Laughter and Other Non-Verbal Vocalisations Workshop 2024
16-17 July 2024, Belfast, United Kingdom

Obtaining Agreement for Conversational Laughter Function Annotation

Bogdan Ludusan

Phonetics Workgroup, Faculty of Linguistics and Literary Studies & CITEC,
Bielefeld University, Germany

bogdan.ludusan@uni-bielefeld.de

Abstract

We present a preliminary study into the use of majority agreement between several raters for laughter function annotation of conversational speech data. Six annotators rated laughter events as belonging to one of eight classes, having had knowledge of the immediately preceding context. Computing three measures of inter-rater agreement, we noticed low values for each one of them, denoting considerable disagreement between raters. We then explored whether taking the majority class label could be used instead. The fact that over 80% of the annotated laughter instances were labelled the same by a majority of raters suggests that this option would be feasible.

Index Terms: laughter, annotation, laughter function, mirthful laughter, social laughter

1. Introduction

Laughter is best-known for its function of expressing humour appreciation (mirthful laughter; e.g., [1]), although it plays a variety of roles in human interaction. Conversational laughter may express a number of social functions (see [2] for an overview), having both positive (e.g., affiliation towards the interlocutor) and negative (laughing at someone) dimensions, or less clearly-defined valence (e.g., softening of previous statements). More recent research has also shown that laughter may have linguistic functions, for instance by demarcating various conversation levels (e.g., topics [3], turns [4]), expressing propositional content [5] or as backchannel instance [6].

Many of the studies exploring laughter do not consider its function. Even the ones that take into account laughter function, the majority do not give information on function annotation agreement, either because it was performed by a single annotator (e.g., [7]) or, in the case when more annotators were employed, no inter-rater agreement was computed/reported (e.g., [8]). Among the few studies that calculated rater agreement, an earlier one [9] reported moderate Cohen's kappa values (ranging from 0.45 to 0.54), for the pairwise agreement between three annotators, when labelling nine different laughter classes (mirthful, social, embarrassed, self-conscious, inviting, contagious, derogatory, dumbfounded, and untrue). More recent investigations have reported low inter-annotator agreement rates for a lower number of laughter classes (2 [10, 11] and 4 [12], respectively). The study by Rychlowska and colleagues [12], involving a perceptual experiment with over 200 participants, asked them to recognize four types of laughter: amused, embarrassed, schadenfreude and other. All laughter types were recognized at chance level in the conditions not giving access to audio-visual data from both interlocutors. Cohen's kappa values of 0.28 [10] and 0.36 [11] respectively, were obtained between two annotators classifying mirthful versus discourse/social laughter. [5] employed a workflow based on decision trees for deciding the laughter function. They reported higher agreement values for each laughter class separately, varying between 0.15 and 0.73 Krippendorff's alpha.

We investigated here the laughter annotation agreement rate obtained on conversational speech, when considering two (mirthful/social) or more (social laughter divided in several subclasses) laughter classes. We employed more than two annotators, in order to explore whether taking the majority class label would be a feasible option for obtaining laughter function annotation for conversational speech data. Moreover, taking into account that acoustic differences between various types of laughter (spontaneous/volitional [13, 14] or mirthful/social [10, 15]) have been found, we analyzed if the acoustic realization of the laughter events had an effect on the majority class label.

2. Analyses

We employed materials from the ALICO corpus [16], containing dyadic interactions between German native speakers. One of the interlocutors was tasked with telling a vacation story, while their conversation partner listened, provided feedback and asked questions about the story. The laughter events (both laughs and speech-laughs) produced during the interaction were identified by an expert annotator and segmented. For this study, we considered only the speech uttered by the storyteller and we selected two of the conversations (5 min-6 min long each) in which the person laughed more. 35 laughter events (19 from the first conversation and 16 from the second one), consisting of 8 laughs and 27 speech-laughs were included in the analysis.

Six annotators listened to the 35 events and their immediately preceding context (ranging between 5 s and 10 s), in the order in which they appeared in conversation. Thus, they had not only the immediate context of the laughter event available to them, but also the larger context of the conversation. They were asked to determine the function of the first laughter event they heard in each stimulus file. They had to choose between several classes that express either mirth or social functions (closeness, embarrassed, pleasantness, polite, remedy, softening, and other [2]) of laughter. We also explored the annotation agreement in the case of only two classes (mirthful/social), by collapsing all labels but mirthful into one class (social).

We computed three measures of inter-rater agreement suitable for more than two raters: percentage (%) agreement, Light's kappa, and Krippendorff's alpha, for each of the two cases (8 and 2 classes). The first measures is defined as the proportion of observations which are assigned the same label by all raters, while the second one represents the mean of obtained Cohen's kappa values of all pairwise combinations of the raters. Cohen's kappa is a more reliable measure than % agreement, as

10.21437/LW.2024-3

Table 1: *The inter-rater agreement obtained for laughter function annotation, across six raters, when considering either eight classes or two classes (merging all non-mirthful laughter classes into one). We considered three measures: percentage agreement (% agree), Light's kappa and Krippendorff's alpha.*

# classes	**% agree**	**Light's κ**	**Krippendorff's α**
8	5.71	0.199	0.185
2	22.9	0.192	0.179

it also considers in its calculation the agreement due to chance. Krippendorff's alpha takes into account the disagreement due to chance, as well, having the additional advantage of being suitable for multiple raters. The implementations offered by the *irr* package [17] of the R software [18] were employed. In order to investigate how much the annotators agree on the annotated laughter events, we determined the majority label assigned by the six raters. If at least half of them chose one label, this was considered as the majority label, otherwise there was no agreement (also when the vote was split 3/3). We then computed how many of the laughter instances received a majority label.

Finally, we extracted several acoustic features (see [13, 14, 15]) from the recordings, namely the fundamental frequency (f0) of the voice, the energy of the speech signal and cepstral peak coefficient (cpp), a measure characterizing voice quality. The VoiceSauce software [19] was employed with default parameters, to obtain f0 values (Straight algorithm), the root-mean-square-energy and cpp. We determined the mean, maximum, and range of these features within the annotated laughter events, as well as the mean of these events normalized by the mean of the same features within the inter-pausal unit immediately preceding the laughter event (meanNorm). We then fitted logistic regression models with the majority label as dependent variable and the f0, energy, cpp as predictors. Separate models (8 total) were fitted for the two cases of majority labels (2 vs. 8 classes), with each of the four measurements (mean, maximum, range, normalized mean). All analyses were run in R.

3. Results

First, we had a look at the distribution of labels across raters. There was substantial variation between raters, with two of them assigning everything into 4 classes, one using all 8 labels for classification, while the remaining three employing 6 or 7 classes. Moreover, there were very few uses of the label "other". Comparing the label distribution pairwise between all raters, by means of Fisher's exact tests with Bonferroni correction, 6 of 15 pairs showed significant differences (three of those involving the same rater).

The inter-rater agreement results are illustrated in Table 1, both when considering the eight original laughter classes, and in the case of two classes (mirthful/social). We can see that all measures exhibited low values. For the Light's κ measure, the employed R package also provided significance testing: neither in the 8 classes case ($z = 0.781, p = 0.435$), nor in the 2 classes one ($z = 0.011, p = 0.991$) was the agreement significantly different from chance level. The rater pairwise Cohen's κ value varied between [0.102, 0.291] for the 8-class case and between [0.028, 0.388] for 2 laughter classes.

Looking next at the results obtained for the majority class (8 classes), we see in Figure 1 that most of the laughter instances rated here (29/35) achieved a majority vote (>= 3 votes, no tie for 3 votes). Although, overall, all the labels were employed by the annotators, only five of them reached majority vote. A similar number of mirthful and social laughter events were assigned a majority label. In the two-class case, 31 of the 35 laughter events were given the same class by a majority of raters.

The analysis of the role of acoustic-prosodic information in the rating of the majority class, revealed only one significant effect: cpp, for the two-class model fitted with the meanNorm values (ANOVA type II: $\chi^2 = 6.63, df = 1, p = 0.01$).

4. Discussion and conclusions

Our raters attained low agreement levels, similar to those reported by [12, 10, 11] (but see [9, 5]), with no change between 8 and 2 classes, indicating confusion between social and mirthful laughter. These results seem to support the hypothesis of an underdetermined nature of laughter when it comes to its classification [20]. It still remains to be ascertained if this ambiguity is due to how laughter function is normally obtained, judged by observers to the interaction, and if it is less ambiguous for the conversation participants themselves. Despite low agreement rates, we observed that by taking the majority class, we can obtain annotations for most laughter instances. While the distribution of mirthful and social laughter majority label did not differ (see Figure 1), some types of social laughter (e.g., embarrassed) exhibited higher agreement than others, in line with results reported by [5].

We noticed individual variation in perceived laughter function, based not only on the low inter-rater agreement values, but also on the rater whose label distribution differed significantly from those of most other raters. This suggests that employing a low number of annotators (such as in [9, 10, 11]) might pose problems for obtaining reliable majority class labels. This can be seen by comparing the number of cases where no majority label could be obtained for 2 classes (11%) with those in [10] (34%). On the other hand, considering a very large pool of raters, as in [12], is not feasible from a resources point of view. Therefore, having 5-10 (or up to 16, as in [8]) ratings per laughter instance might be sufficient to obtain cost-effective and reliable laughter function based on majority label, while avoiding the risks posed by high individual variation.

Only voice quality (meanNorm cpp) had an effect on the majority class (two-class case), implying that the raters based their choice almost exclusively on higher-level (e.g., semantic, pragmatic) information. It might be that each rater uses a different weighting of the acoustic features to decide on a class (similar to other speech phenomena, such as prominence [21]) and by taking a majority vote across speakers, we cannot untangle these strategies. Alternatively, it may be due to our dataset containing a higher proportion of speech-laughs (77%) than found in other conversations (23%-44% as in [4]) and laughter type differences have been studied and documented only for laughs [13, 14, 15]. It would be worthwhile investigating if laughs and speech-laughs tend to be used with the same function in conversation and if the acoustic marking of the different function classes is less discernible for speech-laughs than for laughs.

To conclude, our preliminary study confirmed findings about low inter-rater agreement for laughter function, while also showing that the use of majority decision between several raters may be feasible to obtain larger amounts of laughter function annotations. We acknowledge the limited size and reduced interactional setting of the employed dataset and future work will try to extend these findings to larger and more spontaneous data. Also, it would be interesting examining how a stricter annotation workflow [5] would improve the annotation reliability.

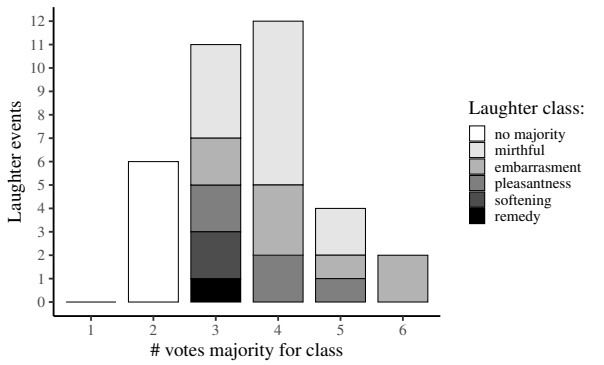

Figure 1: *Number of laughter instances obtaining a majority vote, for each of the eight laughter function classes. No instances were assigned a different class by each rater (first bar of the plot) and only six instances received a maximum of two for the most rated class, meaning no majority obtained.*

5. Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number 461442180. The author would like to thank the six annotators for their help.

6. References

[1] H. Foot, Ed., *Humor and laughter: Theory, research and applications*, 2nd ed. Routledge, 1996.

[2] P. Glenn, "Towards a social interactional approach to laughter," in *Laughter in Interaction*. Cambridge University Press, 2003, pp. 7–34.

[3] F. Bonin, N. Campbell, and C. Vogel, "Laughter and topic changes: Temporal distribution and information flow," in *Proc. of CogInfoCom*. IEEE, 2012, pp. 53–58.

[4] B. Ludusan and P. Wagner, "Laughter entrainment in dyadic interactions: Temporal distribution and form," *Speech Communication*, vol. 136, pp. 42–52, 2022.

[5] C. Mazzocconi, Y. Tian, and J. Ginzburg, "What's your laughter doing there? a taxonomy of the pragmatic functions of laughter," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1302–1321, 2022.

[6] B. Heinz, "Backchannel responses as strategic responses in bilingual speakers' conversations," *Journal of pragmatics*, vol. 35, no. 7, pp. 1113–1142, 2003.

[7] S. Petridis, B. Martinez, and M. Pantic, "The mahnob laughter database," *Image and Vision Computing*, vol. 31, no. 2, pp. 186–202, 2013.

[8] G. McKeown, R. Cowie, W. Curran, W. Ruch, and E. Douglas-Cowie, "Ilhaire laughter database," in *Proceedings of 4th International Workshop on Corpora for Research on Emotion, Sentiment & Social Signals*, 2012, pp. 32–35.

[9] C. Ishi, H. Hatano, and N. Hagita, "Analysis of laughter events in real science classes by using multiple environment sensor data," in *Proc. of INTERSPEECH*, 2014, pp. 1043–1047.

[10] M. Koutsombogera and C. Vogel, "Understanding laughter in dialog," *Cognitive Computation*, vol. 14, no. 4, pp. 1405–1420, 2022.

[11] B. Ludusan, M. Schröer, M. Rossi, and P. Wagner, "The co-use of laughter and head gestures across speech styles," in *Proc. of INTERSPEECH*, 2023, pp. 3592–3596.

[12] M. Rychlowska, G. McKeown, I. Sneddon, and W. Curran, "The role of contextual information in classifying spontaneous social laughter," *Journal of Nonverbal Behavior*, vol. 46, no. 4, pp. 449–466, 2022.

[13] G. Bryant and A. Aktipis, "The animal nature of spontaneous human laughter," *Evolution and Human Behavior*, vol. 35, no. 4, pp. 327–335, 2014.

[14] N. Lavan, S. Scott, and C. McGettigan, "Laugh like you mean it: Authenticity modulates acoustic, physiological and perceptual properties of laughter," *Journal of Nonverbal Behavior*, vol. 40, pp. 133–149, 2016.

[15] V. Krepsz, V. Horváth, A. Huszár, T. Neuberger, and D. Gyarmathy, ""Should we laugh?" Acoustic features of (in)voluntary laughters in spontaneous conversations," *Cognitive Processing*, pp. 1–18, 2023.

[16] Z. Malisz, M. Wlodarczak, H. Buschmeier, J. Skubisz, S. Kopp, and P. Wagner, "The ALICO corpus: Analysing the active listener," *Language Resources and Evaluation*, vol. 50, no. 2, pp. 411–442, 2016.

[17] M. Gamer, J. Lemon, I. Fellows, and P. Singh, *irr: Various Coefficients of Interrater Reliability and Agreement*, 2019, r package version 0.84.1. [Online]. Available: https://CRAN.R-project.org/package=irr

[18] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://www.R-project.org/

[19] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "VoiceSauce: A program for voice analysis," in *Proc. of ICPhS*, 2011, pp. 1846–1849.

[20] S. Dupont, H. Çakmak, W. Curran, T. Dutoit, J. Hofmann, G. McKeown, O. Pietquin, T. Platt, W. Ruch, and J. Urbain, "Laughter research: a review of the ilhaire project," *Toward Robotic Socially Believable Behaving Systems-Volume I: Modeling Emotions*, pp. 147–181, 2016.

[21] B. Ludusan, P. Wagner, and M. Włodarczak, "Cue interaction in the perception of prosodic prominence: The role of voice quality," in *Proc. of INTERSPEECH*, 2021, pp. 1006–1010.