Investigating inter-coder agreement: Log-linear models as alternatives to percentages and the kappa statistic

Petra Saskia Bayerl
Applied and Computational Linguistics
Justus-Liebig University, Giessen

15. October 2003

Abstract

In quality considerations of manual annotations, the calculation of inter-coder agreement plays an important role. Usually measures such as percentages or kappa are used for this purpose, since they provide an easy way to obtain an overview of the overall annotation quality or reliability. These well-known methods, however, don't provide insights into patterns of agreement. Yet, such information is the basis for well-founded decisions about how to improve annotation quality. Log-linear models are a way to obtain such information as they try to 'reproduce' the data patterns obtained by two or more annotators. This paper shows how log-linear models can be applied for systematic investigations of inter-coder agreement. After a brief discussion of well-known summary indices, it presents basic assumptions of log-linear models demonstrating their actual application in a short example.

1 Introduction

Quality of annotations in computational and corpus linguistics often are operationalized as the amount of agreement of two or more human annotators. This so-called *inter-annotator* or *inter-coder agreement* in most cases is measured by percentage agreement, i.e. percentage of cases annotated with the same category according to the total sum of cases, or alternatively by means of the kappa statistic (Cohen, 1960, 1968). Both methods result in a single number which indicates the degree of agreement among coders. The advantage of these indices is that they are easy to calculate and seem to be comparable across studies. Since both methods are primarily interested in indicating the fit between classifications they can be thought of as summary or overall measures (Agresti, 1992; Rapallo, 2002). Thus, these summarization approaches discard valuable information, that could be used for discovering underlying reasons of disagreement among annotators, such as biases and systematical differences in the interpretation of categories. Such information, however, is most important when deeper insights in the annotation process is needed in order to improve inter-coder agreement.

Since neither percentages nor kappa can provide such information, alternative methods must be applied. Log-linear models recommend themselves for this purpose. In order to show in what respect log-linear methods differ from summarization approaches, this paper starts with a short discussion of two common agreement measures, namely percentages and kappa. Thereafter, basic assumptions of log-linear models and their application in investigations of inter-coder agreement are presented.

2 Summary statistics as agreement measures

For the calculation of inter-coder agreement in case of categorical data a variety of measures have been proposed such as χ^2, percentages, kappa (Cohen, 1960), the G-index (Holley & Guilford, 1964), or I_r as 'index of reliability' (Perreault & Leigh, 1989; Rust & Cooil, 1994). All of these indices are so-called *summary statistics* (Agresti, 1992; Schuster, 2002), which provide one single value as indicator of agreement, i.e. they aggregate information about agreement and disagreement among coders in one coefficient.

Of the summary statistics cited above only two, namely percentages and kappa, can be said to be widely used in the area of linguistics. Their advantage lays in their easy calculation and interpretability. But, despite their popularity, they are not without problems when using them as agreement measures.

2.1 Percentage agreement

Percentage agreement (PA) refers to the proportion of judgments that two coders make in agreement in relation to the total number of judgments:

$$PA = 100 \times \frac{\text{number of agreements}}{\text{number of agreements} + \text{number of disagreements}} \tag{1}$$

As can be seen in a wide range of studies, percentages are still a frequently used agreement measure (Melamed, 1998; Halteren, 1998). They are easy to calculate, intelligible, and seemingly comparable with indices from other studies. But, as a number of authors have pointed out so far, their value as an agreement index is nonetheless questionable. One of the most argued points is that these measures are influenced by chance agreement or guessing (Cohen, 1960, 1968; Green, 1981). Especially in cases with very high or very low frequencies of cases per category one can obtain high levels of chance agreement, because judges may reach agreement even if their coding decisions are merely based on the assignment of the most likely category. This in turn will lead to an overestimation of the reliability of judgments (Mitchell, 1979; Wakefield, Jr., 1980; Hayes & Hatch, 1999). Since this chance factor is not controlled for and may thus account for a varying part of the overall agreement, percentages are not truly comparable among different studies or applications (Yelton, Wildman, & Erickson, 1977). In accordance with Perreault and Leigh (1989, p. 137) it is therefore "difficult to set quality standards based on percentage agreement statistics".

2.2 Kappa statistics

The original kappa statistics was developed by Cohen (1960) as an answer to the rigorous criticisms of percentage agreement indices. It is conceptionalized as a fully chance corrected measure of agreement which is defined as follows:

$$\kappa = \frac{p_o - p_c}{1 - p_c} \tag{2}$$

with p_o proportion of observed agreements and p_c proportion of agreements expected by chance.

The kappa statistic originally was thought as an agreement index for two coders and categorical data, but a series of extensions and variations have been proposed to adjust kappa to cases with more than two coders (Conger, 1980; Uebersax, 1982), or ordered categories (Cohen, 1968; Kválseth, 1985; Janson & Olsson, 2001).

Kappa can be considered to be one of the most widely used indices for inter-coder agreement. Carletta (1996) explicitly recommended the use of kappa for linguistic studies as an valuable alternative to percentages. Reasons for its popularity lay not only in the ease of calculation, but also in its simple interpretability. Kappa as a summarization statistic leads to one index κ, the value of which can be interpretated as an indicator of the 'goodness of agreement' between coders. Rules of thumb introduced, for instance, by Landis and Koch (1977) give a framework for comparing indices among studies.

The widespread use of kappa has caused a fierce discussion about its actual value as agreement measure. In most critics the dependency of kappa on marginal distribution is addressed. Cohen conceptualized kappa in such a way that it reaches perfect agreement, i.e. a value of $+1$, only in case of marginal homogeneity. As Feinstein and Cicchetti (1990) observe this may lead to the absurd case, where high agreement goes along with low kappa values. An other critical point is that kappa values increase as the number of categories in the contingency table increases (Munoz & Bangdiwala, 1997). This leads to concerns in interpretability and comparability. Other critics concerning comparability are the base rate problem discussed by Spitznagel and Helzer (1985), where kappa values depend in the proportions of positive and negative cases in the sample. Consequently Uebersax (2001) assumes that general interpretation guidelines are not appropriate. Basing on the same assumptions Thompson and Walter (1988) and Feinstein and Cicchetti (1990) presume that kappa is not comparable among studies. Furthermore, concerns about the kind of chance calculation or chance correction are formulated (Brennan & Perdiger, 1981; Uebersax, 2001). Problematic is also that kappa is insensitive to patterns of agreement (Light, 1971).

2.3 Conclusions

While summary statistics may be valuable for providing an overall index of agreement, neither percentages nor kappa seem to be as reliable as would be desirable for an agreement measure. Moreover, none provide information about the reasons or patterns of disagreement. The latter, however, would be valuable in order to detect starting points for the improvement of inter-coder agreement. This leads to the conclusion that alternatives to summary statistics might be worth considering. Such an alternative approach will be introduced in the following section.

3 Log-linear models for inter-coder agreement

3.1 Basic assumptions

In contrast to summary statistics log-linear models try to explain the way in which the specific pattern of a data set evolves. The general approach is to find a model with which the pattern of observed frequencies n_{ij} can best be explained. For this, frequencies for each category are calculated according to the chosen model (expected frequencies, μ_{ij}) and compared to the observed frequencies. If the fit is not good enough, the model is rejected.

The effects of rows, columns, and their interaction on n_{ij} are additive when considered on a logarithmic scale. Thus, the definition of a model for two coders is achieved by specifying single effects:[1]

$$log\ \mu_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \tag{3}$$

where log denotes the natural logarithm, λ is a constant, λ_i^A the row effect, λ_j^B the column effect, and λ_{ij}^{AB} the interaction of row and column effects. The fit of a model is determined with means of Pearson chi-square χ^2 or by the likelihood-ratio statistic L^2 defined as follows

$$\chi^2 = \Sigma \frac{(O_i - E_i)^2}{E_i} \tag{4}$$

$$L^2 = 2\,\Sigma\, O_i ln\left(\frac{O_i}{E_i}\right) \tag{5}$$

with O_i observed cell frequencies and E_i expected cell frequencies. If the statistic is significant, the model fit is poor and thus the model has to be rejected. The question of which coefficient to choose for determining model fit, i.e. χ^2 or L^2, has been discussed in a variety of articles. Both test statistics approximate to the chi-square distribution, with number of degrees (df) as total number of cells minus the number of estimated free parameters. Given that the total number of annotated objects is big enough, both values are nearly identical. Yet, when data is sparse the approximation might be poor. This is true especially for L^2 when $n/IJ < 5$. Several studies found χ^2 to be preferable to L^2 when N is small (Agresti & Yang, 1987; Hosmane, 1986; Upton, 1982). However, a wide range of studies use L^2 since it is exactly partitionable, i.e. models can easily be compared. As the thorough discussion by Cressie and Read (1989) shows, a definite answer to this issue is difficult to give. In the present paper L^2 will be used.

3.2 Modelling agreement

As Bruce and Wiebe (1999) have shown in their case study of manual tagging, log-linear models can easily be applied for investigations of inter-coder agreement in linguistic tasks. In contingency tables like table 1 agreement between

[1] For ease of presentation, models will be restricted to two coders, but they can easily be adjusted to cases with three of more coders. See e.g. Tanner and Young (1985) or Schuster (2002).

two coders is determined by cells in the main diagonal, whereas disagreement is placed in cells off the main diagonal.

Table 1: General contingency table for two coders

	Coder B				
Coder A	1	2	...	J	**Totals**
1	n_{11}	n_{12}	...	n_{1j}	n_{1+}
2	n_{21}	n_{22}	...	n_{2j}	n_{2+}
⋮	⋮	⋮	⋮	⋮	⋮
I	n_{i1}	n_{i2}	...	n_{ij}	n_{i+}
Totals	n_{+1}	n_{+2}	...	n_{+j}	n_{++}

A baseline model for agreement is the assumption of independence, i.e. no association between two ratings:

$$log\ \mu_{ij} = \lambda + \lambda_i^A + \lambda_j^B \qquad (6)$$

Usually this model will show a poor fit for the data, as it can be assumed that coders will show some degree of agreement. Independence can be tested against the quasi-independent model, which adds parameters for the main diagonal. Here independence is assumed only for cells off the main diagonals which should lead to a much better fit of the model.

$$log\ \mu_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \delta_I(i = j) \qquad (7)$$

with $I(i = j) = 1$ for $i = j$, and $I(i = j) = 0$ for $i \neq j$.

Of course quasi-independence models will not fit perfectly, if bias is present in the data. In order to test for biases marginal homogeneity is considered. If marginal homogeneity must be rejected, the presence of biases is assumed. As their is no log-linear model to directly test marginal homogeneity, the comparison of symmetry and quasi-symmetry models is used as an alternative way. This bases on the fact that

$$\text{symmetry} = \text{quasi-symmetry} + \text{marginal homogeneity}$$

Hence, if the fit of a symmetry model minus the fit of a quasi-symmetry model shows a significant improvement of fit for the data, marginal heterogeneity must be assumed. The symmetry model is specified in (8), quasi-symmetry in (9).

$$log\ \mu_{ij} \quad = \quad \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}, \quad \lambda_{ij} = \lambda_{ji} \qquad (8)$$

$$log\ \mu_{ij} \quad = \quad \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}, \quad \lambda_{ij} = \lambda_{ji} \text{ for all } i < j \qquad (9)$$

Further valuable information can be obtained by investigating the impact of single categories for agreement. There are several different approaches to achieve this. Haberman (1973) proposes the analyses of standardized residuals e_{ij} of categories under independence assumption. Cells with residuals greater than 2 must be considered to deviate from independence. An other approach bases on the quasi-independence model, in which the main-diagonal parameters δ_i represent the strength of agreement for category i (Agresti, 2002). The higher δ_i the higher is the agreement. The odds that coders agree rather than disagree on whether an object is to be assigned in category i rather than category j and vice versa is

$$\tau_{ij} = \frac{\mu_{ii}\,\mu_{jj}}{\mu_{ij}\,\mu_{ji}} \tag{10}$$

Under assumption of quasi-independence τ_{ij} can be calculated as

$$\tau_{ij} = exp(\delta_i + \delta_j) \tag{11}$$

Under assumption of quasi-symmetry τ_{ij} is

$$\tau_{ij} = exp(\hat{\lambda}_{ii} + \hat{\lambda}_{jj} - \hat{\lambda}_{ij} - \hat{\lambda}_{ji}) \tag{12}$$

where $\hat{\lambda}_{ij} = \hat{\lambda}_{ji}$.

In contrast, Darroch and McCloud (1986) defined *indistinguishability* of categories. Under assumption of quasi-symmetry diagonal cross-product ratios τ_{ij} here is given as

$$\tau_{ij} = \frac{\mu_{ij}\,\mu_{ji}}{\mu_{ii}\,\mu_{jj}} \tag{13}$$

The *degree of distinguishability* for categories i and j is $\delta_{ij} = 1 - (\tau_{ij})^{-1}$ with $0 \leq \delta_{ij} \leq 1$. Indistinguishability exists, when $\delta_{ij} = 0$.

3.3 An example

Table 2 presents (fictitious) data of two coders assigning 222 objects to four mutually exclusive categories. Overall agreement in this case is $PA = 74.8\,\%$ and $\kappa = .67$.

Table 2: Contingency table for two coders (fictitious data)

	Coder B				
Coder A	**1**	**2**	**3**	**4**	n_{i+}
1	51	1	2	1	55
	(10.29)^a	(-2.74)	(-3.97)	(-3.27)	
2	2	34	0	10	46
	(-2.75)	(9.31)	(-4.05)	(-0.18)	
3	0	3	42	1	46
	(-3.35)	(-1.74)	(6.34)	(-2.94)	
4	1	0	35	39	75
	(-4.04)	(-3.58)	(1.61)	(5.25)	
n_{+i}	54	38	79	51	222

^{a} standard residuals under independence

As expected the independence model (M_0) fits poorly ($L^2 = 354.83$; $df = 9$)[2]. Examining quasi-independence (M_1) this model fits better ($L^2 = 37.29$; $df = 5$), but still not perfectly. This is mainly due to bias, since the assumption of marginal homogeneity must be rejected: $L^2(S|QS) = 61.90 - 28.87 = 33.03$ with $df = 5 - 3 = 2$. The analysis of residuals (parenthesis in table 2) gives first suggestions for agreement and disagreement patterns. As expected main diagonal cells show high positive residuals, indicating higher agreement than expected under independence. Highly negative values off the main diagonals indicate lower agreement as expected under independence. Cells *24* and *43*, however, show values near zero or even above zero indicating two instances of disagreement. Eliminating these two cells from the independence model (M_4 for cell *24*, M_5 for cell *43*) leads to the conclusion, that only cell *43* represents an instance of relevant disagreement since only $L^2(M_0|M_6) = 354.83 - 348.85 = 5.98$; $df = 1$ proofs to be significant in contrast to $L^2(M_0|M_5) = 354.83 - 354.78 = 0.05$; $df = 1$. Thus, M_6 (quasi-independence with elimination of cell *43*) can be considered as an appropriate representation for the data in table 2 ($L^2 = 13.19$; $df = 4$). A practical implication would be the further clarification of category definitions *3* and *4* and additional training of coders with respect to these categories.

Table 3: Model fits

Model		L^2	**df**	**p**
M_0	independence	354.83	9	.001
M_1	quasi-independence	37.29	5	.001
M_2	symmetry	61.90	5	.001
M_3	quasi-symmetry	28.87	3	.001
M_4	independence without cell 24	354.78	8	.001
M_5	independence without cell 43	348.85	8	.001
M_6	quasi-independence without cell 43	13.19	4	.15

4 Conclusions

Unlike summary statistics log-linear models provide a systematic way to investigate patterns of inter-coder agreement. Where percentages and kappa only present an overall picture, log-linear models give deeper insights into the process and problems of manual annotations. Since their range of application is not limited to inter-coder agreement of two coders, but can also be extended to intra-coder agreement or cases of three and more coders, they represent a valuable addition or even an alternative to summary statistics.

[2] All analyses were done by means of Vermunt's computer program ℓem (Vermunt, 1997). It can be obtained as freeware from `http://www.kub.nl/faculteiten/fsw/organisatie/departementen/mto/software2.html`.

References

Agresti, A. (1992). Modelling patterns of agreement and disagreement. *Statistical Methods in Medical Research*, *1*(2), 201 – 218.

Agresti, A. (2002). *Categorical data analysis* (2 ed.). New York: Wiley.

Agresti, A., & Yang, M.-C. (1987). An empirical investigation of some effects of sparseness in contingency tables. *Computational Statistics and Data Analysis*, *5*(1), 9 – 21.

Brennan, R., & Perdiger, D. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurements*, *41*, 687 – 699.

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

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, *22*(2), 249 – 254.

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

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial agreement. *Psychological Bulletin*, *70*(4), 213 – 220.

Conger, A. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, *88*(2), 322 – 328.

Cressie, N., & Read, T. (1989). Pearson's x^2 and the loglikelihood ratio statistic g^2: A comparative review. *International Statistical Review*, *57*(1), 19 – 43.

Darroch, J., & McCloud, P. (1986). Category distinguishability and observer agreement. *Australian Journal of Statistics*, *28*(3), 371 – 388.

Feinstein, A., & Cicchetti, D. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543 – 549.

Green, S. (1981). A comparison of three indexes of agreement between observers: Proportion of agreement, G-index, and kappa. *Educational and Psychological Measurement*, *41*, 1069 – 1072.

Haberman, S. (1973). The analysis of residuals in cross-classified tables. *Biometrics*, *29*, 205 – 220.

Halteren, H. van. (1998). The feasibility of incremental linguistic annotation. *Computers and the Humanities*, *32*, 389 – 409.

Hayes, J., & Hatch, J. (1999). Issues in measuring reliability: Correlation versus percentage of agreement. *Written Communication*, *16*(3), 354 – 367.

Holley, J. W., & Guilford, J. (1964). A note on the G-index of agreement. *Educational and Psychological Measurement*, *24*, 749 – 753.

Hosmane, B. (1986). Improved likelihood ratio tests and Pearson chi-square tests for independence in two dimensional contingency tables. *Communications in Statistics - Theory and Methods*, *15*, 1875 – 1888.

Janson, H., & Olsson, U. (2001). A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological Measurement*, *61*(2), 277 – 289.

Kválseth, T. (1985). Weighted conditional kappa. *Bulletin of the Psychonomic Society*, *23*(6), 503 – 505.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159 – 174.

Light, R. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, *76*, 365 – 377.

Melamed, I. (1998). *Manual annotation of translational equivalence: The Blinker project* (Tech. Rep. No. 98-07). Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, Pennsylvania.

Mitchell, S. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin*, *86*(2), 376 – 390.

Munoz, S., & Bangdiwala, S. (1997). Interpretation of kappa and B statistics measures of agreement. *Applied Statistics*, *24*(1), 105 – 112.

Perreault, W., & Leigh, L. (1989). Reliability of nominal date based on qualitative judgements. *Journal of Marketing Research*, *26*, 135 – 148.

Rapallo, F. (2002). *Algebraic exact inference for rater agreement models* (Tech. Rep. No. 467). Dipartimento di Matematica, Genova, Italy.

Rust, R., & Cooil, B. (1994). Reliability measures for qualitative data: Theory and implications. *Journal of Marketing Research*, *31*, 1 – 14.

Schuster, C. (2002). A mixture model approach to indexing rater agreement. *British Journal of Mathematical and Statistical Psychology*, *55*, 289 – 303.

Spitznagel, E., & Helzer, J. (1985). A proposed solution to the base rate problem in the kappa statistics. *Archives of General Psychiatry*, *42*, 725 – 728.

Tanner, M., & Young, M. (1985). Modeling agreement among raters. *Journal of the American Statistical Association*, *80*(389), 175 – 180.

Thompson, W., & Walter, S. (1988). A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology*, *41*(10), 949 – 958.

Uebersax, J. (1982). A generalized kappa coefficient. *Educational and Psychological Measurement*, *42*, 181 – 183.

Uebersax, J. (2001). *Statistical methods for rater agreement.* (online available: http://ourworld.compuserve.com/homepages/ jsuebersax/agree.htm)

Upton, G. (1982). A comparison of alternative tests for the 2 x 2 comparative trial. *Journal of the Royal Statistical Society Series A*, *145*, 86 – 105.

Vermont, J. (1997). *ℓem: A general program for the analysis of categorical data* (Tech. Rep.). Tilburg University.

Wakefield, Jr., J. (1980). Relationship between two expressions of reliability: Percentage agreement and phi. *Educational and Psychological Measurement*, *40*, 593 – 597.

Yelton, A., Wildman, B., & Erickson, M. (1977). A probability-based formula for calculating interobserver agreement. *Journal of Applied Behavior Analysis*, *10*(1), 127 – 131.