Methodology for Reliable Schema Development and Evaluation of Manual Annotations

Petra S. Bayerl
Applied and Computational Linguistics
Otto-Behaghel-Strasse 10D, 35394 Giessen
Justus-Liebig-University, Germany
Petra.S.Bayerl@psychol.uni-giessen.de

Harald Lüngen
Applied and Computational Linguistics
Otto-Behaghel-Strasse 10D, 35394 Giessen
Justus-Liebig-University, Germany
Harald.Luengen@germanistik.uni-giessen.de

Ulrike Gut
Department of English
Fahnenbergplatz, 79085 Freiburg
Albert-Ludwigs-University, Germany
ulrike.gut@anglistik.uni-freiburg.de

Karsten I. Paul
Organizational and Social Psychology
Lange Gasse 20, 90403 Nürnberg
Friedrich-Alexander-University, Germany
Paul.Karsten@wiso.uni-erlangen.de

ABSTRACT

The quality of manual annotations of linguistic data depends on the use of reliable coding schemas as well as on the ability of human annotators to handle them appropriately. As is well known from a wide range of previous experiences annotations using highly complex coding schemas often lead to unacceptable annotation quality. Reducing complexity might make schemas easier to handle, but in this way valuable information needed for more sophisticated applications is excluded as well. In order to deal with this problem, we developed a systematic approach to schema development, which allows for developing coding schemas for fine-grained semantic annotations while systematically securing the quality of such annotations. For illustration, we present examples from two projects where text and speech data are annotated.

Keywords

schema development, reliability, kappa, semantic annotation, speech data

INTRODUCTION

Despite efforts to automatize annotations of linguistic data [33, 3] manual annotations still play an important role in the compilation of corpora and linguistic research material. The quality of such manual annotations depends on the use of adequate and reliable coding schemas, which define the categories underlying annotations of linguistics data. Their development and evaluation must therefore be seen as one of the major tasks in annotation projects. Schema development is crucial because unreliable schemas may lead to inconsistencies in the labeling of objects not only by a single coder over time, but also to inconsistencies in the labeling among different coders. Both types of inconsistencies indicate a reduced usability of annotated data.

K-CAP'03, October 23–26, 2003, Sanibel Island, Florida, USA.

As many annotation projects have shown, especially highly complex coding schemas are difficult to use and will thus often lead to an unacceptable low quality in manual annotations [29, 4]. Most of the time, researchers will choose to reduce the complexity of schemas to make them more manageable for human annotators. [29], for instance, reduced her schema from originally 31 to seven basic categories for these reasons. But even when the evaluation of an annotation schema leads to good results in terms of reliability, the number of categories might still be reduced to yield even better results [20]. This procedure, however, has the severe drawback of also reducing the amount of information which can be represented in linguistic data and which may be valuable for more complex research questions or applications such as information extraction, word-sense disambiguation, document layout, or in the context of the semantic web [24, 32, 27] Hence, it seems vital to develop a systematic approach to schema development with the aim to create highly complex reliable coding schemas which are nonetheless manageable for human annotators. The approach presented here consists of measuring the reliability of the newly developed schema, systematically considering and identifying sources of unreliability by statistical means, and thus iteratively evaluating and improving the schema. The resulting methodological framework is believed to be fruitful, especially as most previous approaches seem to lack a systematic methodology of schema development and evaluation, which often makes complex schema usage so dissatisfying. Approaches like the one presented by [8] for a dialogue coding scheme still seem to be an exception.

In the remainder of the article, we first want to describe the basic principles of the methodological framework and thereafter demonstrate its applicability by presenting data from two separate annotation projects.

Critical Issues in Schema Development

The task of annotating linguistic data can be seen as a categorization task in which objects (e.g. morphemes, words, phrases, sentences) have to be assigned to a single category. The assignment of one object (usually) has to be exclusive and independent from the categorization of other objects. To be valuable, the prerequisite of such an assignment is that identical objects will be assembled in the same category, which leads to the following conclusions:

1. The categories of the coding schema must be defined in a way that enables humans to adequately differentiate among them.
2. The schema must be usable in a consistent way by several persons as well as by one person over time.

Although the first point seems rather trivial, it constitutes exactly the way in which most complex coding schemas fail. Especially in the case of semantically close concepts the interpretations of single coders often vary considerably [2, 35]. This leads to the second point, which refers to the question of consistency in manual annotations and is thus directly related to the issue of reliability. Reliability can be defined as "the complex property of a series of observations or of the measuring process that makes it possible to obtain similar results if the measurement is repeated" [15, p. 51].

In contrast to the use of existing schemas where inconsistency in annotations are usually attributed to differences in the application of the schema by annotators or even to characteristics of the human annotators, the sources of inconsistency in schema development must be attributed to a lack of reliability of the schema itself. Accordingly, the steps to be taken when reliability is not high enough are supposed to be different. Whereas in the case of text annotation intensive training and the application of supporting tools are appropriate measures to secure annotation quality [21], in schema development improvement of the coding schema must be the primary goal.

When asking why schemas might be used inconsistently by different coders or even one coder over time, several reasons may play a role. On a theoretical basis, two major problematic aspects in schema development can be differentiated that lead to systematic variance in annotation behavior [16, 31]. As stated above the interpretation of categories might be ambiguous. Secondly, the probability of assigning an object to a category may differ among coders. In addition, annotators might develop new or slightly aberrant coding habits over time. The latter point refers mainly to the reliability of the annotation process, but it can be hypothesized that such aberrations occur mostly when category definitions and boundaries are diffuse.

As a consequence, the actions taken to improve the quality of a schema must be based on the specific reasons responsible for systematic variance in manual annotations. Problematic features in the schema or its application must thus be analyzed thoroughly in order to ensure or improve the reliability of the coding instrument.

OUTLINE OF METHODOLOGY

The above considerations led to the design of a methodological framework for systematic schema development and evaluation. It comprises five successive steps, which are repeated as long as the results are not considered to be sufficient. These steps are

Step 1: let two or more coders annotate a sufficient amount of data with a preliminary coding schema
Step 2: repeat the annotation with the same coders and material after a certain period of time
Step 3: check both annotations for inter- and intra-coder agreement as a measure of reliability
Step 4: identify sources of inconsistency
Step 5: take appropriate steps to improve the coding schema

Methodological considerations concerning single steps will be considered further in the following sections.

Preliminary Considerations (Step 1)

Before starting the evaluation process some basic facts have to be considered such as the number of coders and the amount of material to be annotated in order to give a sufficient data basis for statistical analyses. Concerning the amount of data needed, to our knowledge, there is no evidence of how many observations dependent on the overall number of categories should be obtained to get reliable statistics.[1] Literature on kappa (see below) usually mentions at least 100 observations to get sufficient data for calculating significance [13, 12]. [12] propose a much larger amount of data if considerable differences in the number of cases per categories are expected.

Second Annotation (Step 2)

The second annotation is done a certain time after the first using the same material. The amount of time elapsing between the two annotations is not easy to chose, however.

As known from social sciences the time between the first and the second testing, i.e. annotation may influence the result or the degree of agreement between the two measurements [9, 14]. Since linguistic annotations do not deal with personal traits or attitudes that may change over time, the impact might not be as severe as in the context of the social sciences. Still, certain effects of time should be taken into account. For instance, if the two annotations lie too close together, the coders may remember large parts of their former annotation which leads to overestimations of retest-reliability. Too long

[1] Some considerations about suitable sample size could be found in case of equal marginal frequencies or with restriction to only two categories [6]. Both cases are not applicable to our problem, however.

a period, on the other hand, might not only cause undesirable delays in annotations. If the annotation process is stopped throughout this period, unrealistically high negative effects could appear as definitions or whole categories of a coding schema tend to be forgotten (especially in the case of highly complex schemas). Obviously, the calculation of the appropriate period of time that should elapse before starting the second annotation is not as trivial as it may first seem. We, unfortunately, do not have a definite answer to what the 'best' span of time is. This part is still open to research.

Measuring Inter- and Intra-Coder Reliability (Step 3)

Types of Reliability Two ways to measure the reliability of a coding schema seem feasible, known also from social sciences for the development of rating instruments. The first possibility is the application of a schema by different coders annotating the same material which leads to the measurement of inter-coder agreement (ICA). The second is its application to two or more times by one coder for the same material which indicates intra-coder agreement or test-retest reliability (TRR) in terms of measurement theory [14]. These two types are comparable to what [18] named *stability* and *reproducibility*. The first approach thus measures consistency among different persons, whereas the second approach measures consistency of one person over time. These measures can be affected by variations in the coding behavior, leading to inconsistencies, which may be both observable in manual annotations, without being necessarily attributable to the same sources. Since features of highly complex schemas may induce both kinds of inconsistencies independently, they should hence be analyzed separately.

Calculating Reliability Inter-coder agreement and intra-coder agreement (test-retest reliability) can both be calculated by the same means. As annotations of linguistic data primarily consist of annotations with mutually exclusive categories without any ordering (i.e. nominal data) calculations can be done with the κ-(kappa)-coefficient developed by [10]. κ measures the agreement between two coders while correcting for chance agreement, which is the reason why it should be preferred to the mere calculation of the percentage of agreement [7]. The value of the resulting kappa-coefficient indicates the degree of agreement. For the interpretation of the resulting kappa one may refer to rules of thumb like those given by [19], where $0 \leq \kappa < 0.2$ means light agreement, $0.2 \leq \kappa < 0.4$ fair agreement, $0.4 \leq \kappa < 0.6$ moderate agreement, $0.6 \leq \kappa < 0.8$ substantial agreement, and $0.8 \leq \kappa < 1.0$ almost perfect agreement. In spite of several problems with this measure of agreement [1], kappa has the advantage of being widely accepted and easy to calculate.

Identifying Sources of Unreliability (Step 4)

To detect possible reasons for lack of agreement, we decided to test the homogeneity of marginal distributions, which can be seen as an indicator of whether coders have different interpretations of the meaning of categories or whether coders just use single categories with different frequencies [31]. The comparison is based on differences in coders' assignments to single categories, e.g. the number of times coder 1 assigned an object to category A (0 times) with the number of times coder 2 used category A (1 times) (cp. table 1). Homogeneity is assumed when the distributions, i.e. marginals do not differ significantly.

		Categories – Coder 2							
		A	B	C	D	E	F	G	Σ
Categories Coder 1	A	0	0	0	0	0	0	0	0
	B	0	1	0	0	0	1	0	2
	C	0	4	26	0	1	0	3	34
	D	0	0	0	0	0	0	0	0
	E	1	1	1	1	46	0	3	53
	F	0	0	0	1	0	2	1	4
	G	0	5	2	0	3	2	13	25
	Σ	1	11	29	2	50	5	20	118

Table 1: Assignment decisions of two coders (imaginary data)

For two coders this check can be done with the non-parametric test by Stuart and Maxwell [28, 22], which calculates the overall homogeneity over all categories. Significance of the test indicates that marginal homogeneity is not given, and thus a different interpretation of categories must be assumed.

Furthermore it is important to identify the problematic categories, i.e. those which are interpreted differently by the coders. This can be done with the aid of the McNemar-test [23]. This test considers the marginal distributions of schemata with only two categories, i.e. the category under consideration and a compound in which the remaining categories are joined together. Significance of the test indicates different interpretations of the category under consideration. For the calculation of both statistics we resorted to the MH-program developed by Uebersax. The tool can be obtained as freeware from `http://ourworld.compuserve.com/homepages/jsuebersax/mh.htm`.

CASE 1: CODING SCHEMA FOR SEMANTIC TEXT ANNOTATIONS

Setting

The methodological framework for schema development was developed within an annotation project at the University of Giessen. The aim of this project is the analysis of the semantics of document structures.[2] For this purpose, English and German scientific articles are manually annotated on multiple levels, namely the structural and two semantic levels, called rhetorical and thematic. The thematic structure of the article describes the 'text world' that is referred to by the article, the article's rhetorical structure describes the rhetorical relations that hold between the discourse units of the article.

[2] Project C1/*SemDoc*, DFG-Forschergruppe 437/*Texttechnologische Informationsmodellierung*. For more detailed information about the project see `http://www.text-technology.de/`

Category	Definition
assumption	theoretical assumption or supposition by the author
theoreticalBasis	well established theoretical knowledge in the research area
hypothesis	concrete formulation of a statistically testable assumption, which is to be either corroborated or refuted by the results of the study

Table 3: Examples of category definitions

While we could resort to existing coding schemas for the structural and rhetorical level, which only needed to be adjusted to our purposes, the thematic schema had to be developed nearly from scratch. Using existing schemas [17, 30] as well as analyses of sample scientific articles as a starting point we compiled a coding schema of originally 71 topics such as `method`, `history`, and `inducements`. By applying the schema to a wide range of documents, it was extended to presently 121 different topics. Some of these categories represent very subtle semantic differences (see table 3), which made it necessary to accomplish the annotation task manually. The annotation itself is done by hand in an XML-format in the style of [25]. A small part of an annotated document is shown in figure 2.

Guidelines defining the topics and clarifying problematic cases were written. At the beginning of the annotation process the quality, measured in terms of inter-coder agreement, was very low. We obtained agreement rates between $\kappa = .09$ and $\kappa = .50$ ($m = .22$) for two coders each annotating the same six documents. Since the annotation quality did not improve much during the following annotation sessions we attributed the problem to the annotation schema itself. Since we did not want to reduce our schema in order to retain as much information as possible we decided to develop a methodology to improve the usability of the coding schema instead.

First Evaluation Cycle

Steps 1 and 2: Annotations As in the project three separate annotation levels are used, the number of annotators for each level was kept to the minimum of two annotators each. In order to meet the requirements for kappa (see above) and to ensure a more or less even distribution in the probability of occurrences of topics we decided to annotate two complete scientific articles for each evaluation cycle. The chosen articles for the first annotation cycle contained between 102 and 192 segments to be annotated leading to an average number of 293 annotated topics for each coder. The two coders annotated independently from each other. The second annotation was done approximately two weeks after the first.

Step 3: Calculating Reliability For the two documents in our first evaluation cycle we obtained kappas at the slight to moderate level of agreement (see table 4), which clearly could not be considered as satisfying. Inter-coder agreement was calculated with data from the first annotations of each coder. The rather low kappa coefficients led to the question of why such a low agreement was obtained and, in turn, where the causes for the lack in agreement could be found.

		TRR	
	ICA	**coder 1**	**coder 2**
text 1	.18	.45	.64
text 2	.26	.55	.62
mean	.22	.50	.63

ICA: inter-coder agreement; TRR: test-retest reliability

Table 4: Degrees of agreement at the first evaluation cycle [kappa-values]

	Number		Significance
Category	**Coder 1**	**Coder 2**	**Level**
2	4	10	0.031
3	9	0	0.004
5	45	79	0.000
6	19	0	0.000
7	19	0	0.000
11	9	0	0.004
12	6	18	0.011
20	9	20	0.001

Table 5: Differently interpreted categories

Step 4: Identifying Sources of Unreliability As our results from the first evaluation cycle show interpretations turned out to differentiate considerably. The Stuart-Maxwell test was highly significant ($\chi^2 = 90.42;\ p < 0.001$). The McNemar-Text for single categories showed that in the first evaluation cycle eight categories were interpreted differently by the two coders (table 5). Two differences, however, occurred because coder 1 introduced new topics which was therefore not known to coder 2 (categories 6 and 7).

By further checking the types and number of categories annotated by both coders we found the effect that the first coder annotated many more different categories than the second coder. In text 1 and text 2 the first coder annotated 71 and 51 categories, respectively, whereas coder 2 chose between 39 categories in text 1 and 34 categories in text 2. In this light the higher TRR-values of coder 2 do not seem so surprising any more.

Step 5: Adjustment of the Schema Starting from the statistical evidence, we now began to adjust our coding schema. First, we discussed the problematic categories from table 5 with the annotators to clarify their understanding. Definitions were adjusted and fixed in the annotation guidelines like in case of category 23 (table 6). The two newly invented categories 6 and 7 were dropped because discussion showed that they could be subsumed in two existing categories. The differences in annotation behavior of the two coders concerning the use of a different amount of categories were also discussed and more rigorous guidelines established.

```
<segment id="s81" parent="g15" topic="assumption"> In these situations, it does not matter
whether he or she mentions any information during discussion.</segment>
<segment id="s82" parent="g7" topic="findings_oth" litref="s529 s547"> However, Stasser(1988;
see also Stasser et al., 1989) identified a types of information distribution in which the
best decision is not apparent to members prior to discussion.</segment>
<segment id="s83" parent="g17" topic="name_cpt"> This is termed a hidden profile.</segment>
```

Table 2: Part of an annotation at the thematic level

Category	Definition
textual (old)	statements of the author's intentions or about the organization of text or text parts
textual (new)	statements of the author's intentions or about the organization of text or text parts, also information for further reading; table captions are excluded

Table 6: Adaptation of definition for category 23

		TRR	
	ICA	**coder 1**	**coder 2**
text 1	.44	.80	.55
text 2	.40	.74	.64
mean	.42	.77	.60

ICA: inter-coder agreement; TRR: test-retest reliability

Table 7: Degrees of agreement at the second evaluation cycle [kappa-values]

Second Evaluation Cycle

After the modifications of the coding schema a new evaluation cycle started, which included the same steps as described above. In the second evaluation cycle we obtained the results stated in table 7. ICA values were nearly twice as high than in cycle 1. Also TRR values for coder 1 increased considerably. (Data for the second annotation of coder 2 was not available in time, but will be ready in short.) According to [19] the test-retest reliability for coder 1 could now be considered as substantial to almost perfect, indicating that the schema may be used consistently over time by a single coder. Inter-coder agreement turned from fair to moderate.

The test for marginal homogeneity still was highly significant ($\chi^2 = 153.02;\ p < 0.001$). The comparison of single categories, however, showed that three instead of the former eight categories were not used in accordance (table 8). Hence, other evaluation cycles will follow in the near future to further improve the coding schema.

	Number		Significance
Category	**Coder 1**	**Coder 2**	**Level**
12	0	15	0.000
23	1	9	0.011
25	10	2	0.021

Table 8: Differently interpreted categories

CASE 2: EVALUATION OF ANNOTATIONS OF SPEECH DATA

We also tested our methodological approach for coding schema evaluation with data from another project. The LeaP (`http://leap.lili.uni-bielefeld.de`) project is concerned with the acquisition of prosody by foreign language learners and has set up a large corpus of annotated speech files. These were annotated by six coders using a six-tier coding schema. On the first tier, type of phrases (e.g. complete, interrupted) and intervening non-speech events such as laughter and noise are coded. The second tier consists of an orthographic annotation of words. On the third tier, syllables are annotated in SAMPA [34], and on the fourth tier vowel and consonant boundaries are annotated. On tier 5, tones are annotated using the ToBI [26] system, and on the sixth tier, initial highs, final lows and intermediate highs and lows of pitch are marked. For one speech file, an average of 1000 annotations are carried out. All annotators were trained for two months at the beginning of the project.

For the calculation of inter-coder agreement one speech file consisting of 368 words was annotated separately by three to four annotators. For a measure of overall agreement the median of all pairwise comparisons per tier (kappa-values) was calculated. Since orthographic environment, i.e. words and syllables can not be considered as categories, no agreement was calculated for the second and third tier. The results of pairwise and overall agreement for each tier are shown in table 9. Kappa-values clearly indicate that certain tiers are more difficult to annotate in agreement than others, e.g. tones and phrases. These differences seem attributable mainly to the complexity of the underlying schemas as the number of categories from tier 1 to tier 6 are seven (phrases), three (vowels), 34 (tones), four (pitch). For the calculation of retest-reliability the first file annotated was annotated again two years later by each coder. Results for tier 1 and tier 4 show that kappa-values are on a moderate level of agreement (table 10). In the light of the long period of time that elapsed between the first and the second annotation this must still be seen as a rather good result.

An evaluation of the reasons for disagreement will be presented here only for the pair coder 1 – coder 3 in the first tier

	Coder Pair						
Tier	1-2	1-3	1-4	2-3	2-4	3-4	**Median**
1- phrases	.40	.39	.43	.57	.63	.60	.50
4- vowels	.46	.46	.52	.46	.46	.49	.46
5- tones	.21	.20	.29	.30	.35	.25	.27
6- pitch	.58	.68	–	.62	–	–	.62

Table 9: Inter-coder agreement at different annotation tiers [kappa-values]

	Coder			
Tier	1	2	3	4
1- phrases	.53	.24	.51	.65
4- vowels	.58	.35	.46	.53

Table 10: Retest-reliability at different annotation tiers [kappa-values]

(inter-coder agreement), since this is the pair with the lowest agreement on this level. Procedure and interpretation are identical to those described in case 1. As the only tendentially significant Stuart-Maxwell Test ($\chi^2 = 12.404;\ p < 0.05$) proposes, the overall interpretation of categories can be considered as nearly identical. This leads to the conclusion that the differences are attributable primarily to a systematic variance in assigning objects to different categories. Additionally, however, the McNemar-Test reveals that there is one category in the schema (category 2) that is interpreted differently ($\chi^2 = 7.36;\ p < 0.05$). The implication in this case would be to first clarify the definition of the problematic category with both coders, and then to resume training with the aim of improving the differentiation between objects.

PRACTICAL PROBLEMS

In applying the methodology to the two projects described above we encountered some practical problems, which might be worth noting, since they are likely to occur in other applications as well and in quite a similar way.

Coder Characteristics

In our case studies we assumed that coder characteristics were stable or had no direct influence on annotation quality. This of course is an overly optimistic view. Individual characteristics of coders such as familiarity with the material, amount of former training, but also motivation and interest may clearly have a varying impact on their work. In both studies we tried to keep these variables as stable as possible by providing equal training for every coder, choosing annotators familiar with the subject or material and giving guidelines for the annotation process aimed at reducing effects of fatigue (e.g. restricting the annotation time to maximally three hours per session). Nonetheless, as interaction effects of coder characteristics and coding task cannot be excluded, the choice of a group of similar coders should be aspired.

	Coder Pairs						
Tier	1-2	1-3	1-4	2-3	2-4	3-4	**Median**
1- phrases	.86	.92	.88	.89	.93	.90	.90
4- vowels	.99	1.00	.99	.99	1.00	.99	.99
5- tones	.44	.44	.58	.56	.58	.51	.54
6- pitch	.96	.94	–	1.00	–	–	.96

Table 11: Inter-coder agreement in case 2 [corrected kappa-values]

Kappa as Measure of Reliability

One of the major problems when employing kappa is that the coefficient depends on the actual marginal distribution [11, 5]. In cases with heterogeneous marginal distributions kappa may not have the originally intended range of -1 to $+1$, but a more restricted one. This will not only reduce the kappa-values obtained, but also the interpretability of the coefficient, since rules of thumb for interpreting the goodness of the coefficient [19] do not apply anymore.

In this case some authors suggest the calculation of the possible maximum that kappa can reach (κ_{max}) with the given marginal distribution [10, 1]. The expression κ/κ_{max} will then lead to a corrected κ with the original range of -1 to $+1$ [10, 1]. Even though this procedure would have the big advantage of not only tremendously improving the kappa-values (see table 11 for an example), but also of restoring the original interpretation of kappa, we refrained from using it in the context of our framework. Severe aberrations from the homogeneity of marginal distributions often indicate underlying problems with the use of the categories. By correcting kappa, valuable information would be discarded.

CONCLUSIONS

The aim of the work presented here was to present hands-on experience with the development of highly complex coding schemas for manual annotations of linguistic data. The methodological framework we created in order to solve our problems with poor annotation quality because of the high complexity of the annotation task proved fruitful not only in the context of our original project aiming at the semantic annotation of text documents, but also in translating it to the annotation of speech data. We therefore feel confident that the systematic and iterative process presented here can profitably be applied in other annotation projects, where complex coding schemas have to be developed and evaluated.

REFERENCES

1. Brennan, R.L. and Prediger, Dale J. (1981). Coefficient kappa: Some uses, misuses, and Alternatives. *Educational and Psychological Measurement, 41, 687-699.*

2. Bruce, R. and Wiebe, J. (1999). Recognizing subjectivity: A case study in manual tagging. *Natural Language Engineering, 5(2), 187-205.*

3. Bulyko, I. and Ostendorf, M. (2002). A bootstrapping approach to automating prosodic annotation for limited-domain

synthesis. *Proceedings of the IEEE Workshop on Speech Synthesis, 11-13 September, Santa Monica, California USA.*

4. Butler, T., Fisher, S., Coulombe, G., Clements, P., Brown, S., Grundy, I., Carter, K., Harvey, K. and Wood, J. (2000). Can a team tag consistently? Experiences on the Orlando project. *Markup Languages, 2(2), 111-125.*
5. Byrt, T., Bishop, J. and Carlin, J.B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology, 46(5), 423-429.*
6. Cantor, A.B. (1996). Sample-size calculations for Cohen's kappa. *Psychological Methods, 1(2), 150-153.*
7. Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics, 22(2), 249-254.*
8. Carletta, J., Isard, A., Isard, S., Kwotko, J.C., Doherty-Sneddon, G. and Anderson, A.H. (1997). *The reliability of a dialogue structure coding scheme, 23(1), 13-31.*
9. Carmines, E.G. and Zeller, R.A. (1979). *Reliability and validity assessment*. Sage Publications: Beverly Hills and London. Paper series on Quantitative Applications in the Social Sciences, 07-017.
10. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20(1), 37-46.*
11. Feinstein, A.R. and Cicchetti, D.V. (1990). High agreement but low kappa: I. The problem of two paradoxes. *Journal of Clinical Epidemiology, 43(6), 543-549.*
12. Flack, V.F., Afifi, A.A., Lachenbruch, P.A. and Schouten, H.J.A. (1988). Sample size determinations for the two rater kappa statistics. *Psychometrika, 53(3), 321-325.*
13. Hanley, J.A. (1987). Standard error of the kappa statistic. *Psychological Bulletin, 102(2), 315-321.*
14. Helmstadter, G.C. (1964). *Principles of psychological measurement.* Meredith Publishing: New York.
15. Hollnagel, E. (1993). *Human Reliability Analysis Context and Control*. Academic Press: London.
16. Hoyt, W. and Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4(4), 403-424.*
17. Kando, N. (1997). Text-level structure of research papers: Implications for text-based information processing systems. *Proceedings of the British Computer Society Annual Colloquium of Information Retrieval Research, Aberdeen, Scotland, 8-9 April 1997, 68-81.*
18. Krippendorff, K. (1980). *Content analysis: An introduction.* Sage Publications: Beverly Hills and London.
19. Landis, J.R. and Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33(1), 159-174.*
20. Maier, E. (1997). *Evaluating a Scheme for Dialogue Annotation.* VERBMOBIL Report 193. DFKI GmbH, Saarbrücken.
21. Marcu, D., Romera, M. and Amorrortu, E. (1999). Experiments in Constructing a Corpus of Discourse Trees: Problems, Annotation Choices, Issues. *The Workshop on Levels of Representation in Discourse, Edinburgh, Scotland, 71-87.*
22. Maxwell, A. Comparing the classification of subjects by two independent judges. *British Journal of Psychiatry, 116, 651-655.*
23. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12, 153-157.*
24. Ng, H.T., Lim, C.Y. and Foo, S.K. (1999). A Case Study on Inter-Annotator Agreement for Word Sense Disambiguation. *Proceedings of the ACL SIGLEX Workshop: Standardizing Lexical Resources, College Park, Maryland, USA, 21-22 June 1999, 9-13.*
25. O'Donnell, M. RST-Tool 2.4 - A Markup Tool for Rhetorical Structure Theory. *Proceedings of the International Natural Language Generation Conference (INLG'2000), Mitzpe Ramon, Israel, 12-16 June 2000, 253-256.*
26. Silverman, K., Beckman, M., Pitrelli, J. Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and HirschbergJ. (1992). ToBI: A standard for labeling English prosody. *Proceedings of the 1992 International Conference on Spoken Language Processing, Denver Colorada, USA, 16-20 September 1992, 867 – 870.*
27. Staab, S., Maedche, A. and Handschuh, S. (2001). *Creating metadata for the semantic web: An annotation framework and the human factor.* Technical Report 412. Institute AIFB, University of Karlsruhe.
28. Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika, 42, 412-416.*
29. Teufel, S. (1999). *Argumentative zoning: Information extraction from scientific text.* PhDThesis, University of Edinburgh.
30. Teufel, S., Carletta, J. and Moens, M. (1999). An annotation scheme for discourse-level argumentation in research articles. *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL-99), Bergen, 8-12 June 1999.*
31. Uebersax, J. (2001). *Statistical Methods for Rater Agreement.* online available: http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm.
32. Veronis, J. (2000). Sense tagging: Don't look for the meaning but for the use. *Workshop on Computational Lexicography and Multimedia Dictionaries (COMLEX'2000), 22-23 September 2000, Patras, Greece, 1-9.*
33. Vorsterman, A., Martens, J.-P. and Coile, B. van. (1996). Automatic segmentation and labeling of multi-lingual speech data. *Computational Linguistics, 19(4), 271-293.*
34. Wells, J.C., Barry, W., Grice, M., Fourcin, A. and D. Gibbon. (1992). *Standard Computer-Compatible Transcription.* SAM Stage Report Sen.3 SAM UCL-037, University College London.
35. Wiebe, J.M., Bruce, R.F. and O'Hara, T.P. (1999). Development and use of a gold standard data set for subjectivity classifications. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99), 20-26 June 1999, University of Maryland, 246-253.*