missing data. Finally, the AC_1 statistics of Gwet [2001] marks an attempt to overcome certain well-known paradoxes of the other statistics. Gwet also develops a second-order variant of his coefficient, i. e. his AC_2 statistics in [Gwet, 2001].

[12] *1.* Substitute the respective expressions. *2.* Substract $P(C)$. *3.* Divide by $(1 - P(C))$.

[13] Think of categories A, B, and C on ordinal scale niveau and suppose that A outranks B (and C), B outranks C, and C outranks nothing. Then a disagreement between two raters involving categories A and C will be worse than a similiar disagreement concerning the categories A and B. This is true because the latter two categories are closer to each other according to the ranking function holding among the members of the scale.

Turning to the inner workings, it is most convenient to arrange the outcomes of the available ratings in form of a contingency table, as shown in Table 2.2 below. It depicts the way to arrange the data annotated by two raters, A and B, with respect to k possible response categories and n tokens rated. For all configurations, the obtained results are added and appear in the appropriate cells of the table, e. g. n_{21} represents the number of phenomena that have been cross-rated as category 1 by rater A and as category 2 by rater B.

Table 2.2: Contingency Table for Two Observers, k Response Categories and n Tokens

Observer B	Observer A				Total
	1	2	$\cdots$	k	
1	n_{11}	n_{12}	$\cdots$	n_{1k}	n_{1+}
2	n_{21}	n_{22}	$\cdots$	n_{2k}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	n_{k1}	n_{k2}	$\cdots$	n_{kk}	n_{k+}
Total	n_{+1}	n_{+2}	$\cdots$	n_{+k}	n

We will start with the formula for calculating Scott's *pi*, that is the K_π statistic. As has been remarked above, the general scheme of equation (2.1) applies. The "golden formula" is instantiated as follows to give the *pi* statistic:

$$K_\pi = \frac{P(A) - P(C_\pi)}{1 - P(C_\pi)} \tag{2.3}$$

The calculation of the $P(A)$ term will, of course, always proceed in the same easy way:

$$P(A) = \sum_{i=1}^{k} \frac{n_{ii}}{n} \tag{2.4}$$

This is very straightforward: we add the agreement proportions given by the values of the diagonal cells (the upper left to lower right ones) in Table 2.2, i. e. the n_{ii}, divided by the total number of cases n and sum across all categories. We turn to something more demanding now—the function for estimating the *pi* chance proportion:

$$P(C_\pi) = \sum_{i=1}^{k} \frac{\left(\frac{n_{i+}}{n} + \frac{n_{+i}}{n}\right)}{2} \cdot \frac{\left(\frac{n_{i+}}{n} + \frac{n_{+i}}{n}\right)}{2} \tag{2.5}$$

Here, for each response category, we determine the average of the presumed propensities of raters A and B to sort tokens into that category as construed along the actual proportions indicated by the marginals in Table 2.2, i. e. the n_{i+} and n_{+i}, against the total number of cases n respectively. We add both values and divide the sum by two, since we are interested in an average across two raters. Finally, this average propensity is

multiplied with itself in order to determine the joint probability representing two raters' chance agreement on a certain category. Furthermore, we determine the sum across all categories used. We comment on the rationale underlying this formula after inspecting Cohen's *kappa*, which is our next step.

By now, we are already well-acquainted with the general scheme of the "golden formula", which is instantiated as follows for the *kappa* statistic:

$$K_\kappa = \frac{P(A) - P(C_\kappa)}{1 - P(C_\kappa)} \tag{2.6}$$

$P(A)$ will be the same as it was for *pi*, hence formula (2.4) above remains correct, and we can omit this step here since the term can be simply substituted. As has been mentioned above, the point of divergence is the heuristics exploited in order to estimate the chance proportion $P(C_\kappa)$:

$$P(C_\kappa) = \sum_{i=1}^{k} \frac{n_{i+}}{n} \cdot \frac{n_{+i}}{n} \tag{2.7}$$

Since we have worked our way through the internals of the somewhat more complicated *pi* estimate above, this should be relatively easy to follow: instead of determining a joint "average" propensity underlying the chance-based ratings, we make direct use of the observed propensities for each single rater. This means, we multiply the observed categorical proportions of each rater in order to determine a joint probability, i. e. we divide the marginals (n_{i+} and n_{+i}) by n to give the presumed probability of the rater choosing a category and multiply it with the respective probability for the other rater. Again, we have to determine the sum across all response categories along these lines.

Now, that we have discussed the inner workings of *pi* and *kappa*, we come to evaluate them. For us, it seems that both *pi* and *kappa* implement their estimates—calculated in terms of summation of joint probabilities—in an all too direct way, i. e. without possibly "cleaning" the propensities involved or without "weighting" the overall result with regard to a reasonably expectable degree of non-chance based agreement. Indeed, with both *pi* and *kappa* the probabilities that determine the extent of chance-based agreement are directly based on those categorical preferences that have been observed, i. e. the bare actual rating proportions for the respective categories. Based on these, the joint probabilities are calculated and added without further ado. Proponents of *kappa* and *pi* may disagree on the exact procedure (determining one "average" propensity for both raters, or taking each rater's propensity as an own parameter), but apart from this detail, we think that the foundational strategy rests on a misled intuition.

Firstly, due to the way the calculation is set up the result for the chance estimates cannot be discerned stochastically from what would be an apt procedure for estimating the amount of chance-based agreement for exclusively chance-based ratings! What could the latter be? Imagine, say, a setting in which all ratings were based only on chance processes according to certain probabilities alone.[14] To us, however, it deems

[14] In principle, nothing prevents the raters from "implementing" all(!) their ratings along the lines of

to be absolutely implausible that all ratings should be of this kind, as seems to be presumed here. This is clearly not what we had in mind, when we said that we should be striving for a correction(!) pertaining to chance. What seems to be missing in the *pi* and *kappa* calculations is a moderating factor, e. g., an attenuation of the proposed proportion of chance-based agreements with an eye towards that proportion of ratings that can be reasonably assumed to be due to chance. Perhaps this is best illustrated with a "streamlined" all-to-easy example: Imagine two raters annotating subjects with respect to two response categories, both making equal use of the categories across the tokens, i. e. $\frac{n_{i+}}{n} = 0.5 = \frac{n_{+i}}{n}$. Furthermore, it shall be taken for granted that both raters' probabilities for determining the outcome of chance decisions come to 0.5 for each category (hence, they are truly random) and that we have further knowledge that exactly every fourth rating is due to chance for both of them simultaneously.[15] Then, a reasonable way to determine the amount of chance-based agreement would consist in the determination of the joint probability of agreement across the two categories for the chance-infected cases $(0.5 \times 0.5 + 0.5 \times 0.5)$ as is proposed by both *pi* and *kappa* heuristics, but additionally weighted, that is, multiplied with the proportion of chance-infected cases (0.25) as a moderating factor. This would leave us with the adequate result of 0.125, while both the *pi* and *kappa* heuristics come up with the unreasonably high result of 0.5 for the chance estimate here.[16] The latter would be a reasonable estimate if all ratings were due to chance, which, however, isn't the case for our example and, as we suspect, also for almost any real-world rating scenario. The point about what proportion of the ratings is likely due to chance and which is not is not taken into consideration by the *pi* or *kappa* heuristics. Indeed, it is a difficult question to answer (we will speculate about possible ways to deal with this point at the end of this report). Taking all of this into account, it seems that the *pi* and *kappa* "corrections" tend to distort the overall result of their statistics on systematic grounds.[17]

randomly determined processes as is implicitly presumed in the *kappa* and *pi* heuristics—e. g. deciding in favor of certain categories according to, say, tossing coins (two categories), throwing dice (e. g., six categories), or the drawing of cards from a set (e. g., 32 categories) all the time. Thereby, each "decision" for a category would be bound by the respective probabilities alone and, hence, also the agreements.

[15] Of course, our knowledge about the latter two propositions is artificial, the difficulty for real-world settings being that we are missing relevant knowledge about these parameters. Nevertheless, we have constructed this example in order to show something completely different: that a factor is needed in order to water down the overcompensated *kappa* and *pi* chance term estimates.

[16] Note that we have not made use of the actual amount of agreement $P(A)$ in our example. Our raters may agree on anything between the extremes of no cases or all of them. Furthermore, it is worthwile to recognize, that the value of chance agreement estimated by both *pi* and *kappa* may exceed the observed amount of agreement thereby leaving us with an overall negative result, which seems rather strange, to say the least. We will come upon another example for this below. With an adequately moderated estimate, however, such a paradoxical outcome would be much more unlikely, since the proportion of chance agreement would be estimated as smaller compared to the *pi* and *kappa* heuristics. This, however, in turn depends on appropriate heuristics for estimating the proportion of chance-infectedness. We will return to this topic at the end of this report.

[17] The result of the distortion, however, may be moderated from case to case, depending on the relative propensities involved as well as the overall level of agreement in the setting at hand. Furthermore, of course, it has to be noticed that the chance term gets subtracted in both the nominator and the denominator of the golden formula (2.1), thereby attenuating the effect of the usual over-estimation to

Our second point of criticism is connected to the role of those propensities that get exploited for the determination of the amount of chance-based agreement with both *pi* and *kappa* in their own modulated ways. We think that it is much too strong a demand to postulate that the probabilities that are involved in cases of spurious agreements must reflect the overall rating results. Imagine an arbitrary case of chance agreement: do the underlying probabilities accord to the overall trait preferences of the raters involved, i. e., are they shaped by the actual results of the raters' previous and forthcoming decisions? We think that the answer is not necessarily "yes" as seems to be presumed by proponents of both *pi* and *kappa*: to the contrary, it has to be remembered that the overwhelming majority of the collected data may well be due to ratings that have not been spurious. In the light of this, it seems that the connection between the underlying probabilities in chance-infected cases and the raters' decisions in serious cases is far from clear. Indeed, we might speculate whether it corresponds to the nature of chance-infected ratings to adhere to or, contrary to what is presumed by both *pi* and *kappa*, to deviate from the observed proportions.[18] At least, we seem to have reason to believe that the overall results do not necessarily reflect the probabilities involved in chance-infected cases. Straightforwardly, we have no measurement from which to infer to those entities in a safe and easy way. The probabilities may be fair (= equal = random, as in our above example), or they may be biased one way or another—we just don't know. In order to summarize our point here: we surely feel the pressing need for a very good argument purporting to show that the probabilities involved in chance-infected cases are framed by the overall trait proportions, i. e. that the chance probabilities reflect the absolute proportions that will regularly subsume spurious and serious tokens alike and which we come to know only *a posteriori*.[19]

It comes to our surprise that nobody seems to have critized *pi* and *kappa* on such grounds before—even Gwet's line of argument, cf. [Gwet, 2001], is somewhat different from ours up to now. However, it is only fair to note here that we have been inspired by Gwet's line of reasoning, compare the details of his analysis below.

In what follows immediately below, we will sketch some further aspects of an original perspective on these matters. Therefore, we shall put our question in terms of what kinds of results would be desirable in different circumstances, i. e., against the background of various imaginable rating scenarios ("ideal" *vs.* "lottery") that can be taken to correspond to different degrees of rater credibility/ability and distorting situative influences that may possibly arise during the rating process.

Firstly, it seems to us that in an *ideal* setting honest raters would follow their instructions[20] to the best of their knowledge in a neutral manner, make no errors in conduct,

some degree.

[18] This line of argument will be taken up in the final chapter of this report, where we sketch our ideas.

[19] Admittedly, taking a psychological stance, it may be possible to think of arguments or come up with empirical records that support a bias with respect to at least previous ratings, chance-infected and deterministic ones alike, e. g., along the lines of a recency bias. However, this does not seem to touch the general point of our criticism here.

[20] In the ideal world, we are presupposing that the instructions themselves do not give leeway for rating decisions, nor that they are stated in a plurivalent way.

and are not prone to disturbing eventualities. In addition, all the items that have to be rated exhibit clear criteria as predicted by the instructions in the coding scheme which manifest the coding scheme's underlying theory. In such an ideal setting we could rightfully make the strongest possible demand that $P(C)$ should equal 0 (regardless of the actual prevalences with respect to the categories): every case of agreement observed should be counted as being "serious". Secondly, as counterpart to ideal settings we can imagine *lottery*-like ones where the rater's decisions are made on purely random grounds, independent from the intentions underlying the scheme—compare elements of our discussion of the *kappa* and *pi* chance estimates just above. In such a setting, it would be desirable to demand that $P(C)$ should exactly equal $P(A)$, since all of the observed agreements have to be regarded as being due to chance (again: regardless of the prevalences observed).[21] Clearly, there is a lot of imaginable space in between such extremes. Therefore, what we are dealing with is a continuum of possible rating scenarios, ordered according to the seriousness of the rating and the benevolence of the rating situation. Of course, it seems all too obvious to us that the place of "ratings in the real world" will be somewhere in between such extremes on the continuum. Therefore, for the real world rating scenarios which we will deal with below, the $P(C)$ term should take on values between the lower bound of 0 and the effective upper bound of $P(A)$. Furthermore, taking into account that our very own raters are serious expert raters (as will often be the case in scientific scenarios, we suspect), we think that we are entitled to expect that our rating scenario will tend more towards the ideal rather than towards the lottery pole on the continuum. Hence, our respective chance estimators should rather tend towards the lower bound of 0 than towards the upper bound of $P(A)$.

Against this background, we shall fall back onto the *kappa* and *pi* statistics now and take a closer look at the mathematical properties of the *pi* and *kappa* chance estimates along the lines of the analysis provided by Gwet [2001]. Figure 2.1 depicts *pi* chance agreement as a function of two raters's classification proportions measured against one of two response categories.[22] Figure 2.2 does the same for *kappa*. As can be verified from inspecting the figures, the *pi* estimate varies between 0.5 and 1, while the *kappa* heuristic ranges between 0 and 1.

In the light of our above remarks, it comes as no surprise that we take the value pattern of the *pi* estimate to be speaking for itself: we take it as absolutely implausible to presume that the proportion of chance-correlated agreement should come to a minimum of 0.5 and a maximum of 1! This is an absurd result, since it implicates (among others) that we would always be tending towards the "lottery" pole.

Turning to the *kappa* estimate, here the range of the value pattern seems to be more in accord with our demands, since it varies between 0 and 1 as might be expected in

[21] In principle, $P(A)$ may grow up to the maximum of 1 for $k = 1$, even in a lottery setting. However, the practical upper bounds will be lower for $k > 1$ due to the various possibilities for the probabilities determining the outcome of the chance processes, i. e., raters in lottery settings will probably disagree on a lot of cases due to the nature of the chance processes involved. We will come back to this point below.

[22] Of course, with the proportions for one response category being determined, the proportions for the other category are determined too, since they are complementary.

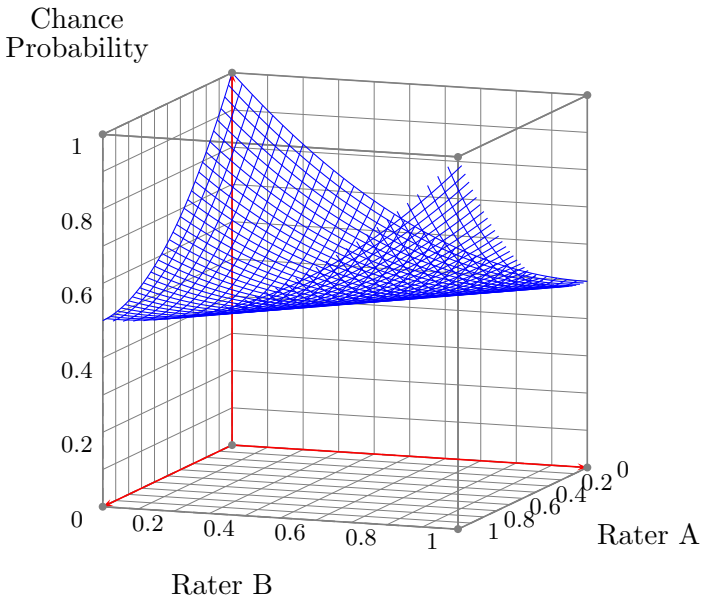

Figure 2.1: *Pi*: estimate of chance agreement as a function of raters' classification proportions

line with our demands above. However, it is illuminating to take a closer look and examine under what circumstances the *kappa* estimate becomes 1 or close to it. As obvious from figure 2.2 below: if both raters sort all or nearly all tokens into one and the same category, the chance proportion becomes 1 or close to 1, respectively. This, however, is not what we would expect! Rather, intuition seems to demand that if all or nearly all cases are classified as being of the same type, we would refrain from ascribing such an enormous amount of agreement to chance alone, cf. [Gwet, 2001]. A plausible hypothesis is that such a result might be due to a strong trait prevalence among the target population to be rated. Why should the raters disagree when most or all of the items to be rated are in fact of the same kind?

What is connected to these points is the fact that the *kappa* statistic has been criticized by several autors on grounds of delivering anti-intuitive results under certain configurations, the so called *kappa paradoxes*, cf. [Feinstein and Cicchetti, 1990, Gwet, 2002a]. Examining those paradoxes, it has been suggested that the *kappa* statistics may be susceptible to marginal inhomogeneities, trait prevalences, and/or bias [Gwet, 2002b, DiEugenio and Glass, 2004]. What is involved here? We know that for a given $P(A)$, the larger $P(C)$, the lower the overall result of *kappa*.[23] Now $P(C)$ and therefore *kappa* may change, even when the value of $P(A)$ remains constant. This happens, *ceteris*

[23] Of course, this holds for all statistics based on the "golden formula".

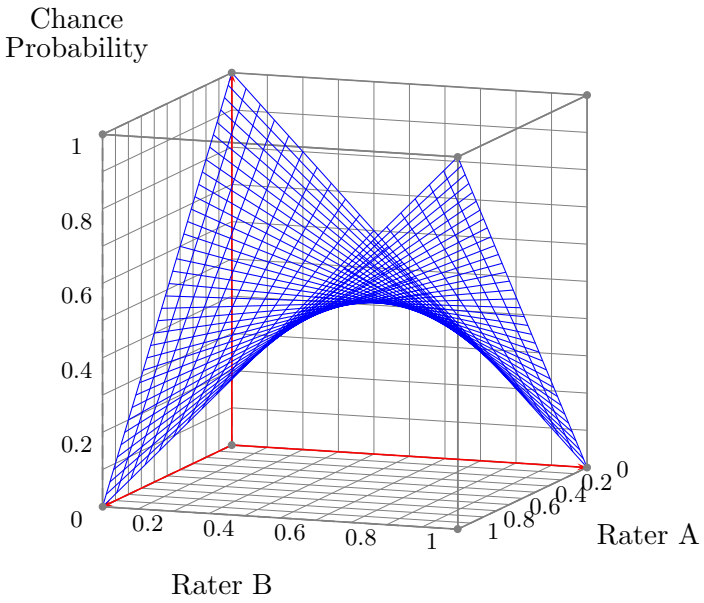

Figure 2.2: *Kappa*: estimate of chance agreement as a function of raters' classification proportions

paribus, when the underlying rating distribution is skewed. For an illustration, compare the contingency tables (a) and (b) displayed in table 2.3 below, an example taken from DiEugenio and Glass [2004].

Table 2.3: Contingency tables illustrating a balanced and a skewed distribution

(a) balanced distribution

Observer B	Observer A		Total
	1	2	
1	45	5	50
2	5	45	50
Total	50	50	100

(b) skewed distribution

Observer B	Observer A		Total
	1	2	
1	90	5	95
2	5	0	5
Total	95	5	100

The overall results are striking: the calculations for the data depicted in table 2.3(a) results in a *kappa* value of 0.80 ($P(A) = 0.9$ and $P(E) = 0.5$), while the respective calculations for the data displayed in table 2.3(b) leave us with the absurd *kappa* value of -0.0526 ($P(A) = 0.9$ and $P(E) = 0.905$).[24] This holds despite the fact that the

[24] Despite the manifest opinion of several practitioners in the field: negative results as such are uninter-

absolute numbers for the disagreements, as well as the proportion concerning the agreed-upon cases remains constant across both settings ($P(A) = 0.90$). However, due to the differing degrees of skewness, the chance estimate varies dramatically: it comes to 0.5 for table 2.3(a), but 0.905 for table 2.3(b). As hinted at above, such non-intuitive results are explicable in the light of Gwet's analysis of the error proportion as a function of the raters' classification proportions as depicted in figure 2.2 above. They are close to 1 for one of the observed categories in (b), thereby resulting in an unreasonably high error estimate that even surpasses the amount of overall agreement which, in turn, leads to a negative and therefore uninterpretable result.[25]

Now, for the mass of reasons that we have outlined above, we think that we are in need of a statistics that does better justice to our intuitions. As should be obvious, we believe that it will be desirable to strive for a coefficient where the amount of chance proportion estimated is smaller compared to the respective *pi* and *kappa* estimates. Furthermore, it seems to be a reasonable demand to remain beneath the upper bound of $\frac{1}{k}$ for k response categories, since such is the expected value that would be achieved by pure random ratings based on fair probability distributions for the chance processes involved. Here "fair" means that all possible outcomes will be equally likely to result from the chance-based process involved in cases of spurious ratings. Indeed, perhaps somewhat contrary to some of the intuitions informing our earlier thought experiments, it seems that such a fair probability distribution is often presupposed when we speak about events being random or due to chance, at least in the ordinary sense of the words. For the very least, this stipulation seems to mark a reasonable upper bound in the context of scientific ratings, where we should expect to tend towards the "ideal" rating pole.[26] Therefore, we now move on to discuss the AC_1 statistic [Gwet, 2001] which conforms to these requirements. We instantiate the "golden formula" as follows:

$$K_\gamma = \frac{P(A) - P(C_\gamma)}{1 - P(C_\gamma)} \tag{2.8}$$

As before, the right hand side of formula (2.4) above can be substituted for the agreement term $P(A)$ here. But the design of the formula used to estimate $P(C_\gamma)$ is very different, also intuitively, from those exploited for *pi* and *kappa* above, compare:

pretable! The point is best understood in analogy to probability results: due to the axioms of probability theory there is no such thing as a negative probability, hence it cannot be interpreted. The same is true for negative *pi* and *kappa* results due to the close connection between proportions and probabilities on the one side and statistics and coefficients on the other. Furthermore, we wonder how DiEugenio and Glass [2004] can come up with examples as the one that we have cited, but still recommend the combined usage of *kappa* and *pi* even in these cases.

[25] There are further issues which we can not deal with in detail here: for example, both $P(C)$ and $P(A)$ can become exactly 1. Under such circumstances, the 'golden formula' yields no result—we would have to divide by 0 (the result in the denominator), hence the overall result is undefined and therefore, again, uninterpretable. Furthermore, there are issues in the applicability of the variance estimators for both *pi* and *kappa*, which can become a drawback for significance testing. See [Gwet, 2001] for the details.

[26] Gwet tries to motivate his demand of "doing better than random" probabilistically, but finally stipulates this restriction on "empirical grounds" without further qualification. He regards it as a foundational truism and it is also his main point of critique concerning the *pi* and *kappa* chance estimators, since both allow for higher chance estimates than $1/k$.

$$P(C_\gamma) = \frac{1}{k-1} \sum_{i=1}^{k} \gamma_i (1 - \gamma_i), \tag{2.9}$$

with γ_i being calculated as follows:

$$\gamma_i = \frac{\frac{n_{i+}}{n} + \frac{n_{+i}}{n}}{2} \tag{2.10}$$

Gwet presumes that agreement by chance comprises the following events: agreement and at least one of the ratings being *random* as opposed to *deterministic* ones, which are taken to be the constituents for agreements not conditioned by chance. Now, it is further presumed that the probability of random ratings is related to the uncertainty of raters' judgements, which in turn is explicated by making use of the technical concept of the *intra-observer variation* [Kjærsgaard-Andersen et al., 1988]: it is estimated as the ratio of the observed intra-observer variance to the maximum possible intra-observer variance. This latter ratio is referred to as the *coefficient of uncertainty*. Finally, the overall chance agreement probability is estimated as the ratio of the coefficient of uncertainty against the total number of response categories. The resulting equations are transformed and we end up with equation (2.9) above.

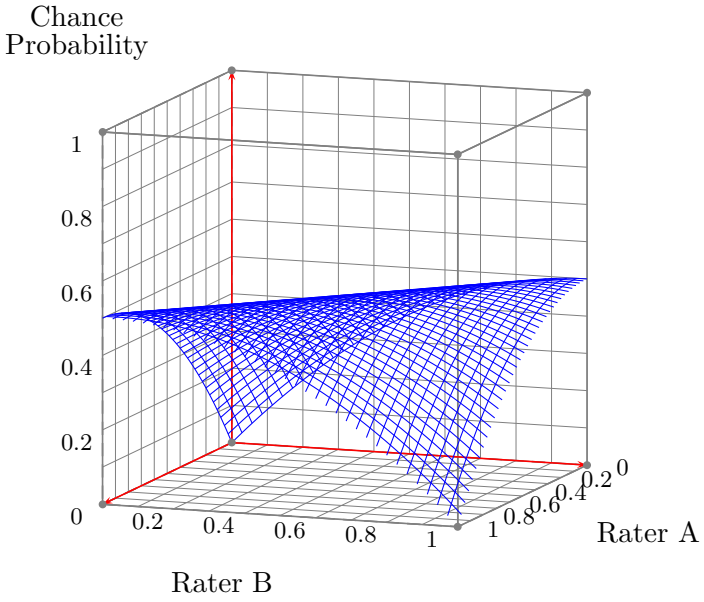

Figure 2.3: AC_1: estimate of chance agreement as a function of raters' classification proportions

Gwet's way of estimating the probability of chance agreement has desirable mathematical properties, compare figure 2.3. It is set up along the same lines as the figures 2.1 and 2.2 above in order to allow for direct comparison. The range of the estimated chance

proportion is as desired, with a minimum value of 0.0 and a maximum of 0.5 for two response categories ($k = 2$). The AC_1 statistics is not subject to *kappa* paradoxes, compare the configurations under which the chance proportion reaches its maxima. There are further advantages that have to do with the way in which variance estimations are set up and applicable, which is important with regard to tests for statistical significance. Another advantage is that Gwet's approach is explicit concerning it's premises: all foundational concepts are well-defined and set up in a systematic way. In opposition, there has been some confusion in the literature on *kappa* and *pi*, especially on how to interpret the obtained results. We shall turn to some of these issues in the following section on interpretations.

However, we think that a caveat is necessary concerning the quality of the AC_1 estimate. The attentive reader will have noticed that finally the amount of chance proportion estimated by AC_1 boils down to a function of the rating proportions for the different categories as depicted in figure 2.3 above. We have criticized *kappa* and *pi* on related grounds, and, of course, AC_1 inherits this possible source of misconception. However, we think that with AC_1 the ascribed amount of chance is measured in a more moderate and therefore more reasonable way. Also, with AC_1 the function is mediated by a prominent factor in the estimation, i. e., the intra-observer variance measuring the dispersion of the individual raters' ratings. In the absence of parameters that could be exploited in order to estimate "true probabilities" involved in chance agreements or the "true share" of chance-infected ratings, we have to make use of the available information in one way or another in order to arrive at a result (we are left with nothing else). What is important with regard to our current aims is that results obtained with AC_1 are reasonably robust and the statistics as such is more appropriate for the setting where it is put to use here. Therefore, we decide in favor of AC_1 on pragmatic grounds and in the absence of a better alternative.

To summarize our perspective concerning agreement statistics on nominal scales: all of the discussed statistics comprise a component to correct for agreement by chance. However, it seems that both *pi* and *kappa* over-estimate the amount of chance-correlated agreement to a significant degree. Thus, the respective overall results are distorted on systematic grounds. Various arguments purporting these points have been raised and the issue has been exemplified by a rating configuration that leads to a strikingly paradox result, thereby indicating further difficulties for the interpretation of the respective statistics. From our perspective *pi* does worst, *kappa* does only little better, but the rather unknown AC_1 statistics delivers moderately adequate estimates. However, none of the statistics can claim to incorporate a "true amount" of chance correction.[27] We have been at pains concerning our arguments here, since, as should be obvious now, the decision in favor of a certain statistics may well have massive implications regarding the results of evaluation efforts as ours! We also think that details of our arguments may be of interest to many researchers who are using one of the mentioned statistics.

[27] Indeed, Gwet might argue that there is no such amount in an absolute sense, since he believes that the concept of a "true score" is generally ill-defined for data resulting from type-ii measurements.

2.2.2 On the Interpretation of the Agreement Coefficients

The application of the introduced statistics results in final values that span the range between negative numbers and the upper bound of 1.0 for all statistics. The lower bound varies depending on the statistics used: for AC_1 it comes to -1.0, while it reaches down to -100.0 for *kappa* and even further down to -200.0 for *pi*.[28] However, it should be noted that values near these lower bounds are seldom, if ever, reached in practice. Now, our question of interest here concerns the significance that can be attributed to such values. As there seems to exist considerable confusion on how to interpret the results of inter-rater agreement statistics, at least among practitioners in the dialogue field, compare, e. g., [DiEugenio and Glass, 2004], we shall discuss our perspective on these matters here.

There are two possible ways to approach this task. One option consists in the interpretation of the result as an estimate for a coefficient that is subject to a sampling variability. This implies a test for statistical significance against a null hypothesis of interest.[29] Such an approach belongs to analytical statistics or statistical inference. Using these techniques, it is possible to come to judgements concerning statements about target populations that reach beyond those subjects and raters that figure in the actual empirical data, i. e., one tries to project beyond the current sample. A completely different approach consists in the direct interpretation of the bare results. Such a perspective can be subsumed under the heading of descriptive statistics, since here the agreement coefficient is construed as an absolute measure constrained to the collected data alone. This, however, implies that we can not generalize beyond our current population of raters and subjects. An approach of this kind can be implemented *via* the application of conventional quality scales or by means of internal ranking according to the relative goodness of the calculated values.

We will begin our discussion with an examination of the approach mentioned second here, i. e., the direct interpretation of the values resulting from the application of *kappa*, *pi*, and AC_1 statistics. In general, it seems that positive values are comparably easy to interpret, since what we set out to determine is the proportion or the probability of agreement not conditioned by chance—compare the description of the "golden formula" in the preceding section. Hence, a positive value of 1.0 indicates that the raters agree on

[28] The interested reader may want to verify our statements here. We recommend the use of a plotting program, e. g. gnuplot, to plot our target function, i. e. the golden formula: $(a - c)/(1 - c)$ (we make use of streamlined symbols here, for the sake of being able to state the function in the program). The range of the agreement proportion a has to be set between 0.0 to 1.0, since such are the limits for the proportion of observed agreement. Of course, the range of the chance estimate c is dependent on the respective statistic. For two raters and two response categories we get the following ranges, compare our explanations in the preceding section of this report: for AC_1 the range is set between 0.0 and 0.5, the *kappa* range goes from 0.0 to 1.0, and *pi* has it from 0.5 to 1.0. These different ranges (due to the differing value pattern of the heuristics for the error estimates) lead to the different overall lower bounds of -1, -100, and -200 for the respective statistics.

[29] The null hypothesis is the negation of the hypothesis we are actually trying to corroborate. Basically, a statistical significance test can be conceived of metaphorically as a proof by contradiction on probabilistic grounds, since we try to establish that our actual result is highly implausible against the assumption of the null hypothesis holding.

all tokens and that none of this is due to chance. In opposition, a value of 0.0 indicates that all actual agreements are conditioned by chance or, possibly, that no agreement can be observed among the raters (and none is expected on grounds of chance). The positive values in between can be interpreted along such lines accordingly, we will be concerned with the details below. Turning to negative results, as we have been at pains to show in the preceding section, we are convinced that both *kappa* and *pi* are infected with serious weaknesses. As a direct consequence, their application will result more often than good in negative values, even for relatively large samples. The results for the data arranged in Table 2.3(b) above mark a fine example. Results as such, i. e., negative values, are simply uninterpretable as is demanded explicitly in the context of the definitions grounding the AC_1 statistics.[30] The latter is a reasonable demand, since there is no such thing as a negative probability, because it is excluded by the axioms of probability theory. In principle, of course, we can imagine what such a result is often taken to indicate: even less actual agreement than what should be expected on grounds of chance. Results as such, however, mark paradox findings that must be taken to indicate that something has gone wrong—perhaps a mistaken calculation, but it might also be due to a failure concerning the foundation of the statistics (as seems to be plausible for *kappa* and *pi*) or we might be dealing with an overly small sample size that does not allow for the projection of a reasonable chance estimate (in case of using AC_1). Since *kappa* and *pi* are weakness-infected, our general advice is to forbear from interpreting their results. Rather, we suggest to repeat calculations on affected data using the AC_1 statistic.

What remains to be done here is to explain our stance on results that span the ground between the extremes of 0 and 1. Many attempts to come to grips with this involve the application of a certain quality scale. A look into the pertinent literature reveals a variety of such ways of categorization that have been devised by "kappa" proponents mostly. We shall summarize two of them here. The suggestions of Krippendorff [Krippendorff, 1980] are among the hardest to be met—he demands that all results worse than 0.67 should be discounted. Definitive conclusions shall only be appropriate in the light of results better than 0.80, while the "middle-ground" of values that span the range between 0.67 and 0.80 shall allow for tentative conclusions to be drawn. In opposition, the suggestions in [Rietveld and van Hout, 1993], are clearly more permissive: results between 0.00 and 0.20 are categorized as "slight", those between 0.21 and 0.40 are counted as "fair". Further, results between 0.41 and 0.60 are seen as "moderate", while values between 0.61 and 0.80 are called "substantial" and, finally, those above 0.81 are regarded as "almost perfect". A full list of all proposals regarding what may be labelled as "best practice heuristics for estimating the goodness of one's results" would be lengthy—suffice it to say that they span the range between the cited extremes. However, all of them are introduced on clearly arbitrary grounds, which is a fact that is also admitted by their proponents.

[30] Indeed, the calculation of the AC_1 statistics may result in negative values down to -1 for samples of small size. However, due to the way the theory is set up such results are excluded from any attempt of interpretation. Generally, with AC_1 negative results occur much more seldom and only for smaller samples as compared to the *kappa* and *pi* statistics. This property is, of course, due to the make-up of the chance estimates—for AC_1 it comes out as generally smaller and therefore more probably less than the agreement proportion, thereby tending to produce a positive overall result.

¿From our perspective, the diversity of proposals may in part be due to the diverging properties of the various statistics. For example, the observation that all-to-high chance estimates lead to comparably lower and sometimes even negative overall results may help to shed some light on why surprisingly weak standards are accepted by at least some researchers who pursue the usage of "kappa" statistics, compare, e. g., Rietveld and van Hout's rather permissive suggestions cited above. In this regard, we think that it will be imperative to apply stricter scales when using AC_1, since that statistics has been designed to estimate the chance term in a generally more moderate way, and therefore it can be expected to deliver higher overall values for the vast majority of configurations.[31] Nevertheless, it has to be stressed that the decision in favor of a certain quality scale is bound to be somewhat arbitrary anyway. This latter point holds, since it will usually be the context of the investigation that dictates the level of goodness one may be willing to accept. What we are dealing with regarding such considerations is a trade-off between the importance of coming to right decisions on the one hand, and pressure to come to any decision at all, i. e. issues in the applicability of such schemes, on the other hand.[32]

Due to the inherent uncertainties with regard to chance estimators and further sampling variability, it may seem natural to fall back on significance tests exclusively, our first option mentioned above. This strategy also marks a prerequisite if we want to generalize from our current rating sample to a larger, possibly infinitely large target population of interest.[33] The usual parametrical technique involves the comparison of a test statistics (which is a function of the agreement coefficient obtained and the underlying variance estimated) with an appropriate critical value, compare any introductory textbook. If the test statistics comes out larger than the critical value, then the null hypothesis can be rejected, otherwise it cannot. However, rejection of the null hypothesis may still happen erroneously: respective errors are known as errors of the first kind. In statistical practice, one tries to minimize the danger of committing such an error. Therefore, often upper bounds of 1 %, 5 %, and/or 10 % are accepted for the corresponding probability. Indeed, we will make use of significance tests below, however, we will not implement a direct parametrical approach, since it presupposes that the test statistics approximates a normal distribution with mean = 0 and variance = 1 (or can be transformed into a normal distribution by means of a z-transformation). As Gwet [Gwet, 2001] notes, this is an inadequate assumption even for large samples. He proposes a non-parametric alternative that functions *via* critical values determined by Monte Carlo simulations.[34]

[31] For skewed distributions, AC_1 values will be much higher than their *kappa* counterparts. However, there are also some configurations where AC_1 will, in fact, deliver smaller overall results as compared to its *kappa* counterparts.

[32] For example, in medical contexts it will usually be imperative to ensure rather high levels of reliability for the sake of the well-being of the patients.

[33] However, we do not have to generalize, if we do not want to or do not have to do so. If all relevant tokens of interest have been rated by the relevant raters, we may rest happy with a final result for a closed domain.

[34] "A Monte Carlo simulation study aims to mimic the random assignment of n subjects into Q categories by r raters and to observe the distribution of the agreement coefficient estimator as well as that of the test statistics. [...] A large number of such simulated values will provide a good approximation of the sampling distribution." [Gwet, 2001, p. 158]

[Gwet, 2002b] contains tables that give the critical values for different amounts of rater and subject samples, as well as different numbers of response categories. So, due to the foundation on Monte Carlo studies, we can determine our AC_1 significance results without resorting to the calculation of a function of the respective variances.

However, it has to be taken into account that the degree of significance that can be attached to a significance test is, of course, inherently connected to the explicit content of the null hypothesis against which it is performed. The usual practice consists in a test against the null hypothesis of "no agreement but chance agreement", i. e., $AC_1 = 0$. However, it is easy to see that comparably little is shown on grounds of a significant result against that proposition.[35] No one should rest assured by the fact, that raters agree with a reliability that is qualified as "distinctly better than chance", since no statement about the degree of "distinctly better" is implied. So one might want to test against further "watermark" null hypotheses, say, e. g., $AC_1 = 0.1$, $AC_1 = 0.2$, ... , $AC_1 = 0.9$. But then, again, we face the old question of what level of distinctiveness can be seen to be constitutive for an appropriate result—as it seems we cannot escape the trap of arbitrariness in principle.

With an eye towards the interpretation of real-world data in the sequel to this report [Lücking and Stegmann, 2005], we decide that we will perform significance tests against the null hypothesis of "null agreement apart from chance agreement", adhering to the usual scientific requirements. However, this approach will be flanked by means of interpretation on direct terms, i.e. for the data as a closed domain, but rather on intuitive grounds. We do not recommend the "strict" application of a certain arbitrary scale, since viewing such constraints as hard ones often does more harm than good. Nevertheless, of course, the following holds and will be taken into account: the better an AC_1 result, the bigger our confidence in that result can be. We shall also feel free to compare AC_1 results against each other. Such internal ranking may provide us with hints on where to discover leaks with regard to reliability matters (at least for the present raters and the subjects rated). However, in interpreting our results, we will also take into account that good reliability results need not be conditioned by intrinsic qualities of the scheme applied alone. That is, in the light of our reliability results, we will speculate about how they can be explained and we will thereby touch upon issues that pertain to validity proper.

2.3 Measures of Correlation on Magnitude Scales

Unlike the classification of observed elements into nominal response categories, the measurement of temporal units, as is the case in appointing word and gesture boundaries, leaves us with a different kind of data. The scale underlying those time marks is that of a rational type. In comparing two (or more) sets of scores on magnitude scale niveau, what we are interested in is whether there is a systematic relationship between those sets. And

[35] Further information is provided by the respective probability p for committing an error of the first kind (rejection of the null hypothesis when it is, in fact, true). Of course, the lower the p, the stronger our confidence can be.

if there is, we want to know how strong they are correlated. Since measurements of this type are so well-known, we only set out a general description. In a first subsection, the three statistical headliners are introduced. The following subsection, which deals with the question of how to deal with the results obtained by applying the correlation coefficients, is restricted to issues which have direct concern to our present cause. Readers interested in more details are referred to any introductory book on statistics.

2.3.1 The Correlation Coefficients r, rho and tau

A first approach to get a value for the relationship of two series of measurements would be to calculate whether they change in a systematic way. This can be carried out, roughly, by summing the products of deviates from the arithmetic mean, divided through the number of cases—often, as here, decreased by one. This method is known as *covariance d*. Applied to two series of measurement X and Y it computes as follows:[36]

$$d_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \tag{2.11}$$

Note that the value of covariance will be negative if, graphically speaking, the corresponding scatter diagram reaches from top-left to bottom-right. It is thus an indicator for the direction of correlation between series of measurements. The value of covariance can be arbitrarily high, since it is sensitive to the amount of the measurement values. Thus, it gives us no hint as to the *strength* of correlation. Since we are interested in that, we have to employ a further technique.

The strength of correlation between two magnitude-scaled sets of data can be calculated using Bravais' and Pearson's coefficient of correlation, generally referred to as the *Pearson product-moment correlation coefficient r*. This coefficient is standardized over the covariance of measurements. It can take a value between -1 and 1, indicating both the direction and the strength of correlation.

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}} \tag{2.12}$$

The Pearson product-moment correlation coefficient is not a distribution-free method. It is associated with measurements which are linearly related to each other. This is the case if, as the measurement series X increases or decreases, the measurement series Y tends to increase or decrease in a similar way. Since we are dealing with scales on a time-bar, we have a clear match to the precondition of a bivariate normal distribution.

We would like to stress, that this coefficient is *not* designed to model the reliability or agreement between raters with respect to response categories. It is a denominator for

[36] Note: the formulae given in this section are empirical estimators, not the model formulae. So, for example, the variable depicting the true mean is instantiated with the arithmetic mean of the actual measurement series.

the *correlation* of measurements, ignoring the possibility of chance and the sampling of raters. Despite the phonetic hassles that arise in determining the boundaries of words (and likewise for gesture phase boundaries), we appoint this kind of measurement as being of type-i. Though classical reliability theory [Lord and Novick, 1968] offers chance-corrected methods even for magnitude scaled data, we refrain from employing them here and use correlation techniques that seem appropriate in handling type-i data.[37]

There are two well-known non-parametric statistical methods appropriate for assessing correlation between at least ordinal-scaled sets of data, independent of their distribution. The first of them, *Spearman's* ρ (*rho*), is just the Pearson product-moment correlation coefficient calculated for ranks; the second is known as *Kendall's* τ (*tau*) and has no obvious parametric counterpart. Like Spearman's ρ, Kendall's τ is computed for ranks. Thus, both are associated with ordinal-scaled data. All three correlation coefficients, r, ρ, and τ, have the common property that their resulting values lie between -1 and 1. Since Spearman's ρ and Kendall's τ are associated with ranks, one has to transpose magnitude scales into rankings in order to apply them to interval- or ratio-scaled measurements.

We obtain Spearman's rank correlation coefficient ρ by substituting in Pearson's product-moment correlation coefficient the values for $x_i, \overline{x}, y_i$ and $\overline{y}$ with the corresponding rank values $r_i, \overline{r}, s_i$ and $\overline{s}$:

$$\rho_{r,s} = \frac{\sum_{i=1}^{n} (r_i - \overline{r})(s_i - \overline{s})}{\sqrt{\sum_{i=1}^{n} (r_i - \overline{r})^2 \sum_{i=1}^{n} (s_i - \overline{s})^2}} \tag{2.13}$$

Although Pearson's r and Spearman's ρ are carried out in an analogous way, there is a petite difference in their power-efficiency . The efficiency of ρ when compared with r is about 91% [Siegel, 1956, p. 213]. This means that for ρ to reach the same power – that is, the probability of rejecting H_0 when it is in fact false – as r, ρ has to be carried out on a population which is 9% larger than that of r.

Kendall's τ is said to be a *concordance coefficient*. It is proposed to have very similar properties as Spearman's ρ, but rests on a different logical basis. Kendall takes a direct approach to a certain tendency in correlated data, namely that high values in one scale are associated with high values in the other one. To put it another way, when we have two pairs of observations $< x_i, y_i >$ and $< x_j, y_j >$, then if $x_j > x_i$, it is very probable that $y_j > y_i$. If this is the case, the pairs of observations are said to be *concordant*. If

[37] Some readers might want to object that word and gesture boundaries are not clear-cut and unique so that there is some space for variation. This in turn makes it possible that two people could "guess" where exaclty to mark the boundary in question, but nonetheless could agree in their guessing. Following this line of thought, it seems to be necessary to apply chance-adjustement for respective gauges of agreement. We meet this objection by stating that the theories underlying the observable phenomena, namely a theory of communicative body motion and phonology, *incorporate* a certain degree of fuzziness themselves. For example, in phonology there is a frequent use of thresholds, e. g., between voiced and voiceless sounds. Thus theory itself predicts a gray area. This is not the kind of uncertainty that would justify employing a chance-corrected measure for agreement. Chance-correction is required if the potential error lies *outside* the theory.

this is not the case, we can face two different situations: (1) $x_i = x_j$ and/or $y_i = y_j$; this is called a *tie*. (2) $x_j > x_i$ and $y_j < y_i$; then the pairs are described as *discordant*. Kendall's τ is determined by the following fomula, where n_c denotes the number of concordant pairs and n_d the number of discordant ones.

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{2.14}$$

The power-efficiency of τ is equal to that of ρ [Siegel, 1956, p. 223]. Thus, compared to the most powerful parametric method, the product-moment coefficient r, both τ and ρ score at the efficiency level of 91%.

2.3.2 On the Interpretation of the Correlation Coefficients

What does the outcome of a correlation coefficient really mean in terms of the strength of the relationship between the sets of data? The only clear outcomes are those where the value of the coefficient $v(c)$ equals -1, 0, or 1, meaning that there is a perfect negative, absolutely no, or a perfect positive correlation, respectively. But what confident information about the association of measurements series can be drawn off if $v(c) \neq 0$ and $-1 < v(c) < 1$?

In the case of Spearman's ρ and Kendall's τ it is possible to calculate critical values[38], against which the actual value of rank correlation has to be compared. In his appendix, Sprent [1989, table A9, table A10] lists the pertainig critical values. If $v(\rho)$ denotes the value obtained for the Spearman test for a given pair of rankings, and $t(\rho)$ denotes the appropriate critical value as fixed in the table, then if $v(\rho) \geq t(\rho)$ the correlation can be considered as significant, meaning that there is a positive association between the two sets of rankings. The critical value for Kendall's τ is determined via the amount of concordant pairs. Let $n_c - n_d$ be the actual result of the Kendall test, and $t(\tau)$ the respective value in the table. If $n_c - n_d \geq t(\tau)$, then the association between the subject ranks counts as significant.

The product-moment correlation coefficient also does not indicate a proportion. That is, an outcome $r = 0.5$ does not mean a fifty percent relationship between the measurements in question. But one can interpret the square of r, r^2, properly as the proportion of the total variance of the measurement series Y accounted for by its linear relationship to measurement series X [cf. Lindemann et al., 1980, p. 57]. The outcome $r = 0.5$ would mean that one fourth ($r^2 = 0.25$) of the variance of Y could be explained by its linear relationship to X.

With the Pearson product-moment correlation coefficient r we can take advantage of the underlying magnitude scale to infer more detailed information. Under the assumption of linearity, we can describe the regression line of two measurement series X and Y with a function of the following form:

[38] The key for getting at those critical values lies in calculating the occurrence probabilities of correlation values. [Siegel and Castellan, Jr., 1988] discusses the test of significance for each non-parametric "measure of association".

$$y = bx + a \tag{2.15}$$

Given the values of the standard deviation σ^2, the arithmetic mean for each measurement series, and the Pearson product-moment correlation coefficient r, we can easily calculate the values for the slope and the intercept of the regression line function:

$$b = r_{xy}\frac{\sigma_y^2}{\sigma_x^2} \tag{2.16}$$

$$a = \overline{y} - r_{xy}\frac{\sigma_y^2}{\sigma_x^2}\overline{x} \tag{2.17}$$

Nota bene: In carrying out the calculation of simple linear regression, one measurement series is reduced onto the other. Using the formulas given above, measurement series X is set as basic. The values for Y are derived on the values of X. This asymmetry has to be carefully observed when we are concerned with an independent and a dependent variable. But in our case no rating can be presumed to be basic (or classified in terms of dependence). Thus, we could also calculate the reversed regression function, depicted by $x = by + a$ and the formulas to compute slope and intercept modified to $b = r_{yx}\frac{\sigma_x^2}{\sigma_y^2}$ and $a = \overline{x} - r_{yx}\frac{\sigma_x^2}{\sigma_y^2}\overline{y}$. To keep things more handy, we only deal with the scheme given in (2.15) and just insist that we could have implemented things the other way round, or both.

Having defined the function for linear regression, we receive a pattern of the degree of one measurement series changing in dependence of the other.[39] In the case of the scores for boundaries of words and gesture phases, we can interpret the scores with respect to tendencies one rater might have in setting the boundaries earlier or later than the other.

[39] McNemar [1969, chapter 9] lists six methods for interpreting the product-moment correlation coefficient, namely in terms of (1) the rates of change (this is what we apply here), (2) the standard error of estimation and the coefficient of alienation, (3) variance, (4) the proportion of common elements, (5) the normal correlation surface, and (6) "success" expectancies. The readers interested in methods (2) to (6) are referred to McNemar's book.

3 Reliability Reloaded

In this last chapter of the present report, we take up some advanced issues in reliability and discuss them against the background of the other parts of this report. Firstly, in section 3.1, we highlight differences between our stance and that of Gwet in more detail than before. This also leads to a discussion of issues related to validity. In the following section 3.2, we sketch some ideas that may lend themselves to the construction of an original statistic. These points, however, can not be worked out to a satisfactory degree of detail here and our respective ideas must be judged as speculative and remaining "under construction" for now.

3.1 Advanced Issues in Reliability and Validity

Early suspicions that something might be wrong with the well-established statistics were nurtured when we came across the analysis of the mathematical properties of the *kappa* and *pi* chance estimators in [Gwet, 2001]. However, we have argued mostly from our viewpoint in section 2.2.1 above, while some of Gwet's points have been sacrificed for the sake of the presentation. Now, we will try to render more transparent some differences between our stance above and the details of his perspective.

Going beyond what is presented in the pertinent literature, especially in the most advanced theory given in [Gwet, 2001], we have to dwell on issues concerning random ratings, validity, true scores, vague objects and independent theories for certain subject-domains. Before we do that, however, we shall have a closer look at the details of Gwet's approach. On a very fundamental level, Gwet makes a distinction between a rating being either *deterministic* or *random*. Agreements where both ratings are deterministic count as actual agreements—in opposition, spurious agreements are such that at least one of the ratings must be random. Furthermore, if a rating is random, he presumes that all response categories will be equally likely to result from the rating process. Recall that accordingly the upper bound of the AC_1 estimator for agreement by chance is set by $\frac{1}{k}$ for k categories. We have looked at the rationale for this calculation above and we have seen that it is, by and large, a sensible measure for spurious agreement. Indeed, it is due to this presumption that Gwets proposition concerning the amount of chance agreement must be regarded as a truism on probabilistic grounds caused by the equal probabilities involved. Nonetheless, we have to state that it does not deliver an appropropriate gauge for every rating situation. Suppose a rating situation, called C, where *all* ratings are random. Here, and this must be emphasized, the rating is *purely chance-driven*. Given that there are two response categories, Gwets reliability theory makes the prediction that the probability for agreement by chance can be maximally 0.5. Nevertheless it is comparatively easy to show that purely chance-driven processes can end up with values

that exceed that upper bound. For the sake of our argument here, we shall presume that the outcome of spurious ratings can be determined by, say, the chance-based process of drawing colored balls from an urn.[1] Imagine there were eight white and two red balls in a container. The raters draw blindly one after the other, whereby the ball is put back into the urn immediately after each draw, and only the color of the ball gets noted. The probability for the event of "drawing a white ball" comes to $P_{white} = 0.8$ in this setting, while the probability for "drawing a red ball" amounts to $P_{red} = 0.2$. Furthermore, the probability for the complex event of "both raters drawing a white ball *or* both raters drawing a red ball" equals the sum of $P_{white \times white} = 0.64$ and $P_{red \times red} = 0.04$, that is 0.68. Interpreted as the probability for a spurious agreement according to our premises here, this result contradicts Gwet's upper bound of 0.5 for sure. And, clearly, this is due to the unequal probabilities exploited in our example. To exemplify the point underlying this rather abstract thought experiment, we will try to bind it back to a more realistic chance-driven rating situation. Here, each rater will have to make a chance-based choice between the two categories. The probability of chance agreement in situation C equals 0.5 (what AC_1 says) *if and only if each rater is as likely to choose the one category or the other in his random decisions*. This condition gets violated, e. g., if at least one of the raters has a bias, say, a special liking towards one of the categories in cases of doubt.[2]

If we try to imagine a couple of rating situations which permit purely random ratings, we will find that they have something in common: the raters will have to cast their votes *regardless of which feature the rated object actually exhibits*. Implicit in this statement is a realistic ontological position pertaining to the features of the objects involved. We shall dwell on this a bit more. Suppose the following modification of the ball-drawing example above: this time the raters classify the colors of balls, but they also have to wear funny glasses which, unbeknown to them, make ball colors invisible. Now suppose a rater holds a certain ball in his hand. What is the color of this ball? Replying to this question, the rater is left with two options: firstly, he can believe that the ball *factual* has no real color, because he cannot determine one. Secondly, he can come to the opinion that the ball has a color, but he just cannot *see* it. The first option makes an *ontological* assumption, the second an *epistemic* one. To cut a long story short: We think that the ontological claim cannot be held, since it implies the existence of vague or uncompleted objects—something that cannot be, [cf. Evans, 1978, Lewis, 1988]. So we have to stick to the epistemic view, which seems to be the only viable alternative. To state it properly: Uncertainty is an epistemic fact in the context of type-ii ratings and is a source for agreement by chance.[3] Intuitively, and from the point of view of an experienced annotator, the conception of type-ii ratings with epistemic uncertainty is a

[1] Other pertinent examples: the rater might toss a coin (if two categories are involved) or roll a dice (for six categories) and thus realize his rating as a purely chance-driven process. However, for the sake of our argument here, we will presume that the chance-driven process is based on unequal probabilities, e. g. a marked dice.

[2] Metaphorically speaking, this would correspond to a marked coin in coin-tossing or, of course, to the distribution of the different colors of the balls in our urn example above.

[3] We say "*a* source" rather than "*the* source" since there are more possible shortcomings like instructions that do not deliver clear rating criteria.

well-suited rendering of codings on a nominal scale. However, it seems that this poses serious problems for validity. Of course, one purpose of doing annotation is to get data that reflects *true facts* about the phenomena in question. But with a type-ii setting and the epistemic ignorance bound up with it, it seems that it will not in every case be possible to determine a true score, i. e., the "right" classification.

So do we have to say goodbye to valid type-ii ratings? Not necessarily, we think. In measurements, we are restricted to type-ii assignments just in case the subject matter is not properly understood—recall the respective remarks in the introductory section. Thus, suppose there is a theory accounting for the phenomenon we want to explore with our ratings. In this case it is probable that a measurement procedure can be found that is independent from the rating procedure and which delivers true values, i. e., a "gold standard". It is advisable to illustrate this rather abstract discussion with an incisive example: measuring the pepperiness of a chili fruit. The substance that is responsible for the "bitingness" of chilis is a complex molecule called *Capsaicin* ($C_{18}H_{22}NO_3$). Capsaicin stimulates the sensory receptors for heat located on the tongue and thus induces a "thermic deception"—what in the English language is accurately called "hot". The more capsaicin is contained in a chili pod, the hottier is the pod. The amount of capsaicin is measured in terms of Scoville units, a measurement scale developed by the pharmacologist Wilbur L. Scoville in 1912. Though the subject matter of a certain pepper fruit being hot is well understood and explicable in chemical terms, until recently the Scoville measurement had to be carried out by several raters which had to rank pods according to the Scoville scale. Of course, this kind of measuring is subject to subjective impact since sensing hotness is relativized by gusto and routine. Nowadays, with the advent of High Performance Liquid Chromatography (HPLC), it is possible to detect the objective amount of capsaicin fairly exactly. Alas, the possibility to determine hotness on a Scoville-scale using HPLC, that is, in a way that is independent from subjective ratings, does not make ratings any better. The point is, that with an objective procedure existing besides the rating procedure we are in a position to gather data depicting the "gold standard". This in turn enables us to determine the validity of our ratings by comparing them to the gold standard. Having "true values", we are also able to get their distribution pattern, which might be of some interest.[4]

3.2 Towards a statistic that does justice to our intuitions?

Of course, our readers will have noticed that theoretical issues in chance-correction have been a main focus of this report. This, however, was not a direct concern when we envisaged our evaluation efforts: originally, we simply wanted to collect some data for our scheme, compute the appropriate reliability statistics, and see how they fit in comparison with the values that have been reported for other schemes in the pertinent literature. However, as soon as we recognized that there were problems with the more popular statistics, we began to take a closer look at their foundations—and things began

[4] Recall that unbalanced distributions of the target features in a rating domain has been an issue in showing the inappropriateness of the *kappa* estimate, cf. page 15.

to develop a dynamics of their own. Now, this definitely is an important issue to address with possible applications ranging from the evaluation of diagnoses in medical facilities to the assessment of corpus-based research in computational linguistics. Therefore, we feel justified in spending a few more lines on related questions here. The discussion of some intricate points will lead to the expression of ideas that might be made use of in order to construct an original statistic. However, these points can not be worked out to a nearly sufficient degree of detail here. Indeed, the topic of chance-infectedness and how to come to terms with it is a jinxed one—and it has surely bedevilled us during the course of our work on this report.

Of course, at times we have been hunted by hunches on how to improve on the established statistics by means of constructing an original coefficient. In this regard, it seems that one fundamental question is how to assess the degree of chance-infectedness (in order to arrive at a moderating weight factor for the chance term in the golden formula), another point pertains to the exact probabilities involved in chance-based ratings (in order to calculate the respective joint probabilites for the determination of the chance term, again); compare our discussion and the detailed critique of the *kappa* and *pi* statistics in chapter 2 above.

Concerning the first point, our main idea is to opt for a measure of the raters' certainty concerning their respective decisions empirically. The rationale for this is as follows: if we get such a measure for each and every rating investigated, we can make use of it in order to weight the respective agreements and disagreements in a sensible way, compare our arguments in section 2.2.1 above. Such a measurement for the degree of chance-infectedness could be implemented along several possible ways: a *rater-based* approach would be to ask the raters directly for each and every rating performed, i. e., *online* during the rating situation. Thus, the rater would have to provide a meta-rating concerning his degree of confidence in his rating, e. g., as an estimate of confidence in percent. However, it is hard to see how this could be realized without disturbing the flow of the rating process considerably.[5] On the other hand, if we decided to ask our questions afterwards, i. e., *offline*, the raters themselves might not remember the answers for the individual cases or, even worse, the answers might be misremembered. However, the perhaps biggest problem for such a rater-based approach is that it can well be doubted whether such meta-ratings can be performed by the rater themselves in a correct and reliable way due to an inherent lack of objectivity and competence with regard to their own ratings. But if we cannot guarantee that the reliability classification itself is sound and stable, we are in danger of running into vicious circles: so we might well end up writing papers on how to assess reliability (or validity) on reliability studies concerning annotation and so forth.[6] However, concerning the possibly viable alternative of a respective evaluation

[5] This might even amount to a distortion of the result of the rating process we have set out to evaluate.

[6] Due to the inherent uncertainties with regard to the chance estimate in the AC_1 statistics, a similar point can be made against our application of Gwet's statistics in this report. However, it seems that *reliability* it not so much the issue here, since calculations using AC_1 over the same data will deliver identical results on repeated trials (apart from calculation errors). Nevertheless, we may well speculate whether AC_1 delivers *valid* results for the reliability assessmemt of the ratings investigated. To put it differently: the question is whether the AC_1 statistics is an appropriate measure of reliability for the kind

from the outside, say, a *supervision-based* approach, it is perhaps equally hard to see how a neutral expert could come to sound results regarding the relative seriousness or spuriousness of his subject raters' ratings, this time due to a lack of knowledge. Therefore, it might be a good idea to try out alternative *data-based* approaches. For example, with our ratings concerning gesture functions in mind, this might seem to be not that difficult at all. Against the background of our discussion in [Lücking and Stegmann, 2005, sec. 3.3], it seems reasonable to assume that ratings concerning gestures towards the middle columns of the pointing table lend themselves rather to an interpretation that assumes chance as an effective factor in the ratings (due to the uncertainties with regard to the underlying rating instructions under such circumstances). We could, for example, construct a "confidence factor" with regard to the proportion of ratings concerning the individual columns and categories, respectively. Such "confidence factors" could be used for weighting the observed agreements accordingly. However, a serious difficulty is the question of how to come up with such criteria for rating dimensions in general. Perhaps data-invasive algorithmic techniques such as data mining might be appropriate for the task of discovering respective regularities in the bare data from category to category. But due to the peculiarities involved, this strategy does not seem to lend itself to a general uniform practice, as to be desired. However, there may be one more promising way left open for us: we could make use of *stability-based* results (over several trials). Recall from section 2.1 that intra-observer reliability can be measured using a test vs. re-test design. Now let us hypothesize that serious ratings should come to identical results over many repeated ratings, which seems to be not implausible at face value. That is, we could repeat the rating procedure with the same raters and the same tokens several times and compare the results on a token- and rater-based basis. Those ratings that remain identical (or at least almost identical) with respect to the same token for the same rater over all ratings should be judged as serious.[7]

Concerning the second point, we might try to think of procedures in order to determine adequate probabilities characteristic of spurious ratings. If successful, we might end up with unbalanced probabilities concerning the different categories (*contra* Gwet) and/or probabilities different from the overall categorial proportions observed for the raters (*contra* Cohen and Scott). Furthermore, the respective probabilities might also be differing from rater to rater, e. g., if bias is involved. However, such procedures are hard to

of settings where it finds application or not. Note that this point does not attack the AC_1 coefficient as a theoretical construct, since that is defined in its own rights as a model, but rather its adequacy in terms of use of the corresponding statistics (= instantiations of the model) in certain settings for a specific aim. However, we think that we have taken up that challenge on our own terms above, since, firstly, we stressed the point that there is no serious alternative yet and, secondly, we argued that AC_1 results seem to be at least better (= "more valid" or "close to valid") than those obtained with the other statistics due to the known reasons.

[7] However, if there is a strong bias that applies in all cases of doubt (= chance-based ratings), we might mistakenly judge such ratings as serious when they are, in fact, just biased. The question, however is, whether there are any empirical means to distinguish between deterministic ratings and chance-based ones that are performed on grounds of strong biases. When there are none, the distinction itself might collapse. However, it seems that gold standards, when available, might be of relevance in this respect, hence, the problem does not seem to disappear.

come up with and we can also imagine to take, e. g., a free ride on Gwet's fundamental premise in the sense of assuming equal probabilities for the categories in chance-based ratings (along the lines of his random = equal chance ratings). This would have the advantage of being much more easy to handle and it also seems to be one plausible line of interpretation concerning what is involved when we talk about chance-based ratings[8]. After all, however, it seems that a correct answer concerning our question depends on what exactly we take chance-infected ratings to be: whether, e. g., biased ratings must be subsumed as chance-infected or not, and further, if we subsume them, whether it seems plausible that such biases will manifest themselves in the overall proportions observed or not (or to which degree). A positive answer with regard to the latter point might even lead to an uptake of the *kappa* or *pi* suggestions concerning the pertinent probabilities (following the overall proportions for the raters individually or on average).[9]

As things stand, it seems that we still have to decide on some critical points. Concerning the estimation of the chance term we have to decide how the probabilities involved in chance-based ratings shall be determined. Furthermore, we would like to incorporate a weighting factor in order to attenuate the chance term due to the actual proportion of chance-based ratings. Hence, we have to measure that proportion. Given appropriate measures or decisions[10], we could construct our original coefficient, say *omega*, along the lines of the golden formula as follows:

$$K_\omega = \frac{P(A) - P(C_\omega)}{1 - P(C_\omega)}, \tag{3.1}$$

with $P(C_\omega)$ instantiated as follows:

$$P(C_\omega) = w \cdot \sum_{i=1}^{k} (p_{1i} \cdot p_{2i}) \tag{3.2}$$

Here, w stands for the attenuating weighting factor, which will be given by the proportion of the chance-based ratings (against the overall number of ratings).[11] Furthermore,

[8] Compare, e.g. table 3.15 in section 3.3 of [Lücking and Stegmann, 2005]: it displays the distribution of gesture function ratings according to the different regions on a table (= relative distance to the person who is gesticulating). For the seemingly unclear cases in the middle columns, we may suspect that the raters have chosen the one category as likely as the other (although there is a clear difference between the raters regarding the columns). Indeed, this might be *the regularity to search for* using data-based techniques ($\backsimeq$ equal proportions in the ratings).

[9] Nevertheless, we would still want to weight them with an attenuating factor, as we would, of course, also do with equal probabilities (along the lines of Gwet)!

[10] Respective decisions might lead to some simplifications concerning the exact form of formula (3.2), e. g., if equal probabilities for all categories were demanded.

[11] We have not yet answered the following question: what shall we do, when only one of the raters performs a chance-based rating for a certain token? We imagine that this will be observed often, given sensible measures for the determination of the spurious ratings. Each rater will perform a certain number of chance-based rating, but not necessarily on the same tokens as the other rater. Shall all cases enter into the attenuating weighting factor or only those in the overlap? We only give a first-shot answer (along the position of Gwet) here: all of them should count—the weighting factor shall be determined by the relative proportion of the "union" of the spurious cases.

the p_{1i} term represents the probability of rater 1 to come to a decision in favor of category i in a chance-based rating (and analogously for p_{2i}, representing rater 2's chance-decision probability concerning category i). Of course, we calculate the sum for the joint probabilities across all categories. Then we weight it, i. e., multiply it with the proportion of chance-based ratings, in order to arrive at a sensible measure for the chance component in the golden formula.

It has to be noted that the above formula ist not necessarily simply a function of the entries of a contingency table, as it has been the case for those in the preceding chapter. Indeed, we have speculated whether our points here can be tackled along the measurement of additional empirical parameters or on grounds of theoretical decisions alone. However, we readily admit that we still feel a bit uneasy concerning these sensible issues due to the complexity of the various arguments and counter-arguments that have been summarized here and in other parts of this report. Furthermore, we still have to systematize our approach in a way comparable to the rigor of the theoretical apparatus of [Gwet, 2001] as an idealized model and a respective statistic. This must also incorporate detailed suggestions regarding pertinent significance tests, i. e. ways of determining appropriate variance measures in order to account for issues such as sampling variability and the likes. Our respective decisions and the systematization issues remain to be solved in detail yet.

Bibliography

Jean Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, pages 249–254, 1996.

Edward G. Carmines and Richard A. Zeller. *Reliability and Validity Asessment*. Quantitative Applications in the Social Sciences. SAGE, Beverly Hills / London, 1979.

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

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

Barbara DiEugenio and Michael Glass. The kappa statistic: a second look. *Computational Linguistics*, 30(1), 2004.

Gareth Evans. Can there be vague objects? *Analysis*, 38:208, 1978.

Alvan R. Feinstein and Domenic V. Cicchetti. High agreement but low kappa: I. the problem of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549, 1990.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382, 1971.

Kilem Gwet. *Handbook of Inter-rater Reliability*. STATAXIS Publishing Company, 2001. URL `http://www.stataxis.com/`.

Kilem Gwet. Kappa statistics is not satisfactory for assessing the extent of agreement between raters, April 2002a. URL `http://www.stataxis.com/`.

Kilem Gwet. Inter-rater reliability: Dependency on trait prevalence and marginal homogenity, May 2002b.

Per Kjærsgaard-Andersen, F. Christensen, S. A. Schmidt, N. W. Pedersen, and B. Jorgensen. A new method of estimation of interobserver variance and its application to the radiological assessment of osteoarthrosis in hip joints. *Statistics in Medicine*, 7: 639–647, 1988.

Stefan Kopp and Ipke Wachsmuth. Synthesizing multi-modal utterances for conversational agents. *Computer Animation and Virtual Worlds*, 15(1):39–52, 2004.

Alfred Kranstedt, Stefan Kopp, and Ipke Wachsmuth. MURML: A multimodal utterance representation markup language for conversational agents. In *Proceedings Workshop Embodied Conversational Agents*, Bologna, Italy, 2002. First International Joint Conference on Autonomous Agents & Multi-Agent Systems.

Klaus Krippendorff. *Content Analysis*, volume 5 of *The SAGE* KommTexT *Series*. SAGE Publications, Beverly Hills / London, 1980.

Peter Kühnlein and Jens Stegmann. Empirical issues in deictic gestures: Referring to objects in simple identification tasks. Technical Report 2003/03, CRC 360 "Situated Artificial Communicators", Bielefeld University, 2003.

Peter Kühnlein, Manja Nimke, and Jens Stegmann. Towards an HPSG-based formalism for the integration of speech and co-verbal pointing. In Jürgen Streeck, editor, *Gesture: The Living Medium. Proceedings of the first congress of the International Society for Gesture Studies (ISGS)*. University of Texas at Austin, June 5-8, 2002 2003. URL `http://www.utexas.edu/coc/cms/International_House_of_Gestures/Conferences/Proceedings/Contents/List_of_Papers.html`.

Peter Kühnlein, Alfred Kranstedt, and Ipke Wachsmuth. Deixis in multi-modal human computer interaction: An interdisciplinary approach. In A. Camurri and G. Volpe, editors, *Gesture-based communication in human-computer interaction*, number 2915 in Lecture Notes in Artificial Intelligence, pages 112–123, International Gesture Workshop 2003, Genua, Italy, 2004. Springer: Berlin, Heidelberg. revised papers,.

David Lewis. Vague identity: Evans misunderstood. *Analysis*, 48:128–130, 1988.

Richard H. Lindemann, Peter F. Merenda, and Ruth Z. Gold. *Introduction to Bivariate and Multivariate Analysis*. Scott, Foresman and Company, 1980.

Frederic M. Lord and Melvin R. Novick. *Statistical Theories of Mental Test Scores*. Behavioral Science: Quantitative Methods. Addison-Wesley, 1968.

Andy Lücking and Jens Stegmann. Assessing reliability on annotations (2): Statistical results for the DEIKON scheme. Technical report, Universität Bielefeld, SFB 360, Projekt B3, 2005.

Andy Lücking, Hannes Rieser, and Jens Stegmann. Statistical support for the study of structures in multi-modal dialogue: *Inter*-rater agreement and synchronization. In Jonathan Ginzburg and Enric Vallduví, editors, *Catalog '04—Proceedings of the Eighth Workshop on the Semantics and Pragmatics of Dialogue*, pages 56–63, Barcelona, 2004. Department of Translation and Philology, Universitat Pompeu Fabra.

Quinn McNemar. *Psychological Statistics*. John Wiley & Sons, New York / London / Sydney / Toronto, 4^{th} edition edition, 1969.

Hannes Rieser. Pointing in dialogue. In Jonathan Ginzburg and Enric Vallduví, editors, *Catalog '04—Proceedings of the Eighth Workshop on the Semantics and Pragmatics*

of Dialogue, pages 93–100, Barcelona, 2004. Department of Translation and Philology, Universitat Pompeu Fabra.

Toni Rietveld and Roeland van Hout. *Statistical Techniques for the Study of Language and Language Behaviour*. Mouton de Gruyter, 1993.

William A. Scott. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, XIX:321–325, 1955.

Sidney Siegel. *Nonparametrical Statistics for the Behavioral Sciences*. McGraw-Hill Series in Psychology. McGraw-Hill Book Company, Inc., Tokyo, 1956.

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

Peter Sprent. *Applied Nonparametric Statistical Methods*. Chapman and Hall, London / New York / Tokyo / Melbourne / Madras, 1989.