Word-Sense Distinguishability and Inter-Coder Agreement *

Rebecca Bruce† and **Janyce Wiebe‡**
†Department of Computer Science
University of North Carolina at Asheville
Asheville, NC 28804-8511
‡Department of Computer Science
New Mexico State University, Las Cruces, NM 88003
bruce@cs.unca.edu, wiebe@cs.nmsu.edu

Abstract

It is common in NLP that the categories into which text is classified do not have fully objective definitions. Examples of such categories are lexical distinctions such as part-of-speech tags and word-sense distinctions, sentence level distinctions such as phrase attachment, and discourse level distinctions such as topic or speech-act categorization. This paper presents an approach to analyzing the agreement among human judges for the purpose of formulating a refined and more reliable set of category designations. We use these techniques to analyze the sense tags assigned by five judges to the noun *interest*. The initial tag set is taken from Longman's Dictionary of Contemporary English. Through this process of analysis, we automatically identify and assign a revised set of sense tags for the data. The revised tags exhibit high reliability as measured by Cohen's κ. Such techniques are important for formulating and evaluating both human and automated classification systems.

Introduction

It is common in Natural Language Processing (NLP) that the categories into which text is classified do not have fully objective definitions. Examples of such categories are lexical distinctions such as part-of-speech tags and word-sense distinctions, sentence level distinctions such as phrase attachment, and discourse level distinctions such as topic or speech-act categorization. This paper presents an approach to analyzing the agreement among human judges for the purpose of formulating a refined and more reliable set of category designations.

We performed a case study of the classification process, involving multiple judges performing a word-sense disambiguation task. Table 1 presents the data for two judges assigning one of six senses to each instance of *interest* used as a noun in the corpus. The data is represented as a *contingency table*, often referred to as a *confusion* matrix; it depicts the "confusion" among the judges' classifications. Evidence of confusion among the classifications in Table 1 can be found in the marginal totals, n_{i+} and n_{+j}, where i and j range from 1 to 6. We see that, on average, judge A has a higher preference for senses 1 and 3 than judge E does, while judge E has a higher preference for sense 2 than judge A does. These *biases* are one aspect of *agreement* (or the lack of it) among judges.

A second aspect of agreement is the extent to which judges agree on the tags of individual words (*category distinguishability*). We see from the diagonal frequencies in Table 1 that these judges agree on 2097 out of 2369 of them, which is 88.5% of the individual tags.

Cohen (1960) proposed the coefficient of agreement, κ, for measuring the agreement between two judges. κ compares the actual agreement to that which would be expected if the decisions made by each judge were statistically independent (i.e., "chance agreement"). A number of previous studies have used κ to evaluate inter-coder reliability (e.g., Carletta 1996, Litman & Passonneau 1995; Moser & Moore 1995; Hirschberg & Nakatani 1996; Wiebe et al. 1997). However, in looking at agreement among judges, we are often not as concerned with describing how well two particular judges

*This research was supported by the Office of Naval Research under grant number N00014-95-1-0776.

sense 1	"readiness to give attention"
sense 2	"quality of causing attention to be given"
sense 3	"activity, subject, etc., which one gives time and attention to"
sense 4	"advantage, advancement, or favor"
sense 5	"a share (in a company, business, etc.)"
sense 6	"money paid for the use of money"

Figure 1: Noun Senses of *Interest* in LDOCE

agree as in measuring how well any observer can distinguish the categories from one another. In other words, the issue is the precision of the *classification process*.

In this paper, we present a study of a classification process. The section *Agreement Among Judges* presents an analysis of the patterns of agreement among the judges. Agreement is a function of the differences among the judges (i.e., their biases) and the distinguishability of the categories themselves. We study bias using the models for symmetry, marginal homogeneity, and quasi-independence (in the subsection *Observer Differences*). We study category distinguishability using Darroch & McCloud's (1986) *degree of distinguishability*, δ_{ij} (in the subsection *Category Distinguishability*). Guided by these analyses, in the section *Modification of the Classification Process* we investigate modifications to the classification process that improve reliability. We analyze the effects both of removing judges and collapsing categories. A technique is presented for formulating a tag set which can be automatically derived from the original tag set. The technique is successful in the study presented here: the derived tag set yields improved reliability, as measured by Cohen's κ.

The Data

The classification process performed in this study involved five human judges independently assigning sense tags to 2369 instances of the noun *interest* taken from the Wall Street Journal Treebank Corpus (Marcus et al. 1993). The senses given to the taggers, shown in Figure 1, are from the Longman's Dictionary of Contemporary English (LDOCE).

The annotation instructions were minimal. They were asked to use their judgment in assigning to each usage of *interest* the single tag that best characterizes its meaning. It is likely that more explicit tagging instructions including examples and default rules would improve agreement among judges. Indeed, an analysis of the classification process such as performed here could be used to formulate and interactively revise a set of tagging instructions, but this application is not considered here.

Five human judges, referred to as A through E, participated in the study. Two of the judges (judges C and D) were involved in the project and had participated in previous sense tagging experiments. The remaining three judges (judges A, B and E) were not members of the project and had no previous background in NLP or linguistics.

Agreement Among Judges

All of the techniques that we present for the analysis of agreement are appropriate for category classifications assigned to multiple objects (in this case, words) by two judges.[1] We analyze the agreement among all five judges by evaluating the agreement between all pair-wise combinations of these judges. We exclusively use maximum likelihood estimates of model parameters.

The Basics

Tables 1-5 present half of the data, in contingency table format. Each table is for one pair-wise combination of the five judges. The rest of the data, for the other five combinations, is available on the World Wide Web at *http://crl.nmsu.edu/Research/Projects/graphling.* In each table, the rows correspond to the senses assigned by the first judge while the columns correspond to those assigned by the second judge. Let n_{ij} denote the number of words that judge one classifies as i and judge two classifies as sense j. If we let p_{ii} be the probability that the judges will agree that a randomly selected usage is sense i, then $\sum_i p_{ii}$ is the total probability of agreement across all senses. p_{ii} can be estimated as $\frac{n_{ii}}{n_{++}}$ (a maximum likelihood estimate), and the total probability of agreement can be estimated as $\sum_i \hat{p}_{ii} = \sum_i \frac{n_{ii}}{n_{++}}$, where $n_{++} = \sum_{ij} n_{ij} = 2369$.

[1]Several of these techniques are also applicable to the analysis of multiple judges.

The simplest measure of agreement is the estimated probability of agreement, i.e., $\sum_i \hat{p}_{ii}$, where the possible values are affected by the marginal totals (i.e., the row and column totals). Cohen's κ compares the total probability of agreement to that expected if the ratings were statistically independent (i.e., "chance agreement"). That value is then normalized by the maximum possible level of agreement given the marginal distributions. The marginal distributions can be estimated from the marginal counts as: $\hat{p}_{i+} = \frac{n_{i+}}{n_{++}}$ and $\hat{p}_{+i} = \frac{n_{+i}}{n_{++}}$. The complete formulation of κ is:

$$\kappa = \frac{\sum_i \hat{p}_{ii} - \sum_i \hat{p}_{i+}\hat{p}_{+i}}{1 - \sum_i \hat{p}_{i+}\hat{p}_{+i}} \qquad (1)$$

κ is 0 when the agreement is that expected by chance, and is 1.0 when there is perfect agreement.

An extension of κ for the case of multiple judges (three or more) is presented in Davies and Fleiss (1982) and used in this study.

Analyzing Patterns of Agreement

In a classification experiment, the two judges are assumed to classify any given usage independently of each other, but it is clear in the formulation of κ that we expect the data to exhibit dependence, i.e., $\hat{p}_{ij} \neq \hat{p}_{i+} \times \hat{p}_{+j}$. Where does this dependence come from? It arises from three factors and their possible interactions: (1) the heterogeneity of the objects being classified (i.e., the usages of *interest*), (2) the heterogeneity of the judges, and (3) the distinctions made in the category definitions.

We focus on the latter two factors and their interaction. Rather than simply measuring agreement we measure the contributions to agreement made by these two factors and propose changes to the classification process based on the analysis. Just as overall agreement can be assessed as a function of the counts in the pair-wise confusion matrices, so can the measures of *observer difference* (bias) and *category distinguishability.*

Observer Differences (Bias) The hypothesis of no difference between two judges is the hypothesis of complete **symmetry** (*Sym* in Table 6), that is, $\hat{p}_{ij} = \hat{p}_{ji}$ or $\frac{\hat{p}_{ij}}{\hat{p}_{ji}} = 1$ for all i, j. If this ratio equals one for all i, j then it follows that the observers' interpretations are indistinguishable.

Complete symmetry implies marginal symmetry, that is, $\hat{p}_{i+} = \hat{p}_{+i}$. Bias of one judge relative to another is evidenced as a discrepancy between these marginal distributions. Bias decreases as the marginal distributions become more nearly equivalent. The measure of bias is the test for **marginal homogeneity** (*M.H.* in Table 6), $\hat{p}_{i+} = \hat{p}_{+i}$ for all i.

It is possible to access the similarity of two judges even when there is evidence of bias. The model for **quasi-independence** (*Q.I.* in Table 6) (Bishop et al. 1975) tests whether two judges' decisions are independent if we consider only the off-diagonal counts—the counts corresponding to disagreement (i.e., $\hat{p}_{ij} = \hat{p}_{i+} \times \hat{p}_{+j}$ for $i \neq j$). Quasi-independence holds when, given that the judges disagree, there is no pattern of association in the categories they assign.

In the tests for symmetry, marginal homogeneity, and quasi-independence, a model is formulated that enforces the hypothesized constraint, e.g., $p_{ij} = p_{ji}$ in the case of symmetry. The degree to which the data is approximated by a model is called the *fit* of the model. In this work, the fit of each model is reported in terms of the likelihood ratio statistic, G^2, and its significance. The higher the G^2 value, the poorer the fit. The fit of a model is considered acceptable if its reference significance level is greater than 0.001 (i.e., if there is greater than a 0.001 probability that the data sample was randomly selected from a population described by the model).

Category Distinguishability The ratio $\tau_{ij} = \frac{\hat{p}_{ij} \times \hat{p}_{ji}}{\hat{p}_{ii} \times \hat{p}_{jj}}$, referred to as the *diagonal cross-product-ratio*, represents the odds for disagreement over agreement on categories i, j. Darroch and McCloud (1986) define the *degree of distinguishability*, δ_{ij}, for categories i, j as:

$$\delta_{ij} = 1 - \tau_{ij} = 1 - \frac{\hat{p}_{ij} \times \hat{p}_{ji}}{\hat{p}_{ii} \times \hat{p}_{jj}} \qquad (2)$$

If $\delta_{ij} = 1$, we say that the categories are completely distinguishable, and, if $\delta_{ij} = 0$, they are completely indistinguishable.

Majority Consensus When multiple judges are involved in a study, it is possible to formulate a majority tag for each object, that is, the tag that the majority of the judges assign to each object. It represents majority opinion and is useful in identifying outlyers, as shown in the next section.

Results

Table 6 presents the results of the tests for observer differences and Table 7 presents the measures of category distinguishability. All evaluations are performed on each pair-wise confusion matrix. The columns labeled $M|A$ through $M|E$ refer to similar tables comparing the majority tag to the assignments made by each judge (e.g., judge A, in the case of $M|A$). These tables are not included in the paper.

While the κ values in Table 6 are reasonably high, the judges display bias and cannot be considered interchangeable. The only exception is the strong similarity between the majority tag and the assignments made by judge C (i.e., the column labeled $M|C$ in Table 6); these tags are symmetric and unbiased. Among the five judges, the most similar are judges C and D, the two experienced judges. While their scores for symmetry and marginal homogeneity are not significant, indicating a relative bias, their score for quasi-independence is significance (i.e., $0.004 > 0.001$, the cutoff we use to judge significance). This indicates that, although judges C and D are not indistinguishable, there is no **systematic** difference of opinion between them. Judge D also shows some similarity to the majority tag.

The judge that is least similar to the others is judge E; this is particularly evident when judge E is compared to the majority tag.

The distinguishability, δ_{ij}, of all pair-wise combinations of tags are evaluated in Table 7. All scores are at or near the maximum of 1.0, with the exception of those measuring the distinguishability of tags 1 and 2. It is particularly low in Table $A|B$ (i.e., Table 2).

Modification of the Classification Process

Based on the results presented above, we modified the classification process in two ways: (1) judge E is removed, and (2) sense tags 1 and 2 are conflated to form a single sense distinction. The poor marks for distinguishability between these senses seem to be reflected in a closeness in meaning (see in Figure 1), supporting the decision to conflate them.

Removing judge E from the study removes the tables with the lowest κ scores. As a result, the agreement among all judges increases from 0.874 to 0.898, as measured by Davies and Fleiss' extension of κ.

The process of conflating two tags is accomplished using the *latent class model* (Goodman 1974)[2]. This procedure has historically been used to identify a set of *latent* categories that explain the interdependencies among the observable categories. In this case, the observable categories are the sense tags assigned by the remaining four judges, while the latent categories correspond to the unobservable *true* meanings of the noun *interest*. Once the desired number of latent categories has been specified, these categories are assigned via the EM algorithm as described in Goodman (1974) and applied in Pedersen & Bruce (1997)[3].

Using the EM algorithm as described above, all usages of *interest* are assigned to one of five latent sense groupings. The mapping between the derived (i.e., latent) categories and the observed senses is established to maximize the correlation between latent categories and observed senses. This correlation for each judge, is estimated as part of the process of assigning latent categories. As an example, Table 10 presents the correlation for judge C. The values recorded in the table are the probabilities of judge C assigning sense tag i and the EM algorithm assigning latent tag j. As can be seen, correlation is maximized when the mapping of observed tags to latent tags is as follows: $1 \Rightarrow 1,\ 2 \Rightarrow 1,\ 3 \Rightarrow 2,\ 4 \Rightarrow 3,\ 5 \Rightarrow 4$, and $6 \Rightarrow 5$. This mapping conflates senses 1 and 2 while leaving all other senses intact. This corresponds to our expectations based on the study of agreement presented in the previous section. Using this mapping, the observer difference measures among the

[2] Also referred to as the *Naive Bayes model* (Langley et al. 1992).

[3] This is a well known unsupervised learning alobserved tagsgorithm; other notable references to this procedure are Lazarfeld (1966), Pearl (1988), and AutoClass (Cheeseman 1990).

		Latent Tag				
		1	2	3	4	5
Judge C	1	0.142	0.010	0.001	0.001	0.002
	2	0.003	0.001	0.001	0.000	0.000
	3	0.000	0.024	0.005	0.000	0.000
	4	0.001	0.000	0.074	0.001	0.000
	5	0.001	0.003	0.000	0.206	0.000
	6	0.000	0.000	0.000	0.000	0.526

Table 10: Tag Correlation for Judge C

four judges for the latent tag set are presented in Table 8, and the distinguishability of latent tags is presented in Table 9. As compared to the original classification process, the agreement among all judges increases from 0.874 to 0.916 for the revised tag set with four judges.

Recent work has proposed various methods for pruning senses for word instances and tuning tag sets to a particular domain using corpus information and existing linguistic knowledge sources (e.g., Yarowsky 1992, Jing et al. 1997, Basili et al. 1997). We have presented an automatic method for refining a tag set using an important additional source of information: the manual annotations assigned by human judges.

Conclusion

There is increasing awareness of the need to manage the uncertainty inherent in many classification systems. We have presented procedures that can be used to analyze and refine any classification system that makes use of nominal categories. These techniques can be used to study and improve the reliability of human judges as well as refine categorizations that can be applied automatically and, in the process, establish an upper bound on the accuracy of automatic classification, i.e., the agreement among the human judges. In future work, we will apply these techniques to the analysis and evaluation of automated classification systems.

References

[1] Bishop, Y. M., Fienberg, S., & Holland, P. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge: The MIT Press.

[2] Basili, R., Della Rocca, M., Pazienza, M. T. (1997). Toward a bootstrapping framework for corpus semantic tagging. In *Proc. SIGLEX Workshop on Tagging Text with Lexical Semantics*, pp. 58-65.

[3] Carletta, J. (1996). Assessing agreement on classification taks: the kappa statistic. *Computational Linguistics 22* (2).

[4] Cheeseman, P. & Stutz, J. (1996). Bayesian Classification (AutoClass): Theory and Results. In Fayyad, Piatetsky–Shapiro, Smyth, & Uthurusamy editors, *Advances in Knowledge Discovery and Data Mining*, AAAI Press/MIT Press.

[5] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psych. Meas. 20*: 37–46.

[6] Davies, M. & Fleiss, J. (1982). Measuring Agreement for Multinomial Data *Biometrics*,38:1047–1051.

[7] Darroch & McCloud. (1986). Category Distinguishability and Observer Agreement. *Austral. Journal of Statistics*, 28(3):371–388.

[8] Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231.

[9] Hirschberg, J. & Nakatani, C. (1996). A prosodic analysis of discourse segments in direction-giving monologues. In *Proc. ACL-96*, pp. 286-293.

[10] Jing, H., Hatzivassiloglou, V., Passonneau, R., and McKeown, Kathleen (1997). Investigating complementary methods for verb sense pruning. In *Proc. SIGLEX Workshop on Tagging Text with Lexical Semantics*, pp. 58-65.

[11] Langley, P., Iba, W. & Thompson, K. (1992). An analysis of bayesian classifiers. In *Proceedings of the 10th National Conference on Artificial Intelligence*, pp. 223–228.

[12] Lazarfeld, P. (1966). Latent structure analysis. In S. A. Stouffer, L. Guttman, E. Suchman, P.Lazarfeld, S. Star, and J. Claussen (Ed.), *Measurement and Prediction*, New York: Wiley.

[13] Litman, D. & Passonneau, R. (1995). Combining multiple knowledge sources for discourse segmentation. In *Proc. 33rd Annual Meeting of the Assoc. for Computational Linguistics*, MIT, pp. 130-143.

[14] Marcus, M., Santorini, B., & Marcinkiewicz, M. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics 19* (2): 313–330.

[15] Moser, M. & Moore, J. (1995). Investigating cue selection and placement in tutorial discourses. In *Proc. 33rd Annual Meeting of the Assoc. for Computational Linguistics*, MIT, pp. 130-143.

[16] Pearl, J. (1988). *Probabilistic Reasoning In Intelligent Systems: Networks of Plausible Inference.* San Mateo, Ca.: Morgan Kaufmann.

[17] Pedersen, T. & Bruce, R. (1997). Distinguishing Word Senses in Untagged Text. *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-97)*, August 1997.

[18] Wiebe, J., O'Hara, T., McKeever, K., and Öhrström-Sandgren, T. (1997). An empirical approach to temporal reference resolution. *Proc. 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP-97)*, Association for Computational Linguistics, Brown University, August 1997, pp. 174–186.

[19] Yarowsky, D. (1992). Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proc. COLING-92.*

		$Judge\ 2 = E$						
		$sense1$	$sense2$	$sense3$	$sense4$	$sense5$	$sense6$	
	$sense1$	$n_{11} = 174$	$n_{12} = 115$	$n_{13} = 11$	$n_{14} = 8$	$n_{15} = 6$	$n_{16} = 2$	$n_{1+} = 316$
	$sense2$	$n_{21} = 7$	$n_{22} = 8$	$n_{23} = 1$	$n_{24} = 2$	$n_{25} = 1$	$n_{26} = 0$	$n_{2+} = 19$
$Judge\ 1$	$sense3$	$n_{31} = 25$	$n_{32} = 24$	$n_{33} = 40$	$n_{34} = 12$	$n_{35} = 4$	$n_{36} = 3$	$n_{3+} = 108$
$= A$	$sense4$	$n_{41} = 3$	$n_{42} = 1$	$n_{43} = 3$	$n_{44} = 156$	$n_{45} = 8$	$n_{46} = 1$	$n_{4+} = 172$
	$sense5$	$n_{51} = 1$	$n_{52} = 1$	$n_{53} = 6$	$n_{54} = 12$	$n_{55} = 474$	$n_{56} = 6$	$n_{5+} = 500$
	$sense6$	$n_{61} = 0$	$n_{62} = 0$	$n_{63} = 1$	$n_{64} = 2$	$n_{65} = 6$	$n_{66} = 1245$	$n_{6+} = 1254$
		$n_{+1} = 210$	$n_{+2} = 149$	$n_{+3} = 62$	$n_{+4} = 192$	$n_{+5} = 499$	$n_{+6} = 1257$	$n_{++} = 2369$

Table 1: Confusion Matrix for Judges A and E

		$Judge\ 2 = B$						
		1	2	3	4	5	6	
	1	242	37	21	7	8	1	316
	2	13	2	1	1	1	1	19
$Judge\ 1$	3	32	5	53	15	1	2	108
$= A$	4	2	0	1	161	6	2	172
	5	3	0	20	16	458	3	500
	6	0	0	1	1	6	1246	1254
		292	44	97	201	480	1255	2369

Table 2: Confusion Matrix for Judges A and B

		$Judge\ 2 = C$						
		1	2	3	4	5	6	
	1	303	2	0	6	3	2	316
	2	10	6	1	1	1	0	19
$Judge\ 1$	3	42	3	56	5	1	1	108
$= A$	4	4	0	8	154	6	0	172
	5	4	0	1	13	480	2	500
	6	5	1	1	1	5	1241	1254
		368	12	67	180	496	1246	2369

Table 3: Confusion Matrix for Judges A and C

		$Judge\ 2 = D$						
		1	2	3	4	5	6	
	1	342	1	2	2	12	9	368
	2	1	10	0	0	0	1	12
$Judge\ 1$	3	2	1	48	12	3	1	67
$= C$	4	8	1	3	160	7	1	180
	5	4	0	0	0	489	3	496
	6	1	0	0	0	0	1245	1246
		358	13	53	174	511	1260	2369

Table 4: Confusion Matrix for Judges C and D

		$Judge\ 2 = E$						
		1	2	3	4	5	6	
	1	206	131	11	6	7	7	368
	2	0	11	0	0	0	1	12
$Judge\ 1$	3	1	6	42	13	2	3	67
$= C$	4	1	1	5	164	8	1	180
	5	1	0	4	7	481	3	496
	6	1	0	0	2	1	1242	1246
		210	149	62	192	499	1257	2369

Table 5: Confusion Matrix for Judges C and E

Test	$A\|E$	$A\|B$	$A\|C$	$A\|D$	$B\|C$	$B\|D$	$B\|E$	$C\|D$	$C\|E$	$D\|E$	$M\|A$	$M\|B$	$M\|C$	$M\|D$	$M\|E$
Sym.: G^2	165	70	77	75	105	101	109	46	226	214	81	84	22	39	212
Sig.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.102	0.001	0.000
M. H.: G^2	150	30	47	58	69	79	90	37	213	210	64	42	15	39	206
Sig.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.010	0.000	0.000
Q. I.: G^2	154	143	79	61	94	81	186	42	135	120	67	82	34	25	120
Sig.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.000	0.016	0.051	0.000
Kappa	0.825	0.866	0.916	0.903	0.882	0.873	0.821	0.951	0.856	0.849	0.929	0.901	0.977	0.964	0.874

Table 6: Tests of Observer Differences (Bias) for Five Judges and Six Senses

Senses	$A\|E$	$A\|B$	$A\|C$	$A\|D$	$B\|C$	$B\|D$	$B\|E$	$C\|D$	$C\|E$	$D\|E$	$M\|A$	$M\|B$	$M\|C$	$M\|D$	$M\|E$
1 − 2	0.422	0.006	0.989	0.986	0.765	0.662	0.183	1.000	1.000	1.000	0.990	0.675	1.000	1.000	1.000
1 − 3	0.960	0.948	1.000	0.997	0.959	0.964	0.950	1.000	0.999	0.997	1.000	0.968	1.000	1.000	0.999
1 − 4	0.999	0.999	1.000	0.999	0.999	0.999	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1 − 5	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1 − 6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2 − 3	0.925	0.953	0.991	0.978	0.964	0.979	1.000	1.000	1.000	1.000	0.988	0.966	1.000	1.000	1.000
2 − 4	0.998	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2 − 5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2 − 6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3 − 4	0.994	0.998	0.995	0.994	0.997	0.994	0.986	0.995	0.991	0.993	0.999	0.998	0.999	0.999	0.996
3 − 5	0.999	0.999	1.000	1.000	1.000	1.000	0.994	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000
3 − 6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
4 − 5	0.999	0.999	0.999	0.999	1.000	1.000	0.999	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000
4 − 6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
5 − 6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 7: Measure of Category Distinguishability for Five Judges and Six Senses

Test	$A\|B$	$A\|C$	$A\|D$	$B\|C$	$B\|D$	$C\|D$	$M\|A$	$M\|B$	$M\|C$	$M\|D$
Sym.: G^2	56	63	63	72	70	44	72	72	17	36
Sig.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.068	0.000
M. H.: G^2	19	39	52	38	53	37	57	43	7	29
Sig.	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.136	0.000
Q. I.: G^2	72	68	50	46	23	37	60	37	26	19
Sig.	0.000	0.000	0.000	0.000	0.016	0.000	0.000	0.000	0.006	0.017
Kappa	0.898	0.924	0.910	0.909	0.902	0.952	0.943	0.926	0.978	0.964

Table 8: Tests of Observer Differences (Bias) for Four Judges and Five Senses

Senses	$A\|B$	$A\|C$	$A\|D$	$B\|C$	$B\|D$	$C\|D$	$M\|A$	$M\|B$	$M\|C$	$M\|D$
1 − 2	0.948	0.997	0.994	0.957	0.964	1.000	1.000	0.968	1.000	1.000
1 − 3	1.000	0.999	0.999	0.999	0.999	1.000	1.000	1.000	1.000	1.000
1 − 4	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000
1 − 5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2 − 3	0.998	0.995	0.993	0.997	0.994	0.995	0.999	0.998	1.000	0.997
2 − 4	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2 − 5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3 − 4	0.999	0.999	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000
3 − 5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
4 − 5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 9: Measure of Category Distinguishability for Four Judges and Five Senses