Determining Intercoder Agreement for a Collocation Identification Task

Brigitte Krenn
Austrian Research Institute for
Artificial Intelligence (ÖFAI)*
Vienna, Austria
brigitte@oefai.at

Stefan Evert, Heike Zinsmeister
IMS, University of Stuttgart
Azenbergstr. 12, 70174 Stuttgart, Germany
evert@ims.uni-stuttgart.de
zinsmeis@ims.uni-stuttgart.de

Abstract

In this paper, we describe an alternative to the kappa statistic for measuring intercoder agreement. We present a model based on the assumption that the observed surface agreement can be divided into (unknown amounts of) true agreement and chance agreement. This model leads to confidence interval estimates for the proportion of true agreement, which turn out to be comparable to confidence intervals for the kappa value. Thus we arrive at a meaningful alternative to the kappa statistic. We apply our approach to measuring intercoder agreement in a collocation annotation task, where human annotators were asked to classify PP-verb combinations extracted from a German text corpus as collocational versus non-collocational. Such a manual classification is essential for the evaluation of computational collocation extraction tools.

1 Introduction

For the extraction of lexical collocations and technical terms from text corpora, a large number of association measures (AMs) have been suggested (see Evert (2004) for an overview). To assess and compare the practical usefulness of these measures, issues such as the following need to be addressed: the types of collocation to be automatically extracted, domain and size of the extraction corpora, the treatment of high frequency versus low frequency data, as well as the comparison of the outcomes of different association measures.

In practice, evaluation results are valid only for data from specific corpora, extracted with specific methods, and for a particular type of collocations. A list of true positives (TPs), against which the extraction results for different AMs are evaluated, plays a key role in an investigation of this kind. Basically there are two approaches to compiling a list of true positives:

1. extract the TPs from collocation lexica,
2. compile a list of true collocations occurring in the extraction corpus.

An essential drawback of the use of collocation lexica for evaluation is that collocation lexica do not tell us how well an AM worked on a particular corpus. They only tell us that some of the collocations listed in the lexicon also occur in the extraction corpus and that a particular AM has found them.

Using a list of true collocations occurring in the extraction corpus, however, requires a good deal of hand-annotation. Moreover the resulting list of TPs is strongly influenced by the intuitions of the particular annotators. In order to minimise the risk of subjectivity,

a. objective criteria for the distinction of collocational and non-collocational word combinations in the list of candidate data are required;

b. a certain degree of intercoder agreement on the reference data is important.

The phenomena subsumed by lexical collocations are manifold, ranging from lexical proximities in texts to syntactic and semantic units showing semantic opacity, and syntactic irregularity and rigidity. Accordingly there is a variety of definitions and terminology. Both are influenced by different linguistic traditions and by the particular computational linguistics applications for which collocations are considered to be relevant. We typically find an opportunistic approach to collocativity, i.e., the definition of TPs depends on the intended application rather than being motivated by (linguistic) theory, and it covers a mixture of different phenomena and

* The Austrian Research Institute for Artifi cial Intelligence (ÖFAI) is supported by the Austrian Federal Ministry for Education, Science and Culture, and by the Austrian Federal Ministry for Transport, Innovation and Technology.

classes of collocations. Moreover, even when well-defined criteria and explicit annotation guidelines are available, annotators may make different decisions for some of the collocation candidates, because of mistakes, differences in their intuition, different interpretation of the guidelines, etc. All this makes it hard to give a systematic experimental account of the true usefulness of a certain AM for collocation extraction.

A widely used means for measuring intercoder agreement is the kappa statistic (Cohen, 1960), where the observed agreement between coders is compared with chance agreement. Kappa values usually lie in the interval $[0, 1]$, with zero kappa indicating chance agreement and positive kappa indicating an increasing level of intercoder agreement beyond chance up to a value of one for perfect agreement. (Negative kappa indicates genuine disagreement between the annotators.) The precise interpretation of positive kappa values between 0 and 1 is still open for discussion, though. One of the widespread interpretations used for natural language coding tasks, mainly dialogue annotation, was suggested by Krippendorff (1980):

$\kappa \leq .67$	to be discarded
$.67 \leq \kappa \leq .8$	shows tentative agreement
$\kappa \geq .8$	definite agreement

(Rietveld and van Hout, 1993) quoted after (Di Eugenio and Glass, 2004) give the following interpretation:

$.20 \leq \kappa \leq .45$	fair level of agreement beyond chance
$.40 \leq \kappa \leq .60$	moderate level of agreement beyond chance

In (Green, 1997) another interpretation is given:

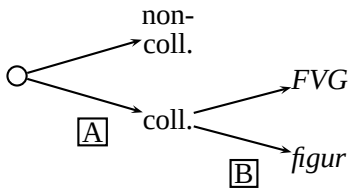
$\kappa \geq .75$	high degree of agreement
$\kappa \leq .40$	low degree of agreement
$.40 \leq \kappa \leq .75$	fair to good level of agreement beyond chance

To reduce inconsistency in the definition of the type of collocation under investigation, we concentrate on two types of PP-verb collocations that are well defined in the literature, namely German support verb constructions (Ge.: Funktionsverbgefüge) and figurative expressions.[1] These two phenomena are the main collocation types occurring in PP-verb cooccurrence data extracted from text corpora.

The paper is organized as follows: in Sections 2 and 6, respectively, we present a collocation annotation experiment and discuss its results based on the theoretical assumptions made in Sections 3 to 5. In particular, Section 3 gives a general account on measuring intercoder agreement, in Section 4 we present the kappa statistic and in Section 5 we describe an alternative approach for estimating intercoder agreement.

2 Experimental setup

In the annotation database of (Krenn, 2000) (which has been further extended since and used in several other publications, e.g. (Evert and Krenn, 2001) and (Krenn and Evert, 2001)), German PP+verb combinations are annotated as *Funktionsverbgefüge* (*FVG*) and *figurative expressions* (*figur*), according to criteria described in (Krenn, 2000). These guidelines suggest that we deal with two (hierarchical) binary classifications:



In Section 6 of this paper, we concentrate on the collocational / non-collocational distinction (step A). However, the measures of intercoder agreement discussed in Sections 3 to 5 can also be applied to step B, or more generally for assessing the decisions of two coders on any binary variable.

The experiment on intercoder agreement is as follows: We test the decisions of Krenn (with respect to step A) against several other annotators, all native speakers of German, students and researchers at the IMS. In particular, Krenn (henceforth BK) is considered the "expert", and we want to test whether the other annotators (henceforth NN) agree with her decisions, employing pair-wise comparisons: BK vs. NN. When looking at the annotated data it may be necessary to exclude some annotators who have clearly not understood the instructions or make obvious mistakes.

BK's annotation database combines candidates from various sources and from different experiments. Rather than using this haphazard collection directly, we want to evaluate agreement on

[1] An overview on the literature can be found in (Krenn, 2000).

a well-defined, reproducible[2], and practically relevant subset of the database. The subset we chose contains high-frequency PP+verb pairs from the *Frankfurter Rundschau* (FR) corpus.[3] The full FR corpus was part-of-speech tagged with TreeTagger (Schmid, 1994), enriched with lemmata and morpho-syntactic information from the IMSLex morphology (Lezius et al., 2000), and chunk-parsed with the YAC parser (Kermes, 2003). In addition to noun phrases (NP) and prepositional phrases (PP), YAC identifies verbal complexes (VC) and subordinate clauses in the text. All chunks are annotated with the corresponding head lemma. PPs are annotated both with the preposition and the nominal head. The head lemma annotations of VCs are particularly useful because they recombine separated particle verbs. Finally, all possible combinations of a VC and a PP (represented by their respective head lemma annotations) within the same main or subordinate clause were extracted as cooccurrences.

A frequency threshold was applied to the resulting data set, retaining only candidates with cooccurrence frequency $f \geq 30$. In a second step, certain "trivial" combinations were filtered out with handcrafted patterns in order to reduced the amount of manual work. Examples are combinations of any verb with a PP whose nominal head is a day of the week (*Montag*, ..., *Sonntag*), a month name (*Januar*, ..., *Dezember*) or another PP that is clearly used as a time adverbial. None of the excluded candidates was marked as a true positive by BK.[4]

The resulting subset contains $3,418$ PP+verb candidates and is called the **test set**. We split the test set randomly into 4 equally-sized parts (3×855 candidates, 1×853 candidates), which were given to different annotators along with a coding manual written by BK (in German, cf. Krenn (2004)). The annotators were told to mark collocations as either *FVG* or *figur*. Unmarked candidates were interpreted as non-collocational. Annotators were forced to make a decision for every candidate.

3 Measuring intercoder agreement

Note that this discussion is restricted to agreement between two coders (A and B) on binary variables (i.e. annotations with two categories).[5]

The agreement data obtained from the evaluation experiment for annotators A and B can be summarised in the form of a contingency table as shown in Figure 1: n_{11} is the number of candidates that were accepted as a TP by both annotators, n_{22} is the number of candidates that were considered a FP by both annotators, etc.; $n_{i\cdot}$ are the row sums (with $n_{1\cdot}$ the total number of TPs marked by A) and $n_{\cdot j}$ are the column sums (with $n_{\cdot 1}$ the total number of TPs marked by B). The sum $n_{11}+n_{12}+n_{21}+n_{22} = n$ is the total number of candidates in the test set (or the evaluated part).

	$B+$	$B-$	
$A+$	n_{11}	n_{12}	$n_{1\cdot}$
$A-$	n_{21}	n_{22}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	

Figure 1: Agreement contingency table for two coders A and B on a binary variable.

It is often more intuitive not to consider the absolute numbers, but rather think in terms of the corresponding proportions $p_{ij} = n_{ij}/n$ (our notation here follows Fleiss et al. (1969), and so do the definitions of p_o and p_c below). The various proportions in this contingency table add up to one: $p_{11}+p_{12}+p_{21}+p_{22} = p_{1\cdot}+p_{2\cdot} = p_{\cdot 1}+p_{\cdot 2} = 1$

An intuitive approach to measuring intercoder agreement between A and B is simply to count the number of candidates for which A and B made the same choice (either both TP or both FP); this is given by the number

$$n_o := n_{11}+n_{22} \tag{1}$$

(the o stands for "observed"); the corresponding proportion

$$p_o := p_{11}+p_{22} \tag{2}$$

is easy to interpret, and should ideally give a value close to 100% for an unambiguous and reproducible annotation task. However, annotation of

[2] In the sense that it does not depend on the status of the annotation database at a specific point in time, i.e., adding new candidates to the database should not make it difficult or impossible to reproduce the set evaluated here.

[3] The *Frankfurter Rundschau* corpus is a German newspaper corpus, and is part of the ECI Multilingual Corpus 1 distributed by ELSNET. ECI stands for European Corpus Initiative, and ELSNET for European Network in Language and Speech. See `http://www.elsnet.org/resources/eciCorpus.html` for details.

[4] Since there can be little doubt that all coders would agree on the candidates that have been filtered out, this step makes the annotation task more difficult, potentially reducing intercoder agreement.

[5] (Green, 1997) gives an account of a generalized measure of intercoder agreement where there are more than two coders and/or categories.

language data is typically not unambiguous, therefore a 100%-level cannot be reached in practice. Another argument comes from mathematics:

Assume that two student assistants were hired for annotation work. Lazy as humans are, each one of them just marks $p = 5\%$ of the candidates as TPs. The students work independently from each other, so they cannot coordinate their choices, which will therefore only agree by coincidence. The average proportion of candidates on which the students agree is $p \cdot p$ (both TP) $+ (1-p) \cdot (1-p)$ (both FP) $= 90.5\%$. A similar argument has also been given by Carletta (1996).

Taken at face value, an agreement rate of 90% seems to indicate high reliability: it is equal to or better than the agreement rates reached by seasoned and well-motivated experts on similar tasks. Carletta (1996) concludes that it is necessary to correct for agreement "by chance", suggesting the widely-accepted kappa statistic (Cohen, 1960). Before we discuss the properties of kappa in Section 4, let us call attention to a terminological confusion which has led many researchers to embrace the intuitive approach outlined above.

What we want to measure is **true agreement**, where both annotators unanimously make the same decisions based on the annotation manual.

What is often counted under the name "agreement" is the number of candidates for which both annotators *make the same choice*, i.e. the **surface agreement**.

Some of these identical choices will reflect true agreement, but in other cases the agreement will be due to annotation errors or the coders may have had *different reasons* for their decisions. The surface agreement of their choices being pure coincidence[6] (which we call **chance agreement**).

We can summarise this situation in the symbolic equation:

surface agreement =
true agreement + *chance agreement*

The problem, however, is that we can only measure surface agreement. There is no direct evidence on whether identical annotations (by *A* and *B*) are due to true agreement or merely due to chance agreement. Therefore, it is necessary to correct for chance agreement in order to arrive at a meaningful interpretation of the observed surface agreement. In Sections 4 and 5 we will discuss two different approaches to this problem.

4 The kappa statistic

The standard method (cf. Agresti (1990)) to correct the observed surface agreement for chance is the kappa statistic $\hat{\kappa}$ (Cohen, 1960; Fleiss et al., 1969). Cohen (1960) estimates the amount of chance agreement from the proportions $p_{1\cdot}$ and $p_{\cdot 1}$ of candidates accepted by *A* and *B*. If their choices were entirely independent, then the average proportion of candidates for which both make the same decision would be

$$\begin{aligned} p_c &= p_{1\cdot} \cdot p_{\cdot 1} + p_{2\cdot} \cdot p_{\cdot 2} \\ &= (p_{11} + p_{12})(p_{11} + p_{21}) \\ &\quad + (p_{21} + p_{22})(p_{12} + p_{22}) \end{aligned} \tag{3}$$

(The first term is the expected proportion of candidates that both coders accept, and the second term the expected proportion that both coders reject.)

The kappa statistic is defined as the observed proportion of agreement p_o minus the expected proportion of chance agreement p_c scaled to a standard range

$$\hat{\kappa} := \frac{p_o - p_c}{1 - p_c} \tag{4}$$

As already stated in the introduction, the values of $\hat{\kappa}$ usually lie in the range $[0,1]$.

Mathematically speaking, $\hat{\kappa}$ is a test statistic and can be used to validate the null hypothesis H_0 that the observed agreement is entirely due to coincidence. In other words, that the annotation is not reproducible at all. A significant amount of evidence against H_0 only tells us that there is *some* true agreement, however small it may be. This is hardly a satisfactory result as H_0 is quite unrealistic in the first place. We would rather like to show that there is a substantial or even high degree of true agreement. As Agresti (1990) puts it: "It is rarely plausible that agreement is no better than expected by chance. Thus, rather than testing $H_0 : \kappa = 0$, it is more important to estimate strength of agreement, by constructing a confidence interval for κ." (Agresti, 1990, 367).

While $\hat{\kappa}$ is confined to an interval whose end points have a clear interpretation (0 = chance agreement, 1 = perfect agreement), it is much less obvious how to classify the values in between. Various

[6] Note that, notwithstanding the tongue-in-cheek example of lazy students above, agreement by coincidence does *not* necessarily imply that the coders make entirely random choices. It rather means that there is no common ground for these decisions (so that they are statistically independent).

scales have been suggested for the interpretation of $\hat{\kappa}$ in different areas of application (see Section 1). These scales are largely (if not solely) based on experience. Many researchers make the mistake of comparing the observed value of $\hat{\kappa}$ directly with the critical values of these scales. For instance, Di Eugenio and Glass (2004) argue that the minuscule difference between (4), which they call κ_{Co}, and a simplified version $\kappa_{\text{S\&C}}$ (Di Eugenio and Glass, 2004, 98) may decide between the rejection and acceptance of a set of annotations, giving an invented example where the values happen to lie on either side of Krippendorff's threshold of .67. These researchers fail to recognise the *uncertainty* inherent in the value $\hat{\kappa}$, which is obtained from the observed data by a *statistical calculation*. Formally, $\hat{\kappa}$ is a random variable in a statistical model of agreement. Therefore, it is inappropriate to compare $\hat{\kappa}$ to a fixed scale of critical values *without* taking this uncertainty into account, which is quantified by its standard error $\hat{\sigma}$.

Fleiss et al. (1969, 325) present the following expression as a large-sample estimate for the variance of $\hat{\kappa}$:

$$\begin{aligned} n(1-p_c)^4 \cdot \hat{\sigma}^2 = & \\ & p_{11}\big((1-p_c) - (p_{\cdot 1} + p_{1\cdot})(1-p_o)\big)^2 \\ & + p_{22}\big((1-p_c) - (p_{\cdot 2} + p_{2\cdot})(1-p_o)\big)^2 \\ & + (1-p_o)^2 p_{12}(p_{\cdot 1} + p_{2\cdot})^2 \\ & + (1-p_o)^2 p_{21}(p_{\cdot 2} + p_{1\cdot})^2 \\ & - (p_o p_c - 2p_c + p_o)^2 \end{aligned} \tag{5}$$

From (5), we obtain $\hat{\sigma}$ as an indicator for the amount of uncertainty in the observed value $\hat{\kappa}$, i.e., to what extent the sample estimate $\hat{\kappa}$ may deviate from the "true" value κ. Under the usual normality assumption, an approximate 95% confidence interval (which means that the "true" κ lies somewhere in this interval with 95% certainty) is given by

$$[\hat{\kappa} - 1.96\,\hat{\sigma}, \hat{\kappa} + 1.96\,\hat{\sigma}]$$

(Porkess, 1991, s.v. *confidence interval*)

The observed values of $\hat{\kappa}$ for the experiment described in Section 2 are presented in Section 6. The ensuing discussion is based on their confidence intervals. Depending on whether the lower or the upper bound is used, conclusions range from mediocre (or even poor) to high agreement. However, this does still not offer us an intuitive interpretation of the "true" kappa value κ. All we have to go by are apparently arbitrary critical points from one of the various scales.

In the following section, we present a different statistical approach that provides a direct estimate of the proportion of true agreement.

5 Estimating the rate of true agreement

The definition of the kappa statistic in Section 4 starts from the assumption that there is only chance agreement between the annotators. This is the null hypothesis when $\hat{\kappa}$ is used as a test statistic.

In the present section, we consider a more realistic model that divides the observed surface agreement into true and chance agreement. In this model, we assume that the test set can be divided into m candidates where the coders A and B reach true agreement (set C_1), and the remaining $n-m$ candidates (set C_2) where any observed agreement is pure coincidence. Let us call this the **dual model** of agreement.[7]

The goal of our model is to estimate the **proportion m/n of true agreement**. The result is a confidence interval including all values m/n for which the dual model provides a satisfactory explanation of the observed agreement data (Figure 1). The basic procedure to construct the confidence interval works as follows:

1. for every possible value $m \in \{0, \ldots, n\}$, divide the test set into sets C_1 (true agreement) with m candidates and C_2 (chance agreement) with the remaining $n-m$ candidates;
2. apply the methods of Section 4 (or a similar statistical test) to the set C_2; if chance agreement cannot be ruled out as an explanation for the observed agreement in C_2, the corresponding proportion m/n is included in the confidence interval.

For illustration, Figure 2 shows an agreement contingency table for an invented test set with $n = 100$ (this example is taken from Di Eugenio and Glass (2004, 99)).

Figure 3 gives a division of this test set into C_1 and C_2 when there is (hypothesized) true agreement on exactly $m = 40$ candidates. Note that the off-diagonal cells are always zero for C_1. Unfortunately,

[7] No need to say that the dual model is a strong simplification of the issue. For a better understanding of the collocation identification task, psycholinguistic studies would be required instead of mere frequency counts on intercoder agreement.

	$B+$	$B-$	
$A+$	40	15	55
$A-$	20	25	45
	60	40	

Figure 2: Invented example of an agreement contingency table for $n = 100$.

this division is not unique. Figure 4 gives another possibility with $m = 40$. Note that the latter is quite "extreme" because the coders reach true agreement only for TPs.

C_1	$B+$	$B-$
$A+$	25	0
$A-$	0	15

C_2	$B+$	$B-$
$A+$	15	15
$A-$	20	10

Figure 3: A division of the test set of Figure 2 into sets C_1 and C_2, assuming that there are $m = 40$ candidates with true agreement.

C_1	$B+$	$B-$
$A+$	40	0
$A-$	0	0

C_2	$B+$	$B-$
$A+$	0	15
$A-$	20	25

Figure 4: Another possible division of Figure 2 into sets C_1 and C_2, also with true agreement on $m = 40$ candidates.

The fundamental problem faced by this procedure is that we do not know *exactly how* to divide the test set for given m, i.e., how many of the TPs the set C_1 should contain (cf. Figures 3 and 4). The results of Step 2 above, however, depend crucially on the particular division that is made.

This ambiguity can be resolved in two different ways: (a) obtain **conservative estimates** by trying *all possible* divisions; (b) make **additional assumptions** about the proportion of TPs in the set C_1. In the absence of any concrete evidence or convincing theoretical motivation for the additional assumptions required by solution (b), most statisticians opt for the conservative approach (a). We are also in favour of (a), in principle, but the range of possible divisions will often include very extreme situations (cf. Figure 4), so that the resulting confidence intervals will be too large to be of any practical use. For this reason, we will compute and present confidence intervals for both approaches. In Sections 5.1 and 5.2, we describe the methods used to determine whether a given amount m of true agreement is consistent with the dual model.

5.1 Conservative estimates

Let m_+ stand for the number of TPs in C_1, and $m_- = m - m_+$ for the number of FPs (recall that the choices of A and B must be unanimous for all candidates in C_1); m_+ and m_- are the diagonal cells of the agreement contingency table of the set C_1 (as shown e.g. in Figure 3); for given m, the contingency tables of C_1 and C_2 are fully determined by the value of m_+.

From the constraint that the diagonal cells of both contingency tables must be non-negative, we obtain the following limits for the number of TPs in m:

$$\max\{0, m - n_{22}\} \leq m_+ \leq n_{11} \; .$$

In the conservative approach, the proportion m/n has to be included in the confidence interval when there is *any* number m_+ for which the resulting contingency table of C_2 does not provide significant evidence against chance agreement.

The statistic $\hat{\kappa}$ from Section 4 can be used (in conjunction with its approximate standard error) to test the null hypothesis $H_0 : \kappa = 0$ on C_2. Alternatively, Fisher's exact test (Fisher, 1970, 96f) can be used, which does not rely on large-sample approximations and is therefore more reliable when the hypothesized amount of chance agreement $n - m$ is small. In our evaluation, we compute two-sided p-values for Fisher's test.

5.2 The homogeneity assumption

The less conservative approach (b) does not consider all possible divisions of the test set into C_1 and C_2. In order to fix the value of m_+, an additional assumption is necessary. An example of such an assumption is *homogeneity* of C_1 and C_2, i.e., the proportion of TPs is the same in both sets (for C_2, the proportion is averaged between coders A and B). Note that we cannot make this assumption for each coder individually, since this would require the overall proportion of TPs to be the same for A and B (i.e., $p_{1\cdot} = p_{\cdot 1}$).

For a given value m, we can now determine m_+ from the homogeneity assumption and continue with the test procedure described in Section 5.1; m/n is only included in the confidence interval when this particular division of the test set is consitent with chance agreement in C_2.

In our implementation, we compute p_c (wrt. C_2) directly from the homogeneity assumption and the observed values $p_{1\cdot}$ and $p_{\cdot 1}$ in the full test set. Then, we apply a two-sided binomial test, comparing the

observed amount of agreement in C_2 with the expected proportion p_c.

6 Presentation and discussion of the evaluation results

For each pairing BK vs. NN we have computed the following values: (i) kappa according to (Cohen, 1960), the standard deviation for kappa according to (Fleiss et al., 1969) and the respective min and max values for the confidence interval (the confidence intervals are not shown here); (ii) a conservative confidence interval for the rate of true agreement (prop.min, prop.max), cf. Section 5.1; (iii) a confidence interval for the rate of true agreement using the homogeneity assumption (homogeneity min./max.), cf. Section 5.2.

Two major results can be derived from the data: (1) the conservative estimate is practically useless because the intervals between prop.min and prop.max are extremely broad, see Table 1; (2) the homogeneity estimate leads to confidence intervals that can be matched to those for the kappa value, see Table 2. Thus, the homogeneity assumption opens up an alternative to the kappa statistic.

Pursuing a conservative approach, i.e., looking at the min values of the homogeneity interval, we find roughly four groups: G1 with $min > 68\%$, G2 with $60\% < min < 65\%$, G3 with $51\% < min < 57\%$, and G4 with $min < 34\%$. G1 and G4 are the two extremes of intercoder agreement in our experiment. As regards G4, the low min values and the broad confidence intervals[8] for the homogeneity estimate indicate poor agreement between BK and the annotators NN8 and N12, whereas high min values and rather small confidence intervals in G1 provide evidence for strong intercoder agreement between BK, and NN7, as well as BK and NN9. In between, there is a broad field with the groups G2 and G3. Adapting Krippendorff's interpretation scheme based on kappa values (see Section 1) to the lower bounds of our homogeneity intervals we have a new cut-off threshold with homogeneity $min < 60\%$ indicating no intercoder agreement beyond chance, and the values above showing tentative to good agreement.

We also find that high intercoder agreement has been achieved by trained linguists while non-expert annotators achieve clearly lower agreement with BK's annotations. This could be an artefact of the data sets, i.e., that the different parts of the test set were more or less hard to annotate. For instance, NN7 and NN9 worked on the same data set. However, the intercoder results in the upper middle field (NN1, NN4, NN10), stem from the remaining three parts of the test set. Moreover, from practical work we know that trained linguists find it more easy to distinguish collocations from non-collocations than non-experts do. This also means that annotating collocation data is a critical task and that we must rely on expert knowledge for the identification of TPs.

BK vs. NN	prop.min	prop.max
1	12.16%	86.08%
2	9.24%	85.03%
3	13.92%	78.95%
4	13.33%	85.26%
5	11.46%	84.33%
6	12.63%	83.74%
7	13.92%	89.36%
8	12.51%	59.18%
9	12.98%	88.19%
10	13.72%	84.76%
11	11.61%	78.55%
12	4.34%	60.73%

Table 1: Conservative estimates for intercoder agreement between BK and 12 annotators (NN1 ... NN12).

BK vs. NN	kappa value	homogeneity min	homogeneity max	interval size
7	.775	71.93%	82.22%	10.29
9	.747	68.65%	79.77%	11.12
10	.700	64.36%	75.85%	11.49
4	.696	64.09%	75.91%	11.82
1	.692	63.39%	75.91%	12.52
6	.671	61.05%	73.33%	12.28
5	.669	60.12%	72.75%	12.63
2	.639	56.14%	70.64%	14.50
11	.592	52.40%	65.65%	13.25
3	.520	51.70%	64.33%	12.63
8	.341	33.68%	49.71%	16.03
12	.265	17.00%	35.05%	18.05

Table 2: kappa and homogeneity estimates for the intercoder agreement between BK and NN1 ... NN12.

[8] The broader the confidence interval the larger is the statistical uncertainty.

7 Conclusion

We have argued that kappa values on intercoder agreement are hard to interpret, mainly for the following two reasons. (1) The kappa statistic is based on the assumption that intercoder agreement is only due to chance (H_0), which is rather implausible from the point of view of linguistics. (2) When interpreting kappa values, confidence intervals are typically ignored, which can easily lead to a mis- or over-interpretation of the results. As an alternative to kappa, we have introduced an approach based on a distinction between true and chance agreement. We have then calculated confidence intervals for the proportion of true agreement (among the observed surface agreement) between to coders on a binary variable (collocation vs. non-collocation). The resulting confidence intervals are comparable to those of the kappa statistic, opening up a way towards an alternative and more plausible and well-founded interpretation of intercoder agreement than previously available with the kappa statistic. Concerning the collocation identification task, we have obtained further evidence that the distinction between collocations and non-collocations requires linguistic expert knowledge. It is important to take this fact into account when association measures (or other collocation extraction tools) are evaluated against candidate lists that are manually classified as true and false positives.

References

Alan Agresti. 1990. *Categorical Data Analysis*. John Wiley & Sons, New York.

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

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

Barbara Di Eugenio and Michael Glass. 2004. The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101.

Stefan Evert and Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 188–195, Toulouse, France.

Stefan Evert. 2004. An on-line repository of association measures. http://www.collocations.de/AM/.

Ronald A. Fisher. 1970. *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh, 14th edition.

Joseph L. Fleiss, Jacob Cohen, and B. S. Everitt. 1969. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323–327.

Annette M. Green. 1997. Kappa statistics for multiple raters using categorical classifications. In *Proceedings of the Twenty-Second Annual SAS Users Group International Conference (online)*, San Diego, CA, March.

Hannah Kermes. 2003. *Off-line (and On-line) Text Analysis for Computational Lexicography*. Ph.D. thesis, IMS, University of Stuttgart. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), volume 9, number 3.

Brigitte Krenn and Stefan Evert. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL Workshop on Collocations*, pages 39–46, Toulouse, France, July.

Brigitte Krenn. 2000. *The Usual Suspects: Data-Oriented Models for the Identification and Representation of Lexical Collocations.*, volume 7 of *Saarbrücken Dissertations in Computational Linguistics and Language Technology*. DFKI & Universität des Saarlandes, Saarbrücken, Germany.

Brigitte Krenn. 2004. Manual zur Identifikation von Funktionsverbgefügen und figurativen Ausdrücken in PP-Verb-Listen. http://www.collocations.de/guidelines/.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA.

Wolfgang Lezius, Stefanie Dipper, and Arne Fitschen. 2000. IMSLex – representing morphological and syntactical information in a relational database. In Ulrich Heid, Stefan Evert, Egbert Lehmann, and Christian Rohrer, editors, *Proceedings of the 9th EURALEX International Congress*, pages 133–139, Stuttgart, Germany.

Roger Porkess. 1991. *The HarperCollins Dictionary of Statistics*. HarperCollins, New York.

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

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP)*, pages 44–49.