Taming our wild data: On intercoder reliability in discourse research

van Enschot, R.; Spooren, W.; van den Bosch, A. ; Burgers, C.; Degand, L.; Evers-Vermeul, J.; Kunneman, F.; Liebrecht, C.; Linders, Y.; Maes, A.

DOI
10.51751/dujal16248

Publication date
2024

Document Version
Final published version

Published in
Dutch Journal of Applied Linguistics

openjournals.nl

Dutch Journal of Applied Linguistics
https://www.dujal.nl | E-ISSN 2211-7253

Taming our wild data: On intercoder reliability in discourse research

Renske van Enschot[1], Wilbert Spooren[2], Antal van den Bosch[3], Christian Burgers[4], Liesbeth Degand[5], Jacqueline Evers-Vermeul[3], Florian Kunneman[3], Christine Liebrecht[1], Yvette Linders[2] and Alfons Maes[1]

[1] *Tilburg center for Cognition and Communication, Tilburg University* | [2] *Centre for Language Studies, Radboud University* | [3] *Institute for Language Sciences, Utrecht University* | [4] *Amsterdam School of Communication Research (ASCoR), University of Amsterdam* | [5] *Institute for Language and Communication, University of Louvain*

Abstract Many research questions in the field of applied linguistics are answered by manually analyzing data collections or corpora: collections of spoken, written and/or visual communicative messages. In this kind of quantitative content analysis, the coding of subjective language data often leads to disagreement among raters. In this paper, we discuss causes of and solutions to disagreement problems in the analysis of discourse. We discuss crucial factors determining the quality and outcome of corpus analyses, and focus on the sometimes tense relation between reliability and validity. We evaluate formal assessments of intercoder reliability. We suggest a number of ways to improve the intercoder reliability, such as the precise specification of the variables and their coding categories and carving up the coding process into smaller substeps. The paper ends with a reflection on challenges for future work in discourse analysis, with special attention to big data and multimodal discourse.

Keywords intercoder reliability, discourse, quantitative content analysis, complex discourse data, hands-on procedures

Article history
Received: July 25, 2023
Accepted: February 2, 2024
Online: March 26, 2024

Corresponding author
Renske van Enschot,
r.vanenschot@tilburguniversity.edu

Acknowledgements
Author contributions
Funding information
Statement of interest
Statement of technology use
See p. 19–20

1 Taming our wild data: On intercoder reliability in discourse research

Many research questions in the field of communication, linguistics and cognition are answered on the basis of manual analyses of data collections or corpora: collections of (transcribed) spoken, written and/or visual communicative messages. Although many different forms of corpus analysis are used (Krippendorff, 2019), the generic base can be defined as assigning interpretative categories to particular variables in the corpus. For example, particular words or expressions can be classified as having an intensifying meaning or as being ironic or metaphoric; gestures or pictures can be classified as representational or decorative; pitch patterns can be categorized as either expressing or lacking a feeling of knowing on the part of the producer; a particular interpretation of

How to cite Van Enschot, R., Spooren, W., Van den Bosch, A., Burgers, C., Degand, L., Evers-Vermeul, J., Kunneman, F., Liebrecht, C., Linders, Y. and Maes, A. (2024). 'Taming our wild data: On intercoder reliability in discourse research', *Dutch Journal of Applied Linguistics* 13: DOI 10.51751/dujal16248

an artful poster or advertisement can be classified as 'matching the intended meaning' or not; a coherence relation between two utterances in a discourse can be labeled as semantic or pragmatic.

Obviously, these categorization tasks are extremely diverse in nature. For one thing, the span of the units can vary enormously: units can be sounds, words, phrases, utterances, images, complete texts or conversations. For another, the type of unit varies: annotating pitch in a telephone interaction is a completely different task from annotating the verbal language in that interaction, and the analysis of words is a task different from the analysis of gestures or images. It becomes even more complicated when the different modalities are combined. It is this variety in span and type of units that makes the analysis of discourse more intricate than the analysis that restricts itself to one linguistic level such as morphology and grammar. Despite the diversity, all of these activities fall under the heading of discourse coding: analyzing discourse in a systematic way in order to gain insight into underlying patterns. Often, this task is performed in a quantitative way, establishing or comparing frequencies of use, or trying to extrapolate frequencies from samples to populations.

In this kind of quantitative content analysis of discourse, the coding of subjective language data often leads to disagreement among raters. This is partly due to coding errors, and partly due to the inherent ambiguity of the language phenomena. Disagreement can occur even after an extensive training phase, even if an explicit code book is used that has been tested and adapted, even if the number of coding categories is limited, and even if experts are employed instead of naive and untrained coders (Spooren & Degand, 2010). Problems increase for the analysis of static and dynamic visuals in discourse. Often, several rounds of coding are necessary to reach a sufficiently high intercoder reliability statistic such as Cohen's Kappa. At the same time, our publication outlets require that such a Kappa is reached after only one round of coding (cf. Neuendorf, 2002, p. 146). Allegedly, only with one round of coding, categorizations can be considered replicable and valid.

In most cases, the discourse units of interest come from a continuous stream of information. As a consequence, these units are often not predefined, and coders need to unitize the discourse before they can actually start categorizing these units. This may give rise to agreement issues as well, since coders can disagree about the number, length and alignment of the units in a discourse. For an analysis of the intricacies of unitizing issues and a valuable suggestion for quantitatively assessing the agreement on the unitizing effort, we refer the reader to Krippendorff (2019), Krippendorff et al. (2016) and Mathet et al. (2015). In the remainder of this paper, we focus on the issue of categorizing predefined units. Our restriction does not imply that we discard unitizing discourse, or the possibility that unitizing and categorization are interdependent, but we do believe that categorizing discourse units is a challenging task in itself, worth a separate discussion. Also worthy of a separate discussion is crowdsourcing. This can be seen as an interesting development that researchers make use of to have their data annotated by large numbers of naive

Table 1 An overview of the key terms

Term	Definition
Coding categories	The different values that a variable can take
Corpus	A collection of instances of discourse, usually enriched with metadata (for example about participants, register, source) and annotation (for example Part-of-Speech tagging, lemmatization)
Discourse	A continuous stream of naturally occurring communication in one or more modalities (for example written, spoken, video, computer-mediated) and using one or more verbal or non-verbal codes.
Discourse coding	The assignment of coding categories to a discourse unit
Intercoder reliability (ICR)	The degree to which different coders agree on assigning categories to discourse units
Levels of measurement	The nature of the information incorporated in a variable; for instance, an ordinal level of measurement allows for a rank order by which data can be sorted, but it does not allow for relative degree of difference between them
Quantitative corpus analysis	A type of research that uses corpus data and aims at finding frequency information about patterns in these corpora in order to be able to establish or compare those frequencies or extrapolate them to populations
Unit	The object of analysis to which a coder can assign a coding category
Variable	A property of interest that follows from the research question. It consists of at least two categories that presuppose a particular level of measurement

coders, for example via ProLific (Suviranta & Hiippala, 2022). Another noteworthy omission is the possibility of biases in cross-cultural analyses (Peña, 2007; Quiros-Ramirez & Onisawa, 2015).

In our paper, we make use of the terminology in Table 1. For example, in her *quantitative corpus analysis* of metaphoric language use in news discourse, Pasma (2011) worked with a *corpus* of newspaper texts from two time periods (1950/1951, 2002; 80,000 words per period). The *units* of her analysis were the words in the newspaper texts. For each word, she assessed the *coding category* of the *variable* metaphoricity: a word was used metaphorically or not; in other words, the *level of measurement* was nominal. She assessed the *intercoder reliability* (*ICR*) by having a second coder perform the same *discourse coding* task and calculating an ICR statistic.

Our paper differs in several respects from related work like Bayerl and Paul (2011), who give a quantitative meta-analysis of factors influencing percentage agreement in a large

number of linguistic corpus studies in different domains (word-sense disambiguation, prosodic transcriptions, and phonetic transcriptions). Our aim is to give a qualitative analysis of the factors involved, using case studies from various areas of discourse analysis. As a consequence, we are not limited – as Bayerl and Paul are – to the restricted number of factors available in the data set that is used.

In the next sections, we first sketch the scope of the reliability problem by describing different degrees of 'wildness' in discourse data. We then describe the tension between reliability and validity. Subsequently, we provide an overview of shortcomings of using Cohen's Kappa scores as a measure of intercoder reliability – ICR from now on –, and suggest alternative statistical ICR metrics. Our paper continues with hands-on advice on how to improve ICR, and concludes with a reflection on where to go from here, with special attention to big data and multimodality.

2 Different degrees of 'wildness' in discourse data

The quality and outcome of corpus analyses, as well as the success of ICR tests, is determined by a large number of factors. In this section, we will discuss the effects of three important factors on the openness and ambiguity ('wildness') of discourse data: different ways of collecting data, different ways of establishing coding categories, and the kind of coding categories that are taken into account.

First, data collections can be elicited experimentally or selected from naturally occurring data. The latter tends to be more difficult to code reliably. For instance, Arts et al. (2011) asked their participants to produce referring expressions describing entities on a screen in a controlled setting. The result was an easy-to-code dataset of referential expressions containing only attributes visible on the screen. This differs sharply from the problems encountered while encoding the stylistic elements in naturally-occurring newspaper and web texts, such as the ones reported in Liebrecht (2015).

Second, coding categories can be established in different ways. On the one hand, they can have a predefined theoretical position and definition. Examples are using an enchiridion – a short handbook – to categorize intensifiers on a lexical base (Van Mulken & Schellens, 2012), or using an unambiguous and theory-based definition of verbal irony (Burgers et al., 2011). Low intercoder reliability could be prevented if researchers would study only variables with such well-defined categories that hardly cause any interpretation noise. On the other hand, coding categories can start from an intuition and emerge gradually and exploratively as the analysis proceeds (Van Enschot & Donné, 2013). Many interesting questions in the field of human communication are not ready for controlled types of research, and therefore require an explorative approach. Examples of such intriguing questions are: 'What makes a visual message metaphorical (Forceville & Urios-Aparisi, 2009)?' and 'Which linguistic or audio-visual cues can we consider to be deceptive

(Hancock et al., 2007)?' The coding of the latter type of data tends to be much more difficult than the coding of the first type.

Note that agreement tests can be useful both in controlled and exploratory studies. Controlled studies require a high degree of precision and validity, and consequently only high levels of ICR outcomes are acceptable. In explorative studies, however, ICR scores can be used as a heuristic tool to objectify or specify individual intuitions, to try out coding category definitions or segmentation options in data collections. During this incremental process, lower ICR rates are acceptable, although one may want to eventually perform a more controlled analysis to validate the final version of the newly developed coding system.

Third, a set of coding categories can vary from closed to open. On the one extreme, the coding categories form a dyadic, mutually exclusive, closed set (for example yes-no, present-absent), clearly defined in terms of objective characteristics. On the other extreme, the number of coding categories is not fixed (for example ways to intensify an utterance) or the coding categories are not mutually exclusive (for example literal versus figurative language). Such coding categories leave room for an exploratory analysis, but at the same time pose problems for ICR.

Our discussion of these three factors illustrates that analytical data can be less or more wild, and that the degree to which our data are wild depends on the choices made by the researchers. The researchers make decisions about the type of questions asked (for example, do they focus on the question whether or not an utterance is subjective or whether a specific word indicating subjectiveness such as *terrific* occurs in the utterance), about the collection method used (for example, do they focus on a corpus of all tweets produced in a one-hour slot in The Netherlands, or on the one-word-production results of a word-listing task), about the theoretical framework and position taken (for example, do they start from the position that all subjective words can be divided into two or three major coding categories or not), and about the way in which theory is operationalized into coding categories (for example, do they define subjective words as a closed set of identifiable elements in discourse, or as an open-ended class).

These choices to a large extent determine the success of the coding task. Once we agree that data can be wild, the question is which precautions can be taken to make intercoder reliability tasks doable, not trivial, and successful or at least informative. We will address such precautions in the section on how to improve intercoder reliability.

3 Reliability versus validity

Data quality is a matter of both reliability and validity. To illustrate the difference between reliability and validity, various scholars use the metaphor of a target or a dart board (Krippendorff, 2019; Trochim, 2006). Imagine your goal is to systematically hit the bullseye. Your performance is unreliable if the darts are randomly spread across the board. It is

reliable if you consistently hit the same spot on the board – irrespective of whether this is the bullseye or some other spot, such as the 'double 20'. However, your performance is only valid if you consistently hit the intended spot on the board, and not if you just hit the bullseye every now and then. Thus, the 'dart board' metaphor demonstrates that reliability and validity both need to be high, and that high reliability does not necessarily imply high validity. Similarly, aiming for high ICR scores is not a guarantee for good validity. On the contrary, it may even yield serious problems for the construct, external and internal validity of the research.

An established viewpoint is that reliability is at least a necessary condition for validity (Moss, 1994). What does this imply for researchers with 'wild' data that frequently result in insufficient intercoder reliability scores? One way out of this problem is to consider the theory or the coding procedure yielding the analysis of such unreliable data to be underdeveloped to such a degree that the researchers should go back to the drawing board. Alternatively, we could restrict the generalizability of our results to the limited set of data that we can code reliably. Such a solution is chosen by Liebrecht (2015) for the analysis of intensified language. She reports analyses on the subset of data on which both coders agreed. Of course, this limits the generalizability of the results. To accommodate that problem, she also reports findings for the intensifiers identified by each coder separately, and only draws firm conclusions when all results point in the same direction.

Below, we discuss the different types of validity, with special attention for how validity can be at odds with reliability. It will become clear that a high ICR – if achievable – is not always a guarantee for having valid data.

3.1 Construct validity

Construct validity refers to the question whether the coded data accurately reflect the theoretical constructs they are supposed to measure (Elmes et al., 2012, pp. 185-187). Aiming for high reliability scores may be at odds with this type of validity. For example, Wallace (2015) argues that the pragmatic context of utterances is a common feature of modern theoretical approaches to irony and sarcasm, but that this aspect is largely ignored in the ways in which various computational linguists have operationalized irony and sarcasm in their corpus-based research. In efforts that guarantee a boost in reliability, many researchers have coded sarcasm by looking for specific words or word combinations, such as variants of the word sarcasm (sarcastic, sarcastically, etc.), or words that are often used to mark sarcasm (for example *Yeah right*). Wallace (2015) therefore qualifies such automatic identification procedures based on word usage as "shallow", because they do not take into account semantic and pragmatic information about the speaker or situation, and are likely to miss many instances of sarcasm.

For instance, an utterance such as '*Gerald Ford was a great president. Yeah right*' is more likely to be interpreted as literal when said by a supporter of the Republican Party, and as sarcastic when it comes from a supporter of the Democrats. Wallace thus calls

upon computational linguists to develop more advanced computational models that take into account not only syntactic aspects, but also semantic and pragmatic aspects. Thus, even if identification procedures achieve satisfactory or high reliability scores – coding instances of *Yeah, right* in a corpus can easily be done very reliably – it is important to assess critically whether the data analysis meets the criterion of construct validity, i.e., whether it captures all cases included in the theoretical definition of the variable.

3.2 External and internal validity

External validity refers to the way in which observations can be generalized to other situations outside of the specific data investigated. The internal validity criterion invites the researcher to search for confounding factors. Both kinds of validity can be at odds with reliability. An example comes from research on the quality of the spelling in students' writing. It makes an enormous difference whether the researcher analyzes the spelling errors in dictations, or in texts that are composed by students themselves. As Van den Bergh et al. (2011, p. 6) point out, analyzing spelling errors in the former text types can be done in a very reliable way (they report 100% intercoder reliability for dictations). However, there are issues of external validity. Although dictations can give a systematic picture of what children are capable of in terms of specific spelling difficulties, children's numbers of spelling errors in dictations are not predictive of the number of spelling errors in their own writing: Van den Bergh et al. report correlations between .11 and .17 (2011, p. 12). Next to external validity, the internal validity is at stake in this analysis as well. First, the number of errors in students' own writing may be determined, at least in part, by their proficiency in coming up with an alternative formulation, allowing them to avoid words that are difficult to spell. Second, variation in numbers of spelling errors in dictations and students' own texts may also be attributed to differences in task. If students take dictations, their main focus is on form, not on content. However, if students write their own texts, their focus is on content, and less so on correctness of form. This confounding of factors makes it hard to compare the outcomes of studies in which different tasks are used, and it illustrates that reliability is not a sufficient condition for validity.

4 How to assess intercoder reliability statistically

In this section, we turn to a selection of settings in which the most widespread intercoder reliability metric, Cohen's Kappa, can be misleading: when each item is annotated by a different selection of annotators, when the categories are in an ordering relation, or when there is an imbalance in the frequency at which different categories are annotated. We provide an overview of metrics that can be applied alternatively to Cohen's Kappa. We will focus on the main characteristics and advantages of these alternative metrics,

without providing formulas.[1] The issues we describe all assume that the discourse units have already been established, that is: we describe the situation in which the unitizing stage has been finished, and the researcher has to categorize the units of interest.

Cohen's Kappa was proposed by Jacob Cohen (1960) and takes into account the prior chance that two annotators agree on the annotation of any coding category. This makes Cohen's Kappa a more realistic metric for intercoder reliability than percentage agreement, which can easily give misleading insights. For instance, one study using two coding categories has a fifty percent chance that coders agree, while another study using four coding categories yields a 25 percent agreement chance. As a result, given a similar percentage agreement, the first study will yield higher ICR scores than the second study simply due to chance (Artstein & Poesio, 2008, pp. 558-559).

As a way to account for chance agreement, Cohen's Kappa presumes all items in a set to be coded by the same two annotators who are considered statistically independent. The chance for annotator 1 to choose certain categories is calculated, as is the chance for annotator 2 to choose certain categories. Based on these chances, the possibility for two annotators to agree coincidentally is calculated. As Krippendorff (2011) explains, this is not a realistic perspective on the role of chance in ICR: Annotators are dealing with the same categories to code and have read the same code book, and hence their annotations should not be seen as independent sets. A consequence of this is that the Cohen's Kappa metric does not penalize patterns where the class distributions of the annotators are highly distinct. For example, if annotator 1 has chosen category A for 80% of the items and annotator 2 has only chosen this category for 40% of the items, the chance that both agree on this category is (0.8*0.4=) 0.36, and the chance that the two agree on the other category is (0.2*0.6=) 0.12, resulting in a chance agreement of (0.36 + 0.12=) 0.48. This is a smaller penalty than when both annotators would for example choose category A for 40% of the times (a change agreement of 0.52). It would make more sense to look at the combined decisions of all annotators as a proxy of the class distribution, which is done in metrics like Fleiss' Kappa (Fleiss, 1971), Scott's pi (Scott, 1955) and Krippendorff's Alpha (Hayes & Krippendorff, 2007; Krippendorff, 2019). These metrics consider the proportion of times that each coding category is chosen by annotators, as well as the agreement per single item. Essentially, annotators are then seen as interchangeable, which is more in line with the notion of reliability of annotations: if the annotations in the end rely on a straightforward decision procedure which is reflected in the agreement, it should not matter who does the annotations and any one annotator could be replaced by another without much effect on the agreement.

An important property of the coding task is the measurement level of the variable. In standard form, the Cohen's Kappa metric presumes a nominal scale, in which no ordering exists between the coding categories. It will give wrong insights when applied to other than nominal variables, with ordinal, interval or ratio levels. With these levels of measurement, it should be penalized when two coders annotate coding categories that have a larger distance to one another. Metrics that do apply such a penalization are the

Weighted Cohen's Kappa (Cohen, 1968) and Krippendorff's Alpha (Hayes & Krippendorff, 2007; Krippendorff, 2019). In these metrics, the agreement, or disagreement in the case of Krippendorff's Alpha, for any pair of levels is weighted by taking into account the distance between the categories, such that more distanced categories add a lot less to the agreement score, or a lot more to the disagreement score. Krippendorff's Alpha also allows for missing data points, by considering the total number of annotations that were made.

Another factor that needs to be considered when assessing intercoder reliability is data skew: the degree to which data is asymmetrically distributed over coding categories. Di Eugenio and Glass (2004) and Jeni et al. (2013) show that Cohen's Kappa and Krippendorff's Alpha are highly sensitive to imbalanced variables: the agreement score will drastically decrease with a bigger data skew. The reason is that the prior chance of annotators making similar annotations is high if there are one or a few dominant coding categories. As the higher chance agreement is subtracted from the percentage agreement, a high percent agreement may still result in a relatively low ICR. In other words, it is unlikely to reach a high Cohen's Kappa or Krippendorff's Alpha when most of the items fall under one coding category. A solution is to calculate the Kappa Max (Umesh et al., 1989), which returns a Kappa value that is relative to the upper bound of the Kappa that follows the strict chance agreement. Alternatively, adaptations have been suggested to correct Kappa for such biases, such as PABAK (prevalence-adjusted bias-adjusted kappa; Byrt et al., 1993).

The Mutual F-score is another metric that can be used to mitigate the influence of category imbalance. Mutual F-scores are based on F-scores, which are often used in Information Retrieval and Machine Learning for system evaluation (Van Rijsbergen, 1979). In contrast to the other metrics, the Mutual F-score estimates the agreement per coding category rather than over the total of annotations. Considering the annotations of one coder as ground truth, for each category the annotations of a second coder can be assessed with respect to this ground truth. At the basis of the F-score are three counts: of True Positives (TP), False Positives (FP) and False Negatives (FN). TPs are instances that both coders annotate with the coding category in focus, FPs are instances that are only annotated by the second coder with the coding category in focus and FNs are instances that are only annotated by the first coder with the coding category in focus. Based on these values, Precision can be calculated as the percentage of instances coded by the second annotator with the coding category in focus, that were actually correct with respect to the ground truth annotations of the first coder (TPs divided by the total of TPs and FPs). Recall can be calculated as all instances that were annotated by the first annotator with the coding category in focus, that were annotated in the same way by the second annotator (TPs divided by the total of TPs and FNs). The F1 score is the harmonic mean of Precision and Recall: 2*(Precision*Recall)/(Precision+Recall). To reach the Mutual F-score between annotators, the role of the ground truth annotator is swapped, and the F-score is computed again in the other direction. Then, Mutual F-score can be computed

Table 2 Example confusion matrix on coding Twitter posts as sarcastic or non-sarcastic, as part of the study by Kunneman et al. (2015)

	Coder B: Sarcastic	Coder B: Non-sarcastic
Coder A: Sarcastic	205	12
Coder A: Non-sarcastic	11	22

as the mean of the two directional F-scores. As the annotations of other categories by both coders are left out of this calculation (the TNs, true negatives), the Mutual F-score allows to zoom in on the agreement per coding category.

To illustrate the informativeness of the Mutual F-score in comparison to other metrics, we provide an example from our own work. Kunneman et al. (2015) aimed to automatically detect sarcasm in Twitter posts, and trained a model based on tweets that are accompanied by Dutch hashtags likely deployed by the sender to signify the sarcasm in their post. To validate this approach, for a random sample of 250 messages that include such a hashtag, three annotators coded whether they actually were sarcastic (Kunneman et al., 2015, p. 505). The confusion matrix for one of the coder pairs is displayed in Table 2.

The percentage agreement between the two coders is 91% ((205+22)/250), but due to a dominance of the coding category *Sarcastic*, both Cohen's Kappa and Krippendorff's Alpha are 0.60. Considering the annotations of coder A as ground truth, the F-score for the coding category *Sarcastic* is 0.95 (TP = 205, FP = 11, FN = 12, Precision = 205/(205+11) = 0.95, Recall = 205/(205+12) = 0.95, F1 = 2*((0.95*0.95)/(0.95+0.95)) = 0.95) and the F-score for the coding category *Non-sarcastic* is 0.66 (TP = 22, FP = 12, FN = 11, Precision = 22/(12+22) = 0.65, Recall = 22/(11+22) = 0.67, F1 = 2*((0.65*0.67)/(0.65+0.67)) = 0.66). Given that the dataset of this variable contains two coding categories, the Mutual F-score on the coding category *Sarcastic* is therefore 0.95, and for *Non-sarcastic* it is 0.66. These two values highlight the difference in agreement on the two categories, where the less frequent coding category is the more problematic.

After having arrived at a reliability metric, the next step is interpreting this metric. Earlier on, we suggested that the ICR depends on the nature of the data and the research objective(s). This suggestion would imply that an ICR metric should be interpreted relative to these objectives. Interpretation may vary per type of study: for hypothesis testing, a higher ICR is required than for explorative studies. Similar suggestions can be found in Grove et al. (1981) and Spooren and Degand (2010) (see also the discussion in Artstein & Poesio, 2008, and Mathet et al., 2012).

In light of these examples, it is important to fully understand the mechanisms of a metric when interpreting its outcomes and be aware of how Cohen's Kappa can mislead. Next to the overview in this section, Artstein and Poesio (2008) give an excellent, elaborate overview of the mathematics and assumptions behind the ICR metrics, and discuss issues

of annotator bias (not all coders use each category equally often) and prevalence bias (some categories are used more often than others). Such overviews enable researchers to interpret the ICR metrics correctly, including Cohen's kappa, allowing them "to make an informed choice regarding the coefficients they use for measuring agreement" (Artstein & Poesio, 2008, p. 590). The discussion also shows that existing heuristics for interpreting outcomes that do not take into account the context of annotation, seem too simplistic. In line with Perreault and Leigh (1989) who state that "different indices reflect different aspects of reliability" (p. 146), we recommend assessing interrater agreement with several metrics. This way, we can achieve a more complete interpretation of the factors in play.

5 How to improve intercoder reliability

Researchers need hands-on advice and practical solutions to ICR problems. In this section, we meet this need and suggest a number of concrete ways to improve intercoder reliability: making categories independent (5.1), reducing the number of categories (5.2), decomposing the analytical process into smaller steps (5.3) and some procedural measures such as optimizing the coding instructions (5.4). Some of these topics are also discussed by Bayerl and Paul (2011), who used a quantitative meta-analysis to identify several factors influencing reported agreement values (like number of annotators and number of categories in a coding scheme). Our discussion is more qualitative, allowing us to discuss factors not touched upon by Bayerl and Paul such as the decomposition of the coding process and the independence of the categories. Bayerl and Paul's discussion is limited to three relatively local domains (word-sense disambiguation, prosodic transcription, phonetic transcription), whereas our discussion focuses on the level of discourse data. Other useful sources are for instance Krippendorff (2019), Neuendorf (2002) and Selvi (2020).

5.1 Make your categories independent

Agreement success is highly determined by the relation between the categories within a given variable. Two major conditions are relevant here: one is whether the categories are mutually exclusive or not, the other is whether the categories are hierarchically ordered. The same unit of analysis can have different functions or interpretations, which may result in scalar categories or in units belonging to more than one of the categories within a given variable (for example, clauses can have different types of relations with other clauses). Obviously, reaching satisfactory agreement scores is easier if categories are mutually exclusive.

Scalar variables can be transformed into binary variables to make the categories mutually exclusive. For example, in a study of the subjectivity of adjectives preceding or following causal connectives, Hoek et al. (2021) used the subjectivity scores on a contin-

uous scale from 0 to 1 that were available in the so-called gold1000 lexicon of subjective adjectives (De Smedt & Daelemans, 2012). For the sake of the analysis, explicit boundaries were used to create subsets of objective adjectives (defined as adjectives with a subjectivity score < .20) and subjective adjectives (adjectives with a subjectivity score > .70). The other adjectives were considered ambiguous and therefore excluded from the analysis.

Categories can also be hierarchically ordered: coders are first asked to determine major categories and then asked to subclassify units within the assigned coding category. An example is coding discourse relations first in terms of semantic versus pragmatic relations, and then the exact relation type within the assigned coding category. Agreement success is endangered if coders have to apply such embedded coding tasks (cf. Scholman et al., 2016).

This issue occurred in a study by Zufferey and Degand (2013) who report percentage agreements of three types of multilingual discourse relation annotation differing in the amount of specificity. For the most generic type of annotation, i.e., distinguishing between four categories of discourse relations (temporal, comparison, contingency, expansion, cf. Webber et al., 2019), agreement is highest, above 90 %. For the second type, which subdivides for example contingency into condition and cause, agreement drops to 60-72 %. The third, most specific type splits for example causal into reason and result, and yields agreement percentages between 39 % and 53 %. Part of the disagreement concerning the second and third type is caused by disagreements concerning the first, more generic type. This is because decisions regarding this first type directly impact decisions that have to be made for the second and third type; the latter decisions are dependent on the decisions for the first type. In example (1), the relation conveyed by *when* could arguably be either temporal or contingent. Disagreement regarding this first, more generic type automatically induces disagreement regarding the more specific type because the available subcategories will be different.

(1) The cliché of a Mediterranean lolling in the sun has become a mental reflex when trying to explain the cause of the crisis in the Eurozone.

How would one estimate the intercoder reliability in such a case? Suppose that intercoder reliability scores are calculated for the most specific type (such as whether the relation is hypothetical or general). To obtain these scores, only the cases are used on which the coders already agreed when looking at the more generic type (i.e., they agreed that the relation was one of condition and not of cause; they also agreed that it was a contingency relation, and not a temporal relation). A high ICR score for a more specific type implies a high ICR score for its overarching type. By contrast, if an ICR score for a more specific type is low, it remains to be established where the disagreement comes from. A recommendation is to proceed stepwise, to start analyzing at the first, most generic type, determine the ICR, and then move on to the second and then the third

type, analyzing just the instances of the more generic type that were agreed upon. This makes it possible to situate the coding problems with more precision.

5.2 Reduce the number of categories

Reducing the number of categories in a coding system is likely to improve intercoder reliability (see Bayerl & Paul, 2011), but this is usually undesirable because a reduced coding system yields less information. An obvious first check is whether all categories are really needed. For example, Van Enschot and Hoeken (2015) originally had two subcategories in their analysis of tropes – a type of rhetorical figures – in TV commercials: one in which the verbal part of the TV commercial explicitly mentioned the trope, as in *This woman is as beautiful as a rose*, and one in which the verbal part did not address this link explicitly, as in *This woman is beautiful*. Both were regarded as explicit explanations of the trope, and during the final phase of the analysis were therefore combined into a single coding category, resulting in higher ICR scores.

An advantage of reducing the number of categories per variable is that the remaining categories occur more frequently, which often avoids the statistical bias of unequal distribution. A case in point is the coding of the syntactic class of discourse markers (Bolly et al., 2014). This class of linguistic expressions is very heterogeneous, consisting mostly of coordinate and subordinate conjunctions such as *but* and *because*, and of adverbials such as *well* and *actually*, but also of less frequent members such as parentheticals (*I mean*, *I think*) or adjectives (*first*, *good*). This results in a variable with many categories, of which some occur infrequently. Depending on the general theory and the research question at hand, researchers can question whether it is useful to keep all possible syntactic categories or whether some of them should be grouped. Should they maintain fine-grained distinctions such as the one between coordinate and subordinate conjunctions, or between prepositions and prepositional phrases, or should they, for instance, choose to distinguish only the most probable syntactic classes (for example adverbials, conjunctions and prepositional phrases) and group all other possibilities in one encompassing other coding category, or even retain only two coding choices (for example between conjunctive and non-conjunctive). Some of these options are illustrated in Table 3.

Let us assume that the coders have a data set of 50 occurrences to annotate. If they choose to code according to option 1 in Table 3 (with ten categories), an equal distribution of all categories would lead to a maximum of five occurrences per coding category. Now, knowing that adjectives or nouns used as discourse markers are very rare in English, it is highly probable that these categories will receive zero counts. This may lead to biases in the statistical analysis as demonstrated by Feinstein and Cicchetti (1990). Therefore, either the sample has to be increased to account for rare events, or the number of categories has to be reduced. Still another recommendation is to use a statistical measure such as Kappa Max or PABAK (Prevalence-Adjusted Bias-Adjusted Kappa) that is sensitive to uneven distributions (see the section on how to assess ICR statistically).

Table 3 Coding options for the variable 'syntactic class' of discourse markers

Syntactic class 1	Syntactic class 2	Syntactic class 3
clause	adverbial	conjunctive
verbal phrase	conjunction	non-conjunctive
adverb	prepositional	
coordinating conjunction	other	
subordinating conjunction		
adjective		
preposition		
prepositional phrase		
noun		
interjection		

5.3 Decompose the analytical process into smaller steps

If reducing the number of categories is not possible without losing too much information, an interesting alternative is to decompose the analytical process into smaller, simpler steps (see Figure 1; see also Fort et al., 2012). Thus, instead of reducing the number of categories to be coded, one can simplify the coding decisions by decomposing coding steps, creating a hierarchical coding system, while at the same time reducing the number of categories that need to be considered during each step. The net result is that the same number of coding categories will be considered. The main advantage of this procedure is that the decision process is split up into smaller decision trees. Note that the decision tree described in Figure 1b seems to suggest a hierarchical variable and not the nominal variable described in Figure 1a. The same bias might hold for a categorization task as described in Figure 1a if the coder always starts with considering category A. To avoid the bias that category A is always considered first, the categories should, if possible, be presented to the coder in a random order.

For example, in a research project on literary criticism (Linders, 2014), coders had to indicate which characteristics of novels were being evaluated by critics. For each evaluative statement, the coders had to choose which of the fifteen listed types of characteristics applied. Linders' task consisted of making coders choose one of these fifteen options at once. An alternative would have been that the coders were confronted with a series of fourteen binary decisions: the first step might have been deciding whether the statement was about the book itself, the second about the effect the book had on the reader, the third about the book in relation to the world, et cetera. The result would have been a coding procedure that has a better chance of achieving a high ICR because each coding decision would be a relatively simple binary one. All in all, decomposing

Figure 1a Schematic analytical process of a complex coding category

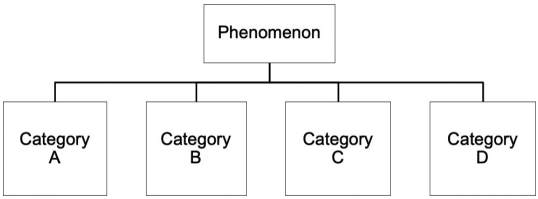
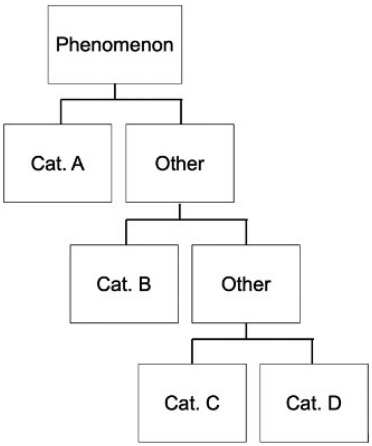


Figure 1b Simplified version of the analysis in Figure 1a



looks like a promising strategy, although it may be more time-consuming than a more straightforward single-step coding procedure.

5.4 Procedural measures

In order to facilitate the coding procedure, a number of elements have to be taken into account. We will discuss the number of coders involved, the instructions given to the coders, and their training.

5.4.1 *Consider the number of coders*

An increase in the number of coders has been shown to correlate negatively with the agreement (Bayerl & Paul, 2011). Two or three coders are standard in most studies in the field of communication, linguistics and cognition (for example Kunneman et al., 2015; Van Enschot & Donné, 2013; Van Mulken & Schellens, 2012). If coding is relatively simple (for example with a few categories to be assigned in well-defined units), one additional coder who recodes part of the data is considered sufficient (for example Mol et al., 2012). If data are wilder or categories more diffuse, coding by three or more coders may be useful, not only to obtain a reliable analysis, but also to gradually develop a sufficient understanding of the research phenomenon. In general, including more coders is more reliable (Potter & Levine-Donnerstein, 1999), but practical considerations make two or three coders reasonable.

Another interesting strategy is to deploy several coders on a part of the data, to assess the intercoder reliability on this part with an agreement measure, and if the value is high, to deploy only one coder (not necessarily the same one) for the rest of the data. An example can be found in Martin et al. (2014), where one third of the data was coded

by two coders, and the remaining two thirds were spread over both of them once the intercoder reliability had been assessed high enough.[2]

5.4.2 *Optimize the coding instruction*

The coding instruction also plays an important role. How specific is the instruction? In the majority of cases, written instructions are used. Such instructions differ in specificity: some only present a description of the task; others contain various examples of the phenomenon under investigation with the risk that coders are biased by those examples. Frequently, coders get the opportunity to ask the researcher for further explanation after they have read the instruction. Then, the coders analyze the materials with the instruction in mind. A risk of this procedure is that coders gradually start leaving out certain analytical steps because of tiredness or subconscious automatic behavior. Possibly, a clearer and more ordered way of instructing the coders and guiding them through the analytical process is to not only let them read a written instruction, but also present the analytical procedure step by step in a decisional flowchart, like Burgers et al. (2011, 2016) and Scholman et al. (2016) did. With the decisional flowchart at hand, coders can follow the analytical process step by step, which prevents tiredness and automaticity.

5.4.3 *Train the coders*

The final factor involving the coding procedure is the degree to which coders are trained. Bayerl and Paul (2011, pp. 713-714) report a strong positive correlation between training and agreement scores, particularly if the intensity of training is high.

Coders can have different degrees of expertise, based on the amount of training and experience. They can be not trained at all, having read the instruction by themselves, they can practice the instruction with a single text, or they can be trained with multiple texts and feedback rounds with the researcher. Coders can also be experts, who have been involved in a particular type of analysis for years and years. In Neuendorf's words (2002, p. 133): "Three words describe good coder preparation: Train, train, and train." The more experience the coders have, the more likely it is that they are doing 'the same' during the actual analysis.

When deploying experts, it is important to check beforehand whether enough coders can be found with a comparable expertise who are unbiased in the sense that they are not aware of the research goal and hypotheses (cf. Artstein & Poesio, 2008, pp. 574-575; Krippendorff, 2019, p. 274). Another issue is that intensive training of coders includes the risk of a coding bias: the trainees code the studied phenomenon the same way because they have learned to do this, which results in a high ICR. It is questionable, though, to what degree this affects the external validity of the study: is the research phenomenon still studied in the analysis, or did the coders learn some kind of superficial 'trick'?

6 Conclusions: Where do we go from here?

For scholars of discourse studies using quantitative content analysis, issues of intercoder reliability are of the highest importance, both for practical reasons (how do we convince our peers that our studies are worthwhile despite our wild data yielding low ICR scores?) and for methodological reasons. In this paper, we have discussed how discourse data can differ in their degree of wildness and – with that – in their risk of a low ICR, how reliability can be at odds with validity, and what to pay attention to when assessing intercoder reliability statistically. We have also demonstrated that intercoder reliability issues are not insurmountable. Implementing the measures as discussed in the previous section will help improve the ICR scores of categorizing predefined units. Since assigning categories to discourse units is an essential part of the discourse coding process, we also believe that these measures will be beneficial to the coding tasks that combine unitizing a continuum and categorizing these units (for example Mathet et al., 2015). Nevertheless, we see important challenges for future work, amongst which big data and multimodal discourse.

6.1 Big data

A first issue is how to maintain insightful analyses when confronted with big data. In present-day corpus-based analyses, the availability of large quantities of discourse data raises all sorts of interesting opportunities compared to small-scale analyses, but also many problems. On the plus side, we have the possibility to look at our phenomena of interest in large numbers of texts, consisting of a wide range of genres, which increases the richness of our analyses and the generalizability of the results. At the same time, the sheer amount of data forces us to complement our manual analyses with automatic procedures, which can lead to ill-informed decisions in comparison to human annotations.

A good example of the problems that automatic analyses can yield is provided by Vis (2011). She wanted to distinguish between words from the journalist and words from quoted sources in a wide variety of news texts from the 1950s and the 2000s. To automate this identification, she used the strategy of searching for quotation marks as the indicator of quoted sources. Although efficient, it is also a very coarse measure for quoted discourse. It neglects all forms of indirect and free indirect speech and writing, and it relies on the systematization with which the journalists made use of quotation marks. Unsurprisingly, such an automated procedure forces the researcher to build in manual checks on the quality of the resulting analysis.

A possible improvement is the use of machine learning. Automatic classification by machine learning can be of assistance in manual coding tasks. Van den Bosch et al. (2006) described the word class or part-of-speech annotation of 50 million words, in which an automatic tagger was applied as a first filtering step. Typical for machine learning, the

tagger combined the classification of a word with a confidence score for each possible part-of-speech tag, and only the words that could be assigned to different part-of-speech tags and stayed under a selected confidence threshold were flagged for manual annotation. The other words were labelled with the automatically assigned tag. This way, the number of units to be annotated manually decreased; Van den Bosch et al. (2006) claim that 28% of the cases in which the automatic part-of-speech tagger was most confident could be safely ignored as correct, and that only 1% of errors were missed. In addition, the confidence scores for different categories, along with information about common mistakes, guided the human annotators in their decision, which increased the ICR. In a similar vein, Reijnierse et al. (in preparation) use machine learning to identify metaphor-related words in a corpus of more than 90,000 Dutch newspaper articles on COVID19. The sheer size of such corpora prevents manual annotation of them.

It should be stressed that such a procedure is especially feasible for low-level linguistic tasks such as part-of-speech tagging, for which most systems offer a sufficiently high base-level of performance and a relatively high reliability of their certainty score (Van den Bosch et al., 2006). Higher-level tasks are more challenging for machine learning. Sarcasm, for example, cannot be inferred from the presence of specific words or word clusters alone. Many markers of sarcasm – such as intensifiers, exclamation points, smileys, quotation marks – are ambiguous or hard to implement in machine learning – such as a change of register. The context is crucial to determine whether "*The weather is fantastic! ;-*)" is meant sarcastically or literally, which makes manual coders indispensable to arrive at reliable categorizations (Kunneman et al., 2015).

That said, recent developments in natural language processing under the header of large language models (more generally referred to as generative AI or foundation models), popularized by systems such as ChatGPT appear to expand the capabilities of previous-generation machine learning methods. With few-shot learning or fine tuning, transformer-based deep learning systems are reported to be successful in many NLP tasks (Vaswani et al., 2018). However, calibrating their self-reported confidence to actual accuracy remains an issue (OpenAI, 2023); moreover, it has been argued that closed and proprietary large language models such as GPT-4 should not be used due to their inherent bias, which is in turn due to the large amount of biased data in their training datasets (of which the details are also kept in the dark). Open academic alternatives are there and are becoming more prevalent and of increasing quality (e.g., the BLOOM[3] family of large language models).

6.2 Multimodal discourse

Another challenge for the near future is the annotation of multimodal discourse. Present-day discourse frequently combines different modes of communication: verbal and visual, static and dynamic. Consider a TV commercial, consisting of text as seen on screen, combined with a voice-over describing the quality of the product, a clip showing a

sequence of events, plus a static depiction of the logo and the product at the end of the commercial. How do we analyze this combination of written language plus visuals plus spoken language plus the interactions between all of these features? We are dealing with the combination of codes that differ fundamentally, in that the verbal code is basically non-iconic, as opposed to the iconic nature of the visuals. Although the study of multimodal discourse is booming (for example Bateman, 2008; Bateman & Hiippala, 2021; Bateman & Wildfeuer, 2014; Mordecai, 2023), attention to the ICR of this multifaceted type of discourse is scarce. An interesting initiative from the Metaphor Lab Amsterdam is to use the existing knowledge of metaphor in verbal language to analyze visual metaphor in their subprojects VisMet (Visual Metaphor, http://www.vismet.org/VisMet/) and CogVim (Cognitive Grounding of Visual Metaphor, https://cogvim.wordpress.com/). The VisMIP (Visual Metaphor Identification Procedure) seeks to identify the metaphorical elements and their relationships in a reliable way (Steen, 2018). Other initiatives are the work by Taboada and Habel (2013) on rhetorical relations in multimodal documents and by Brône and Oben (2015) on gesture annotation. Other than that, there is to our knowledge no methodological work that particularly addresses the intercoder reliability of coding dynamic visuals, let alone the interaction between visuals and verbals, although the development of annotation tools such as ELAN (Sloetjes, 2014) and the MAST (Cardoso & Cohn, 2022) will undoubtedly contribute to such work.

6.3 Concluding remarks

Naturally occurring discourse data are messy. And in some cases, we need to annotate large amounts of data, in a restricted amount of time. It is no wonder researchers engaged in the quantitative corpus analysis of natural discourse sometimes feel they are in one of Augeas' stables, not cleaned in over thirty years. We hope that the suggestions made in this paper contribute to dealing with such messiness and help discourse analysts to tame their wild data.

Acknowledgements

We thank the Centre for Language Studies at Radboud University for letting us organize the workshop on intercoder reliability that inspired this paper. We also thank the reviewers for their valuable feedback.

Author contributions

Renske van Enschot & Wilbert Spooren: Conceptualization, Writing – original draft, Writing – review and editing, Supervision, Project administration. Antal van den Bosch, Christian Burgers, Liesbeth Degand, Jacqueline Evers-Vermeul, Florian Kunneman, Christine Liebrecht, Yvette Linders & Alfons Maes: Conceptualization, Writing – original draft, Writing – review and editing.

Funding information

No funding was received for this study.

Statement of interest

All authors declare that they do not have any competing interests.

Statement of technology use

No AI-based generative technology was used in the preparation of this manuscript and the execution of the research that the manuscript reports upon.

Notes

1 To apply these metrics, we recommend the NLTK toolkit in the framework of the Python programming language, the packages offered in the JASP or JAMOVI environment and in R, and the package written by Andrew Hayes in the framework of SPSS or SAS.

2 Interestingly, Bayerl and Paul (2011, p. 713) suggest that high agreement established by a few number of coders may be due to chance agreement, and consequently they recommend to use a large number of coders when possible: "The higher the number of annotators who are able to agree, the less bias and distortion can be expected". This issue is resolved when using metrics incorporating chance agreement.

3 https://huggingface.co/docs/transformers/model_doc/bloom

References

Arts, A., Maes, A., Noordman, L.G.M., & Jansen, C. (2011). Overspecification in written instruction. *Linguistics*, *49*(3), 555-574. https://doi.org/10.1515/ling.2011.017.

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555-596. https://doi.org/10.1162/coli.07-034-R2.

Bateman, J. (2008). *Multimodality and Genre: A Foundation for the Systematic Analysis of Multimodal Documents*. Springer.

Bateman, J.A., & Hiippala, T. (2021). From data to patterns. In J. Pflaeging, J. Wildfeuer, J.A. Bateman (Eds.), *Empirical Mulitmodality Research. Methods, Evaluations, Implications* (pp. 65-90). De Gruyter. https://doi.org/10.1515/9783110725001-003

Bateman, J.A., & Wildfeuer, J. (2014). A multimodal discourse theory of visual narrative. *Journal of Pragmatics*, *74*, 180-208. https://doi.org/10.1016/j.pragma.2014.10.001

Bayerl, P.S., & Paul, K.I. (2011). What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics*, *37*(4), 699-725. https://doi.org/10.1162/COLI_a_00074

Bolly, C., Crible, L., Degand, L., & Uygur, D. (2014). Towards a model for discourse marker annotation in spoken French: From potential to feature-based discourse markers. In A. Sansó &

C. Fedriani (Eds.), *Pragmatic markers, discourse markers and modal particles: What do we know and where do we go from here?* (pp. 71-97). Benjamins. https://dial.uclouvain.be/pr/boreal/object/boreal:161997

Brône, G., & Oben, B. (2015). InSight Interaction: A multimodal and multifocal dialogue corpus. *Language Resources and Evaluation*, *49*(1), 195-214. https://doi.org/10.1007/s10579-014-9283-2

Burgers, C., Konijn, E.A., & Steen, G.J. (2016). Figurative framing: Shaping public discourse through metaphor, hyperbole, and irony. *Communication Theory*, *26*(4), 410-430. https://doi.org/10.1111/comt.12096

Burgers, C., Van Mulken, M., & Schellens, P.J. (2011). Finding irony: An introduction of the Verbal Irony Procedure (VIP). *Metaphor and Symbol*, *26*(3), 186-205. https://doi.org/10.1080/10926488.2011.583194

Byrt, T., Bishop, J., & Carlin, J.B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, *46*(5), 423-429. https://doi.org/10.1016/0895-4356(93)90018-V

Cardoso, B., & Cohn, N. (2022). The Multimodal Annotation Software Tool (MAST). Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), 6822-6828.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213-220. https://doi.org/10.1037/h0026256

De Smedt, T., & Daelemans, W. (2012). Pattern for python. *The Journal of Machine Learning Research*, *13*, 2063-2067.

Di Eugenio, B., & Glass, M. (2004). The Kappa statistic: A second look. *Computational Linguistics*, *30*(1), 95-101. https://doi.org/10.1162/089120104773633402

Elmes, D.G., Kantowitz, B.H., & Roediger, H.L.I. (2012). *Research Methods in Psychology* (9th ed). Wadsworth Cengage Learning.

Feinstein, A.R., & Cicchetti, D.V. (1990). High agreement but low Kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543-549. https://doi.org/10.1016/0895-4356(90)90158-L

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378-382. https://doi.org/10.1037/h0031619

Forceville, C. (Charles), & Urios-Aparisi, E. (Eds.). (2009). *Multimodal Metaphor*. Mouton de Gruyter.

Fort, K., Nazarenko, A., & Rosset, S. (2012). Modeling the complexity of manual annotation tasks: A grid of analysis. In *Proceedings of the International Conference on Computational Linguistics (COLING 2012)* (pp. 895-910). https://hal.science/hal-00769631

Grove, W.M., Andreasen, N.C., McDonald-Scott, P., Keller, M.B., & Shapiro, R.W. (1981). Reliability studies of psychiatric diagnosis: Theory and practice. *Archives of General Psychiatry*, *38*(4), 408-413. https://doi.org/10.1001/archpsyc.1981.01780290042004

Hancock, J.T., Curry, L.E., Goorha, S., & Woodworth, M. (2007). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, *45*(1), 1-23. https://doi.org/10.1080/01638530701739181

Hayes, A.F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77-89. https://doi.org/10.1080/19312450709336664

Hoek, J., Sanders, T., & Spooren, W. (2021). Automatic coherence analysis of Dutch: Testing the subjectivity hypothesis on a larger scale. *Corpora*, *16*(1), 129-155. https://doi.org/10.3366/cor.2021.0211

Jeni, L.A., Cohn, J.F., & De La Torre, F. (2013). Facing imbalanced data – Recommendations for the use of performance metrics. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 245-251. https://doi.org/10.1109/ACII.2013.47

Krippendorff, K. (2011). Agreement and information in the reliability of coding. *Communication Methods and Measures*, *5*(2), 93-112.

Krippendorff, K. (2019). *Content Analysis: An Introduction to its Methodology* (4th ed.). SAGE.

Krippendorff, K., Mathet, Y., Bouvry, S., & Widlöcher, A. (2016). On the reliability of unitizing textual continua: Further developments. *Quality & Quantity*, *50*(6), 2347-2364. https://doi.org/10.1007/s11135-015-0266-1

Kunneman, F., Liebrecht, C., Van Mulken, M., & Van den Bosch, A. (2015). Signaling sarcasm: From hyperbole to hashtag. *Information Processing & Management*, *51*(4), 500-509. https://doi.org/10.1016/j.ipm.2014.07.006

Liebrecht, C. (2015). *Intens Krachtig. Stilistische Intensiveerders in Evaluatieve Teksten* [*Intensely Powerful. Stylistic Intensifiers in Evaluative Texts.*]. [Doctoral dissertation, Radboud Universiteit]. https://hdl.handle.net/2066/141116

Linders, Y. (2014). *Met Waardering Gelezen. Een Nieuw Analyse-instrument en een Kwantitatieve Analyse van Evaluaties in Nederlandse Literaire Dagbladkritiek, 1955-2005* [*Read with Appreciation. A New Instrument of Analysis and a Quantitative Analysis of Evaluations in Literary Reviews in Dutch Daily Newspapers*]. [Doctoral dissertation, Radboud Universiteit]. https://hdl.handle.net/2066/131544

Martin, L.J., Degand, L., & Simon, A.-C. (2014). Forme et fonction de la périphérie gauche dans un corpus oral multigenres annoté [Form and function of the left periphery in an annotated multi-genre oral corpus]. *Corpus*, *13*. https://doi.org/10.4000/corpus.2509

Mathet, Y., Widlöcher, A., Fort, K., François, C., Galibert, O., Grouin, C., Kahn, J., Rosset, S., & Zweigenbaum, P. (2012). *Manual corpus annotation: Giving meaning to the evaluation metrics*, *809*. https://hal.science/hal-00769639

Mathet, Y., Widlöcher, A., & Métivier, J.-P. (2015). The unified and holistic method Gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, *41*(3), 437-479.

Mol, L., Krahmer, E., Maes, A., & Swerts, M. (2012). Adaptation in gesture: Converging hands or converging minds? *Journal of Memory and Language*, *66*(1), 249-264. https://doi.org/10.1016/j.jml.2011.07.004

Mordecai, C. (2023). #anxiety: A multimodal discourse analysis of narrations of anxiety on TikTok. *Computers and Composition*, *67*, 102763. https://doi.org/10.1016/j.compcom.2023.102763

Moss, P.A. (1994). Can there be validity without reliability? *Educational Researcher*, *23*(2), 5-12. https://doi.org/10.3102/0013189X023002005

Neuendorf, K.A. (2002). *The Content Analysis Guidebook*. SAGE.

OpenAI. (2023). *GPT-4 Technical Report*. https://doi.org/10.48550/ARXIV.2303.08774

Pasma, T. (2011). *Metaphor and register variation: The personalization of Dutch news discourse* [Doctoral dissertation, VU University]. https://research.vu.nl/en/publications/metaphor-and-register-variation-the-personalization-of-dutch-news

Peña, E.D. (2007). Lost in translation: Methodological considerations in cross-cultural research. *Child Development*, *78*(4), 1255-1264. https://doi.org/10.1111/j.1467-8624.2007.01064.x

Perreault, Jr., W.D., & Leigh, L.E. (1989). Reliability of nominal data based on qualitative judgments. *Journal of Marketing Research*, *26*, 135-148.

Potter, W.J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research*, *27*(3), 258-284. https://doi.org/10.1080/00909889909365539

Quiros-Ramirez, M.A., & Onisawa, T. (2015). Considering cross-cultural context in the automatic recognition of emotions. *International Journal of Machine Learning and Cybernetics*, *6*(1), 119-127. https://doi.org/10.1007/s13042-013-0192-2

Reijnierse, G., Grunwald, J., & Spooren, W. (in preparation). *MetRobbert: Automatic metaphor identification in Dutch*.

Scholman, M.C.J., Evers-Vermeul, J., & Sanders, T.J.M. (2016). Categories of coherence relations in discourse annotation. *Dialogue & Discourse*, *7*(2), 2. https://doi.org/10.5087/dad.2016.201

Scott, W.A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, *19*(3), 321-325. https://doi.org/10.1086/266577

Selvi, A.F. (2020). Qualitative content analysis. In H. Rose & J. McKinley (Eds.), *The Routledge Handbook of Research Methods in Applied Linguistics* (pp. 440-452). Routledge.

Sloetjes, H. (2014). ELAN: Multimedia Annotation Application. In J. Durand, U. Gut, & G. Kristoffersen (Eds.), *The Oxford Handbook of Corpus Phonology*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199571932.013.019

Spooren, W., & Degand, L. (2010). *Coding coherence relations: Reliability and validity*. *6*(2), 241-266. https://doi.org/10.1515/cllt.2010.009

Steen, G.J. (Ed.) (2018). *Visual Metaphor: Structure and process*. John Benjamins. https://doi.org/10.1075/celcr.18

Suviranta, R., & Hiippala, T. (2022). *Commercial crowdsourcing in digital humanities: Digital Humanities*, 576-578. https://dh2022.dhii.asia/dh2022bookofabsts.pdf

Taboada, M., & Habel, C. (2013). Rhetorical relations in multimodal documents. *Discourse Studies*, *15*(1), 65-89.

Trochim, W.M.K. (2006). *Reliability & Validity*. https://conjointly.com/kb/reliability-and-validity/

Umesh, U.N., Peterson, R.A., & Sauber, M.H. (1989). Interjudge agreement and the maximum value of Kappa. *Educational and Psychological Measurement*, *49*(4), 835-850. https://doi.org/10.1177/001316448904900407

Van den Bergh, H., Van Es, A., & Spijker, S. (2011). Spelling op verschillende niveaus: Werkwoordspelling aan het eind van de basisschool en het einde van het voortgezet onderwijs [Spelling at different levels: Verb spelling at the end of primary education and at the end of secondary education]. *Levende Talen Tijdschrift*, *12*(1), 3-14.

Van den Bosch, A., Schuurman, I., & Vandeghinste, V. (2006). Transferring PoS-tagging and lemmatization tools from spoken to written Dutch corpus development. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation,* (*LREC-2006*).

Van Enschot, R., & Donné, L. (2013). Retorische vormen in gezondheidsvoorlichting [Rhetorical figures in health communication]. In R.J.U. Boogaart & H. Jansen (Eds.), *Studies in Taalbeheersing 4* (pp. 91-101). Van Gorcum.

Van Enschot, R., & Hoeken, H. (2015). The occurrence and effects of verbal and visual anchoring of tropes on the perceived comprehensibility and liking of TV commercials. *Journal of Advertising*, *44*(1), 25-36. https://doi.org/10.1080/00913367.2014.933688

Van Mulken, M., & Schellens, P.J. (2012). Over loodzware bassen en wapperende broekspijpen – Gebruik en perceptie van taalintensiverende stijlmiddelen [On weighty basses and fluttering pant legs. Use and perception of intensifying stylistic devices]. *Tijdschrift Voor Taalbeheersing*, *34*(1), 26-53. https://doi.org/10.5117/TVT2012.1.OVER418

Van Rijsbergen, C.J. (1979). *Information Retrieval* (2nd ed). Butterworths.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2018). Attention Is All You Need. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, R. Fergus, S.V.N. Vishwanathan, & R. Garnett (Red.), *Advances in neural information processing systems 30: 31st Annual Conference on Neural Information Processing Systems* (*NIPS 2017*)*: Long Beach, California, USA, 4-9 December 2017* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Vis, K. (2011). *Subjectivity in News Discourse: A Corpus Linguistic Analysis of Informalization* [Doctoral dissertation, Vrije Universiteit Amsterdam]. https://research.vu.nl/en/publications/subjectivity-in-news-discourse-a-corpus-linguistic-analysis-of-in

Wallace, B.C. (2015). Computational irony: A survey and new perspectives. *Artificial Intelligence Review*, *43*(4), 467-483. https://doi.org/10.1007/s10462-012-9392-5

Webber, B., Prasad, R., Lee, A., & Joshi, A. (2019). *The Penn Discourse Treebank 3.0 Annotation Manual*. https://catalog.ldc.upenn.edu/docs/LDC2019T05/PDTB3-Annotation-Manual.pdf

Zufferey, S., & Degand, L. (2013). Annotating the meaning of discourse connectives in multilingual corpora. *Corpus Linguistics and Linguistic Theory*, *13*(2), 399-422. https://doi.org/10.1515/cllt-2013-0022