Taming our wild data: On intercoder reliability in discourse research

Renske van Enschot, Wilbert Spooren, Antal van den Bosch, Christian Burgers, Liesbeth Degand, Jacqueline Evers-Vermeul, Florian Kunneman, Christine Liebrecht, Yvette Linders & Alfons Maes

PRELIMINARY VERSION, DO NOT QUOTE

Version: 5-1-2016

ABSTRACT

Many research questions in the field of linguistics, communication and cognition are answered by manually analyzing data collections or corpora: collections of (transcribed) spoken, written or visual communicative messages. In this kind of quantitative content analysis of discourse the coding of subjective language data often leads to disagreement among raters. In this paper we discuss causes of and solutions for disagreement problems in the analysis of discourse. We discuss the effects of three sources of difficulty in coding discourse variables. We discuss the sometimes tense relation between reliability and validity. We describe the advantages and disadvantages of using a popular formal assessment of intercoder reliability, namely Cohen's Kappa, and some of its alternatives. We suggest a number of ways to improve the reliability, such as the precise specification and carving up the coding process into smaller substeps. The paper ends with a reflection on challenges for future work in discourse analysis, with a special attention to big data and multimodal discourse.

KEYWORDS intercoder reliability, discourse, quantitative content analysis

1. Introduction

Many research questions in the field of communication, linguistics and cognition are answered by manually analyzing data collections or corpora: collections of (transcribed) spoken, written or visual communicative messages. Although many different forms of corpus analysis are used (Krippendorff, 2013), the generic base may be defined as assigning interpretative levels 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 a metaphoric poster or advertisement can be classified as ‘matching the intended meaning’ or not; a relationship between two utterances in a discourse can be labeled as semantic or pragmatic, or with one label out of a list of 25; et cetera.

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 (Spooren & Degand, 2010). Disagreement can occur even when there has been an extensive training phase, even when an explicit code book is used that has been tested and adapted, even when the number of coding categories is limited, and even when experts are used instead of naive and untrained coders (Spooren & Degand, 2010). Often, several rounds of coding are necessary to reach a sufficiently high intercoder reliability statistic such as Cohen's kappa. Problems increase for the analysis of static and dynamic visuals in discourse. At the same time our publication outlets require that such a kappa is reached after only one round of coding and that only naive coders are employed. Allegedly only then the variables would be sufficiently concrete and the categorizations could be considered replicable and valid.

One may prevent low intercoder agreement results by suggesting researchers to study only clear-cut variables. They will be less stubborn than explorative variables to be detected in randomly selected data produced in uncontrolled conditions. Interrater agreement tests are superfluous when variable levels can be assigned without any interpretation noise. However, too many interesting questions in the field of human communication are not ready for such types of controlled research. Examples of such intriguing questions are: What makes a visual message metaphorical? Which linguistic or audiovisual cues can we consider to be deceptive? The second question is addressed by Hancock, Curry, Goorha and Woodworth (2007) using elicited data.

In this paper we first sketch the scope of the problem by describing different degrees of messiness in discourse data. We will then address how validity is at stake as well and describe the tension between reliability and validity. We will then move on to 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. The paper continues with hands-on advice on how to improve ICR and concludes with a reflection on where to go from here, discussing big data and multimodality in particular.

2. Different degrees of messiness in discourse data

The quality and outcome of corpus analysis, as well as the success of interrater agreement tests, is determined by a large number of conditions. In this section, we will discuss the effects of three factors: different ways of collecting data, the ways in which coding categories are established, and the number of discourse levels that are taken into account.

First, data collections can be elicited experimentally or selected randomly in the field, for example by collecting the materials by random or representative sampling. 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 the 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 variables 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, van Mulken & Schellens, 2011). On the other hand, coding variables can start from an explorative intuition and emerge gradually as the analysis proceeds (van Enschot & Donné, 2013). The coding of the latter type of data tends to be much more difficult than when the coding variables and their levels have been defined in a precise way. Agreement tests can be useful in both kinds of studies. Controlled studies require a high degree of precision and validity, and consequently only high levels of ICR outcomes are acceptable. In explorative studies, ICR scores can be used as a heuristic tool to objectify or specify individual intuitions, to try out coding level definitions or segmentation options in data collections. In this case, lower ICR rates are acceptable, although one may want to perform an additional more controlled analysis to validate the new coding system.

Third, coding categories variables can consist of a more or less closed set of levels. On the one extreme, the variable levels are a dyadic, mutually exclusive, closed set (e.g., yes-no, high-low, figurative-literal, etc.). In case of such closed variables, it is relatively easy to determine conditions with near-guaranteed agreement success: a small number of levels attached to one variable, clearly defined in terms of objective characteristics. An ICR score is hardly relevant in these cases. On the other extreme, the amount of levels is not fixed: There are different ways to intensify an utterance: from typographical elements and word parts to multiple words and syntactic constructions, or the levels are not mutually exclusive. Such variables leave room for an exploratory analysis but at the same time pose problems for ICR. Extra complicating are the cases in which the unit of analysis is underspecified, e.g., when both a lexical item and the sentence in which it is included can be categorized as intensifiers.

The above suggests that analytical data can be less or more messy, and that the degree to which our data are messy depends on the choices made by the researcher. The researcher makes decisions about the type of questions asked (e.g., does she focus on the use of the conjunction ‘because’ or on ‘epistemic cues’ in discourse), about the collection method used (e.g., does she 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 recognition task), about the theoretical framework and position taken (e.g., does she start from the position that all gestures can be divided in two or three major categories or not), and about the way in which theory is translated into definitions of coding levels (e.g., does she define epistemic cues as a closed set of identifiable elements in discourse, or as an open ended class).

These choices in turn determine to a large extent the success of exercises in which different individuals are asked to do the same coding job in order to obtain a satisfactory level of interrater agreement. Data can vary. Once we agree that data can vary to a very large extent, the

question is which measures can be taken to make interrater agreement exercises doable, not trivial, and successful or at least informative. We will address this issue in section 5.

3. Reliability versus validity

Quality of data is not only a matter of the reliability of the data but also of their validity. Increasing reliability means reducing the level of random (coding) errors. Increasing validity refers to a reduction at the level of systematic errors, and hence to a more accurate reflection of reality. Validity and reliability can be at odds with each other. Aiming for high intercoder reliability scores is not a guarantee for good validity and may even yield serious problems for the construct, external and internal validity of the research.

Construct validity. Construct validity refers to the question whether the coded data accurately reflect the theoretical constructs they are supposed to measure (Elmes, Kantowitz, & Roediger III, 2011, pp. 185-187). In a recent paper, Wallace (2015) contests the way in which various computational linguists have operationalized irony and sarcasm. Many operationalizations aimed to automatically code sarcasm are based on 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 (e.g., *yeah right*). Wallace (2015) argues that such automatic identification procedures based on word usage are "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 "Barack Obama is a great president" is likely to be literal when said by a supporter of the Democratic Party, and sarcastic when said by a supporter of the Tea Party. 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 when identification procedures achieve satisfactory or high levels of reliability – coding instances of "Yeah, right" in a corpus can easily be done very reliably– it is important to critically analyze whether specific examples of the variable of interest are not systematically excluded or overrepresented.

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, and is particularly relevant for corpus-based studies in which textual features in two or more (sub)corpora are compared. 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, van Es and Spijker (2011, p. 6) point out, analyzing these text types can be done in a very reliable way (e.g., they report 100% intercoder agreement 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). The internal validity is at stake in this analysis as well. First, the number of errors in students' own writing may (also) be determined 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 a dictation, 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.

The above shows that reliability is not a sufficient condition for validity. A traditional viewpoint is that reliability is at least a *necessary* condition for validity (Moss, 1994). An interesting issue is whether this viewpoint is tenable, i.e., whether we can imagine research that is valid but unreliable (Moss, 1994). The issue is even more pressing given that we often find it difficult to establish the reliability of our codings. If our reliability scores are lagging behind, can we still establish validity (cf. van Enschot & Hoeken, 2015)? Should the answer to this question be negative, we anticipate insurmountable problems for our discipline. A possible viewpoint is that the theory or the coding procedure yielding the analysis of such unreliable data is 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 our 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 the 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. She only draws firm conclusions of all three sets of results point in the same direction.

4. How to deal with Cohen's kappa

In this section, we describe a selection of settings in which the most widespread metric, Cohen's Kappa, can be misleading: when the selection of annotators is alternated, the levels are in an ordering relation, or the annotation of levels is highly imbalanced. 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 alternative metrics, without providing raw formulas. To apply

these metrics, we recommend the package written by Andrew Hayes[1] in the framework of SPSS or SAS, or the NLTK toolkit[2] in the framework of the Python programming language.

Cohen's Kappa was proposed by Cohen (1960), and takes into account the prior chance that two annotators agree on the annotation of any level. This makes Cohen's Kappa a more realistic metric for interrater agreement than percentage agreement, which can easily give misleading insights. For instance, one study using two coding levels has a fifty percent chance that coders agree whilst another study using four coding levels yields a 25 percent agreement chance. As a result, the first study will probably yield higher ICR scores than the second study simply due to chance (Artstein & Poesio, 2008, pp. 558-559). Still, the prior chance of any two annotators to agree is not the only factor that might influence agreement. In this sense, Cohen's Kappa has its own biased focus on reliability. Artstein and Poesio (2008) describe two biases of Cohen's Kappa: the annotator bias -- the case where annotators prefer using different levels of a variable to be coded -- and the prevalence bias -- the case where one level of a variable is used much more than the other. Both lead to different and invalid estimates of the 'true' reliability.

Perreault and Leigh (1989, p. 146) state that "...different indices reflect different aspects of reliability". To give a complete insight of any interrater agreement, it is therefore valuable to show experimental outcomes with other reliability metrics in addition to, or in some cases as replacement of, Cohen's Kappa.

Cohen's Kappa presumes all items in a set to be coded by the same two annotators. The metric does not take into account settings in which the items are annotated by more than two annotators and/or different constellations of annotator sets. A better option in such a context is to use Fleiss' Kappa (Fleiss, 1971). Unlike Cohen's Kappa, which takes into account the answers of any specific *coder*, Fleiss' K is calculated based on the proportion of times that each *category* is chosen by annotators, as well as the agreement per single *item*. These two components are used to calculate the chance of agreement. By focusing on single items rather than the whole of items per annotator, Fleiss' K allows the items to be annotated by any combination of annotators, as long as the number of annotators per item remains the same.

An important property of the coding task is the scale of the variable(s). In standard form, the Cohen's Kappa metric presumes a nominal scale, in which no ordering exists between the levels. It will give wrong insights when applied to other than nominal variables, with ordinal, interval or ratio levels. With these kinds of variables, it should be penalized when two coders annotate levels that have a larger distance to one another. Metrics that do apply such a penalization are the Weighted Cohen's Kappa (Gwet, 2010, pp. 34-36) and Krippendorff's Alpha (Hayes & Krippendorff, 2007; Krippendorff, 2013). 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 levels, such that more distanced levels add a lot less to the

[1] Available at http://www.afhayes.com
[2] http://www.nltk.org/_modules/nltk/metrics/agreement.html

agreement score (or a lot more to the disagreement score). These two metrics also allow for missing data points, by taking into account the total number of annotations that were made.

Another factor that needs to be taken into account when assessing interrater agreement is data skew: the degree to which data are annotated as belonging to the same levels. Jeni, Cohn and De La Torre (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 to make similar annotations is high when there is a dominant class. Consequently, the percentage agreement is subtracted by a higher number and any disagreement has a large effect. A solution is to calculate the Kappa max (Umesh, Peterson & Sauber, 1989), which returns a kappa value that is relative to the upper bound of the kappa that follows the strict chance agreement.

The Mutual F-score is another metric that can be used to conceal the influence of class imbalance. It applies to the agreement about *specific* levels rather than the overall agreement. Mutual F-score is based on F1, which is often used in Information Retrieval and Machine Learning for system evaluation (van Rijsbergen, 1979). When focusing on the annotations of one level, we can regard the annotations of a first annotator as ground truth and the annotations of a second annotator as output of the system. The F1 score evaluates the agreement of the second annotator with the first annotator in terms of recall and precision. The mutual F-score first regards annotator 1 and annotator 2 as ground truth and system output (recall), and then the other way around (precision). It can be typically calculated for the predominant level, and is especially useful for comparing the agreement for multiple datasets.

A number of factors can influence the outcome of any metric to assess interrater agreement. In section 2, we suggested that the ICR depends on the nature of the data and the research question. This suggestion would imply that an interpretation of, for example, Cohen's Kappa should be used relative to the research question. Instead of following Landis and Koch's (1977) proposal to consider kappa's .41< kappa < .60 as moderate and kappa's .61 < kappa < .80 as substantial, interpretation may vary per type of study: for hypothesis testing kappa's >.61 are required, whereas for explorative studies lower kappa's down to .41 can be sufficient. Similar suggestions can be found in Grove et al. (1981) and Spooren and Degand (2010).

In light of these examples, it is important to fully understand the mechanisms of a metric when interpreting its outcomes. Furthermore, existing heuristics to interpret the outcomes, that do not take into account the context of annotation, seem too simplistic. We advise to assess interrater agreement with several metrics, so as to achieve a more complete interpretation of the factors that are in play.

5. How to improve 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 reliability.

5.1. Specify the unit of analysis

A coder can be asked to code predefined units (words, sentences, pictures, audio fragments, etc.), consider them as indivisible, and generate one code for each unit. Again, this situation makes agreement tests easily successful. For example, Pasma (2011) uses the word as a unit of analysis that is coded as being metaphorical or not, with impressively high ICRs as a result.

However, many research topics are embedded in larger contexts (e.g., words and sentences in discourse, turns in conversations, objects in visual scenes, etc.), and the units of analysis can differ (e.g., the valence of a conversation contribution can be defined based on a complete conversation turn, on clauses in one turn, or on words in one clause). A case in point is the study by van Enschot and Hoeken (2015) in which the unit of analysis is the entire TV commercial, without any further specification; unsurprisingly, the ICRs started off low, and increased only after a second round of coding.

In case of hypothesis testing specific units of analysis are preferable. But during an explorative phase, it may well be useful to leave it to the coders to determine which unit of analysis is most appropriate. Such a first coding and agreement exercise may provide more insight into the way in which a more controlled agreement test should deal with the presentation of and instruction about units of analysis.

5.2. Make your categories independent

Agreement success is highly determined by the relation between coding levels. Two major conditions are relevant here: one is whether the coding levels are mutually exclusive or not, the other is whether coding levels are hierarchically ordered.

The same unit of analysis can have different functions or interpretations, which may result in units belonging to more than one of the coding levels (e.g., clauses can have different relations with other clauses) or in scalar levels. Obviously, agreement success is easier when coding levels are mutually exclusive.

For example, in a study of the subjectivity of adjectives preceding or following causal connectives Hendrickx and Spooren (in preparation) used the subjectivity scores on a continuous 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 (subjectivity score <. 20) and subjective adjectives (subjectivity score > .70). The other adjectives were considered ambiguous and therefore excluded from the analysis.

The same holds for hierarchical variables, with which coders are first asked to determine major classes and then to subclassify units within the assigned level. An example is coding

discourse relations first in terms of semantic vs. pragmatic relations, and then within the assigned level the exact relation type. Again, agreement success is endangered when coders have to apply such embedded coding tasks. Splitting up the agreement tasks is an easy solution in such situations. For instance, Zufferey and Degand (in press) report percentage agreements of three types of multilingual discourse relation annotation differing in the amount of specificity. For the least specific type of annotation, i.e., distinguishing between four types of differentiated discourse relations (*temporal, comparison, contingency, expansion,* cf. PDTB Research Group, 2007), agreement is highest, above 90%. For the second type, which subdivides, for example, contingency into conditional and causal, agreement drops to 60-72%. The third, most specific, type yields agreement percentages between 39% and 53%. Part of the disagreement concerning the second and third type is caused by disagreements concerning the first type, because decisions regarding this first type directly impact decisions that have to be made for the second and third type. An example is (1), in which the relation conveyed by *when* could arguably be either temporal or conditional. Disagreement regarding this first type automatically induces disagreement regarding the second and third type because the available decision features 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.

A way of circumventing the problem of combined variables is by asking the coders to decide for each option or level in the coding system whether it applies or not. By doing so, chances become very small that they code a case with the first level that comes to mind while ignoring other relevant levels. In the above case, disagreement regarding the more specific types 2 and 3 would not appear because some of the options have become non-applicable.

5.3 Reduce the number of coding levels

Reducing the number of levels in a coding system might improve intercoder agreement, but usually is undesirable because a reduced coding system yields less information. An obvious first check is whether all levels are really needed. For example, van Enschot and Hoeken (2015) originally had two levels in their analysis of tropes - a subcategory 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 were therefore combined in the final phase of the analysis, resulting in higher ICRs.

An advantage of reducing the number of levels per variable is that the levels 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, Crible, Degand & Uygur-Distexhe,

forthcoming). This class of linguistic expressions is very heterogeneous, consisting mostly of coordinate and subordinate conjunctions such as *but* and *because* and 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 levels some of which occur infrequently. Depending on the general theory and the research question at hand, coders can question whether it is useful to keep all possible syntactic categories or whether they should group some of them. 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 (e.g., adverbials, conjunctions and prepositional phrases) and group all other possibilities in one encompassing "other" class, or even retain only two coding choices (e.g., between 'conjunctive' and 'non-conjunctive'). Some of these options are illustrated in Table 1.

Table 1. 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	
coord. conj.	other	
subord. conj.		
adjective		
preposition		
prep. phrase		
noun		
interjection		

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 1 (ten levels), an equal distribution of all levels would lead to a maximum of five occurrences per level. Now, knowing that adjectives or nouns used as discourse markers are very rare in English, it is highly probable that these levels will receive zero counts. This may lead to biases in the statistical analysis. Therefore, either the sample has to be increased to account for rare events, or the number of levels has to be reduced, or a statistical

measure such as Kappa Max should be used that is sensitive to uneven distributions (see section 4). A simpler coding schema such as that in option 3 of Table 1, with only two levels, simplifies both the coding decisions and the statistical analysis.

5.4 Decompose the process of analysis in smaller steps

If reducing the number of levels is not possible without losing too much information, an interesting alternative is to decompose the analytical process into smaller, simpler steps, by dividing the coding system into several steps. Thus, instead of reducing the number of levels to be coded, one can simplify the coding decisions by increasing the number of coding steps, while at the same time reducing the number of levels that need to be considered during each step. The net result is that the same number of coding levels will be considered. The main advantage of this procedure is that the decision process is split up into smaller decision trees.

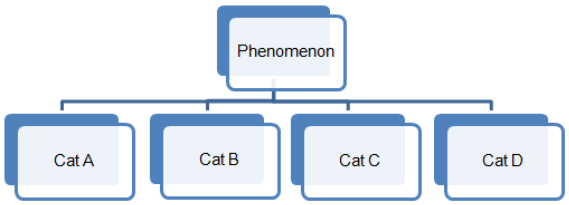

Figure 1a. Schematic process of analysis of a complex category.

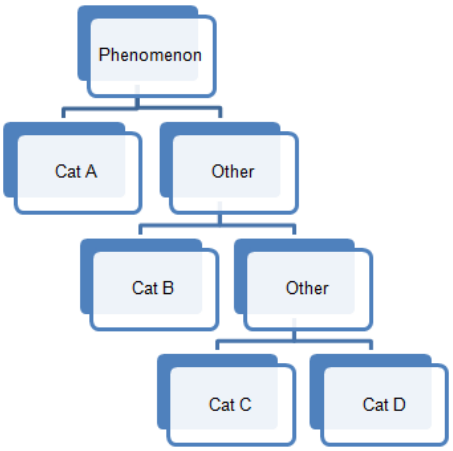

Figure 1b. Simplified version of the analysis in Figure 1a.

For example, in a research project on literary criticism, coders had to indicate what aspects (such as *style* and *structure*) and characteristics (*efficiency* or *clarity*) of novels were being evaluated by critics. For each evaluative statement they had to choose which of the fifteen listed types of characteristics applied. Characteristics varied from *efficiency*, to *emotiveness* to *religious value* (Linders, 2014). Instead of making coders choose one of these fifteen options, they could have been confronted with a number of decisions: the first step might have been deciding whether the statement was about the book itself, the effect the book had on the reader, or the book in relation to the world. The next step would then be to decide which specific subcategory within this larger category applied. If they had coded the evaluation as being a statement about the effect the novel had on the reader, they would have been asked to then choose from a limited number of

characteristics that belong to the category *characteristics about the effect on the reader*, i.e. *humor*, *emotiveness* and *didactic value*. This would have narrowed down the number of options and structures they were coding.

Decomposing looks like a promising strategy. However, it also has some disadvantages. One is that splitting up analyses into smaller steps may be more time-consuming than a more straight-forward coding procedure. Another problem is that decomposing an analysis into smaller steps may lead to an inaccurate estimate of reliability. Suppose that reliability scores are calculated for the most specific levels (such as whether an evaluation in a book review is about *humor*, *emotiveness*, or *didactic value*). To obtain these scores only the cases are used in which the coders already agreed at more generally (i.e., they agreed that evaluation was about the effect the novel had on the reader). The result is that, while agreement for more specific levels is high, the picture is incomplete because this high agreement score does not reflect the difficulties at the more general coding levels. An obvious recommendation is to report agreement scores for all steps.

The conclusion is that decomposing an analysis into several smaller steps should only be used to guide coders through the analysis and, by doing this, to improve reliability, not to enhance the intercoder agreement scores without actually improving the reliability itself.

5.5 Procedural measures

A number of elements have to be taken into account to facilitate the coding the procedure, among these are, at least, the number of coders involved, the instructions given to the coders, and their training.

Consider the number of coders. Two or three coders are standard in most studies in the field of communication, linguistics and cognition (e.g. Kunneman, Liebrecht, van Mulken, & van den Bosch, 2014; Phillips & McQuarrie, 2002; Renkema, 1997; van Enschot & Donné, 2013; van Mulken & Schellens, 2006; 2012). When coding is relatively simple (a few coding levels to be assigned in well-defined units) one additional coder who recodes part of the data is considered sufficient (e.g., Mol, Krahmer, Maes, & Swerts, 2012). When data are more messy or levels more diffuse, coding by two 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 seems to be more reliable (Potter & Levine-Donnerstein, 1999), but practical considerations make two or three coders reasonable.

Optimize the coding instruction. The coding instruction also plays an important role. How specific is the instruction? In the majority of cases, a written instruction is used. The 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 more clear and 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 to present the analytical procedure step by step in a decisional flowchart, like Burgers et al. (2011) did for example. With the decisional flowchart at hand, coders can follow the analytical process step by step, which prevents tiredness and automaticity.

Train the coders. The final factor involving the coding procedure, is the degree to which the coders are trained. Coders can be not trained at all - if they only read the instruction by themselves -, they can practise the instruction with a single text, or they can be trained with multiple texts and feedback rounds with the researcher. The more trained coders are, the more likely it is that they are doing 'the same' during the actual analysis. However, intensive training of coders includes the risk of a coding bias: they code the studied phenomenon the same way because they learned to do this which results in a high internal validity. 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'? This relates to Potter and Levine-Donnerstein's (1999) remark that there should be projective content, i.e. some room for the coders' own interpretations. In such an approach, the goal of the coder training should be to recognize the phenomenon and to analyze the materials based on their own interpretation. Of course, room for own interpretations may have a negative influence on the interrater agreement.

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 low ICRs?) and for methodological reasons. In this paper, we have shown that intercoder reliability issues are not insurmountable. Nevertheless, we see important challenges for future work.

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 groups 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 automatize 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 systematicity 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 helpful for some coding tasks. For example, van den Bosch, Schuurman and Vandeghinste (2006) describe the word class or part-of-speech annotation of 50 million words. Rather than manually annotating all words, which would take a very long time, an automatic tagger was applied as a first filtering step. The tagger combined the classification of a word with a certainty score for each possible part-of-speech tag, and only the words that might be assigned to different part-of-speech tags and surpass a selected certainty threshold were extracted for manual annotation. The other words were labeled with the automatically assigned tag. This way, the number of units to be annotated manually decreases drastically. In addition, the certainty scores for different categories, along with information about common mistakes, guides the human annotators in their decision, which increases the ICR.

Automated analysis can thus help when a coding task encompasses a large dataset. Because an automated system, such as a machine learning classifier, often lacks the analytic skills of a human expert, part of the data will still need to be corrected on the basis of manual annotation. The automated system can, however, provide certainty scores for its decisions, including the certainty for categories that were not chosen. These certainty scores both help to select the data for manual annotation, typically the uncertain and ambiguous ones, and to provide the human annotators with additional context to make their decision. It should be stressed that such a procedure is especially feasible for tasks that do not require a lot of world knowledge.

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 (e.g., Bateman, 2011; Bateman & Wildfeuer, 2014; Forceville & Urios-Aparisi, 2009; Jewitt, 2009; Kress, 2010;

Royce & Bowcher, 2007), attention to the ICR of this multifaceted type of discourse is scarce. An interesting initiative to use the existing knowledge of metaphor in verbal language to analyze visual metaphor are the Metaphor Lab Amsterdam's subprojects VisMet (Visual Metaphor, vismet.org) and CogVim (Cognitive Grounding of Visual Metaphor, cogvim.org). The VisMIP (Visual Metaphor Identification Procedure) seeks to identify the metaphorical elements and their relationships in a reliable way. Other initiatives are the work by Taboada et al. (2013) on rhetorical relations in multimodal documents and by Brone et al. on gesture annotation. Other than that, there is to our knowledge no methodological work that particularly addresses the reliability of coding dynamic visuals, let alone the interaction between visuals and verbals.

Naturally occurring discourse data are messy. It is no wonder researchers engaged in the quantitative corpus analysis of natural discourse sometimes feel they are in one of Augeias' stables, not cleaned in over thirty years. We hope that the suggestions made in this paper contribute to dealing with that messiness and help discourse analysts to tame their wild data.

References

ARTS, A., MAES, A., NOORDMAN, L.G.M., JANSEN, C. 2011. Overspecification in written instruction. *Linguistics, 49*(3), 555-574.

BATEMAN, J. (2011). *Multimodality and genre. A foundation for the systematic analysis of multimodal documents.* London, New York: Palgrave Macmillan.

BATEMAN, J. & WILDFEUER, J. (2014). A multimodal discourse theory of visual narrative. *Journal of Pragmatics, 74*, 180-208.

BIBER, D., CONRAD, S., & REPPEN, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

BOLLY, C., CRIBLE, L., DEGAND, L., & UYGUR-DISTEXHE, D. (forthc.). Towards a Model for Discourse Marker Annotation in spoken French: From potential to feature-based discourse markers. In C. Fedriani & A. Sansó (Eds.), *Discourse Markers, Pragmatic Markers and Modal Particles: New Perspectives*. Amsterdam: Benjamins.

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.

CLARIDGE, C., & WILSON, A. (2002). Style evolution in the English sermon. In T. Fanego, B. Mendez-Naya, & E. Seoane (Eds.), *Sounds, words, texts, and change: Selected*

papers from 11 ICEHL, Santiago de Compostela, 7-11 September 2000. Volume 2 (pp. 25-44). Amsterdam/Philadelphia: John Benjamins.

COHEN, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), pp. 37–46.

DE SMEDT, T. & DAELEMANS, W. (2012). Pattern for Python. *Journal of Machine Learning Research, 13*, 2063-2067.

ELMES, D.G., KANTOWITZ, B.H., & ROEDIGER III, H.L. (2011). *Research methods in psychology* (9th Ed.). Belmont, USA: Wadsworth.

EVERS-VERMEUL, J., DEGAND, L., FAGARD, B., & MORTIER, L. (2011). Historical and comparative perspectives on subjectification: A corpus-based analysis of Dutch and French causal connectives. *Linguistics*, *49*(2), 445-478.

FLEISS, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378-382.

FORCEVILLE, C. & URIOS-APARISI, E. (Eds.) (2009). *Multimodal Metaphor*. Berlin, Boston: Mouton de Gruyter.

FRASER, B. (1999). What are discourse markers? *Journal of Pragmatics, 31*(7), 931-952.

GRIES, S.Th. (2006). Exploring variability within and between corpora: some methodological considerations. *Corpora*, *1*(2), 109-151.

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*, pp. 408–413.

GWET, K.L. (2010). *Handbook of Inter-rater Reliability* (2nd Ed.). Gaithersburg, MD: Advanced Analytics.

HAYES, A. & KRIPPENDORFF, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*(1), 77-89, DOI: 10.1080/19312450709336664.

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, 4*5(1), 1-23.

HENDRICKX, I. & SPOOREN, W. (In Preparation). Beyond manual analyses of discourse coherence.

HERRING, S.C., VAN REENEN, P.Th., & SCHOSLER, L. (2000). On textual parameters and older languages. In S.C. Herring, P.Th. van Reenen, & L.Schøsler (Eds.), *Textual parameters and older languages* (pp. 1-31). Amsterdam/Philadelphia: Benjamins.

JENI, L.A., COHN, J.F., & DE LA TORRE, F. (2013, September). Facing imbalanced data-- Recommendations for the use of performance metrics. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference* (pp. 245-251). IEEE.

JEWITT, C. (Ed.) (2009). *The Routledge handbook of multimodal analysis*. London: Routledge.

LANDIS, J. R., & KOCH, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

KRESS, G. (2010). *Multimodality: A social semiotic approach to contemporary communication*. London: Routledge.

KRIPPENDORFF, K. (2013). *Content analysis: An introduction to its methodology* (3rd Ed.). Thousand Oaks, CA (USA): Sage.

KUNNEMAN, F., LIEBRECHT, C., VAN MULKEN, M., & VAN DEN BOSCH, A. (2014). Signaling sarcasm: From hyperbole to hashtag. *Information Processing and Management*, dx.doi.org/10.1016/j.ipm.2014.07.006.

LEWIS, D. (2006). Discourse markers in English: A discourse-pragmatic view. In K. Fischer (Ed.), *Approaches to discourse particles* (pp. 43-59). Amsterdam: Elsevier.

LIEBRECHT, C. (2015). *Intens krachtig. Stilistische intensiveerders in evaluatieve teksten [Intensely powerful. Stylistic intensifiers in evaluative texts.]*. PhD Dissertation, Radboud University Nijmegen.

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 reviewsin Dutch daily newspapers]*. PhD Dissertation, Radboud University Nijmegen.

MOL, L., KRAHMER, E., MAES, A., & SWERTS, M. (2012). Adaptation in gesture: Converging hands or converging minds? *Journal of Memory and Language, 66*, 249-264.

PADILLA-WALKER, L.M., COYNE, S.M., FRASER, A.M., & STOCKDALE, L.A. (2013). Is Disney the nicest place on earth? A content analysis of prosocial behavior in animated Disney films. *Journal of Communication, 63*(2), 393-412.

THE PDTB RESEARCH GROUP (2007). The Penn Discourse Treebank 2.0 Annotation Manual. IRCS Technical Reports Series, 99p.

PASMA, T. (2011). *Metaphor and Register Variation: The Personalization of Dutch News Discourse*. PhD Dissertation, VU University Amsterdam.

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

PHILLIPS, B.J., & MCQUARRIE, E.F. (2002). The development, change, and transformation of rhetorical style in magazine advertisements 1954-1999. *Journal of Advertising, 31*(4), 1-13.

ROYCE, T.D. & BOWCHER, W.L. (Eds.) (2007). *New directions in the analysis of multimodal discourse*. Mahwah, NJ: Lawrence Erlbaum.

SCHOLMAN, M.C.J., EVERS-VERMEUL, J., & SANDERS, T.J.M. (submitted), A step-wise approach to discourse annotation: Towards a reliable categorization of coherence relations. Submitted to Dialogue & Discourse.

SPOOREN, W. & DEGAND, L. (2010). Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory*, *6*(2), 241-266. DOI 10.1515/cllt.2010.009

POTTER, W.J., & LEVINE-DONNERSTEIN, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research, 27*, 258-284.

RENKEMA, J. (1997). Geïntensiveerd taalgebruik: een analyseschema [Intensified language: A scheme of analysis]. In H. van den Bergh, D. Janssen, N. Bertens, & M. Damen (Eds.), *Taalgebruik Ontrafeld* (pp. 495-505). Dordrecht: Foris.

TABOADA, M. & HABEL, C. (2013). Rhetorical relations in multimodal documents. *Discourse Studies, 15*(1), 59-85.

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.

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 *Proc. of the Fifth International Conference on Language Resources and Evaluation, LREC-2006*.

VAN ENSCHOT, R. & DONNE, L. (2013). Retorische vormen in gezondheidsvoorlichting [Rhetorical figures in health communication]. In R. Boogaert & H. Jansen (Eds.), *Studies in Taalbeheersing 4* (pp. 91-101). Assen: 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, 24*(1), 25-36. doi: 10.1080/00913367.2014.933688

VAN MULKEN, M., & SCHELLENS, P.J. (2006). Overtuigend? Een stilistische analyse van persuasieve teksten [Persuasive? A stylistic analysis of persuasive texts]. In H. Hoeken, B. Hendriks & P. J. Schellens (Red.), *Studies in Taalbeheersing 2* (Vol. 2). Assen: Van Gorcum.

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), 28-55.

VAN RIJSBERGEN, C.J. (1979). *Information retrieval* (2nd Ed.). London: Butterworths.

VIS, K. (2011). *Subjectivity in news discourse: A corpus linguistic analysis of informalization*. PhD Dissertation VU University Amsterdam.

WALLACE, B.C. *(*2015*).* Computational irony: A survey and new perspectives. *Artificial Intelligence Review, 43*(4), 467-483.

ZUFFEREY, S. & DEGAND, L. (In Press). Annotating the meaning of discourse connectives in multilingual corpora. *Corpus Linguistics and Linguistic Theory*, 1-24. Available online at http://www.degruyter.com/view/j/cllt.ahead-of-print/cllt-2013-0022/cllt-2013-0022.xml.