The role of linguistically encoded emotional characteristics for cooperativeness in a one-shot prisoner's dilemma

Andreas Baumann[*1], Theresa Matzinger[2,6], Roland Mühlenbernd[3], Slawomir Wacewicz[5], Michael Pleyer[5], Stefan Hartmann[4], and Marek Placiński[5]

[*]Corresponding Author: andreas.baumann@univie.ac.at

[1]Department of German Studies, University of Vienna, Vienna, Austria

[2]Department of English Studies, University of Vienna, Vienna, Austria

[3]Leibniz-Zentrum Allgemeine Sprachwissenschaft, Berlin, Germany

[4]German Department, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

[5]Center for Language Evolution Studies, Nicolaus Copernicus University in Toruń, Toruń, Poland

[6]Cognitive Science Hub, University of Vienna, Vienna, Austria

We investigate linguistically encoded emotional alignment in pairs of players in a TV game show that is set up as a one-shot prisoner's dilemma. We measure which linguistically encoded emotional characteristics are relevant for choosing between cooperative and defective behavior in that game. We show that cooperativeness depends on interactions between emotional characteristics of both players. In contrast to research on emotional synchrony and cooperation, however, we find that players are more likely to cooperate if their emotions do not align. We interpret this as an instance of deceptive linguistic behavior.

1. Introduction

The cooperative character of language is a key tenet in linguistics, and indeed sharing honest information technically qualifies as cooperation. This presents a well-known evolutionary problem, since, generally, cooperation with biologically unrelated individuals is not evolutionarily stable and can only evolve under very rare circumstances. This is why "the cooperative sharing of information [...] remains a central puzzle in language evolution" (Fitch, 2010: 417). Across the behavioral sciences, the special conditions that enable the emergence and stability of cooperation are typically modeled using the classic game-theoretic tool of the prisoner's dilemma (PD; Nowak & Sigmund, 1993). In this study, we use a PD-structured game show to determine which emotional characteristics may influence the decision to cooperate or to defect.

Previous research indicated that emotions can indeed play a role in maintaining cooperative behavior in PD. Chen et al. (2021) demonstrate that cooperation is promoted in the iterated PD if enough individuals display emotions in a non-competitive way. Similarly, de Melo and Terada (2020) study the effect of non-verbal emotional expression on decision making in the PD.

The alignment of emotions in linguistic interactions was shown to be indicative of cooperation (Arimoto & Okanoya, 2014), and more fundamentally, has been argued to be crucially relevant for the emergence of language in general (Tomasello, 2019). There is robust evidence for emotional alignment and synchrony in parent-child interactions (Lee et al., 2017, Leclère et al., 2014), and among partners (Randall et al., 2013), which are both highly cooperative social relationships. Connected to this, Shilton et al. (2020) argue that emotional synchrony and social bonding are associated and that both have been promoted by coordinated music-making in the social evolution of humans.

Given the close connection of emotion and cooperation, we would expect linguistically encoded emotional alignment to promote cooperativeness, i.e., the tendency to display cooperative behavior in the PD, if players in that game were allowed to communicate before making a decision. This is exactly the hypothesis that we examine in this study. We do so by analyzing linguistically expressed emotional behavior and cooperativeness in a text corpus.

2. Data and preparations

Our study is based on a corpus of 17 transcribed episodes of the TV show ‘Golden Balls’, a game show that has been the subject of various behavioral studies (e.g., Burton-Chewell & West, 2012). In each episode of this show, four players interact, two of which eventually engage in a final round that effectively represents a one-shot PD, i.e., a variation of the PD in which two individuals play only once. In this game, players can choose to ‘split’ (cooperate) or ‘steal’ (defect) the ‘jackpot’. The combination of the chosen strategies determines the final reward in line with payoffs in the PD.

In our analysis[1], we only considered utterances from players entering the final round. Since we are interested in emotional characteristics, each utterance was automatically annotated with numeric scores for the following emotional dimensions (Russel & Mehrabian, 1977): valence (V, negative—positive), arousal (A, calm—agitated), and dominance (D, submissive—dominant). We adopted a lexicon-based bag-of-words approach (Taboada et al., 2011) employing VAD norms from Warriner et al. (2013).

[1] Data and code available at https://gitlab.com/andreas.baumann/emo_coop_golden_balls

Next, a smooth time-series model (generalized additive model, Wood, 2017) was fit for each emotional dimension and each player in each episode, thereby describing the trajectory of that emotional property through the episode (Fig. 1, left). Multiple summary measures of the dynamics of VAD of both players were derived from these models: 'alignedness' (do the trajectories of both players match?), 'alignment' (do the trajectories converge/diverge?), 'own' VAD scores, and VAD scores of the 'other' player. All measures are listed in Fig. 1 (right).

More specifically, 'alignedness' is determined by measuring, for each emotional dimension, dynamic time-warping distance between the trajectories of both players.[2] Low distance, i.e., a high similarity between trajectories, corresponds to high alignedness of both players with respect to that emotional dimension. Measuring emotional 'alignment' involves two steps. First, pairwise distances between points on the trajectories for all time-steps (i.e., utterances) in the conversation[3] are computed, i.e., yielding a sequence of distances. Second, a linear regression model is computed in which this distance depends on time. The slope of this model is used for measuring alignment.[4] If the measure is positive, the trajectories of both players start being distant from each other and converge to become more similar in the course of the conversation. If it is negative, the trajectories diverge. In this way, we can differentiate between effects from aligning emotions through the whole conversation and effects of being emotionally synchronized right from the start.

Finally, for each player and each emotional dimension, the 'own' value is computed as the average across all scores in the trajectory of that player. The 'other' measure is computed, *mutatis mutandis*, by taking the average of all scores of the other player.

3. Importance of emotional features for cooperativeness

To check which measure is most important for predicting player behavior, a linear support vector machine (SVM) with 'split/steal' as binary outcome variable was trained and optimized using 5-fold cross-validation. Area under the ROC curve (AUC) was used as a measure of variable importance (Fig 1, right). The model displays an above-chance, albeit not particularly high, accuracy of

---

[2] Dynamic time-warping was chosen to account for potentially shifted emotional reactions in the (pairwise) sequence of utterances.

[3] Note that this is possible since the time-series models interpolate emotion scores so that these models yield predictions for each utterance-step and each player.

[4] Formally, for a linear model of pairwise distance d depending on time t, $d(t) = bt + c + \epsilon$, we define alignment as $-b$. Positive alignment corresponds to convergence, negative alignment to divergence.

0.71 (chance being 0.5).[5] More interestingly, the analysis shows, first, that emotional interactions and the emotions of the other player are considerably more important for behavioral decisions than this is the case for one's own emotions. This is evident since measures of a player's 'own' emotions (valence_own, arousal_own, dominance_own) display low importance. Measures that relate emotions of both players to each other rank higher, on average. Second, we find valence alignedness as well as dominance alignment and alignedness seem to be most important with an AUC score above 0.70, while all other measures are less important (Fig 1, right).

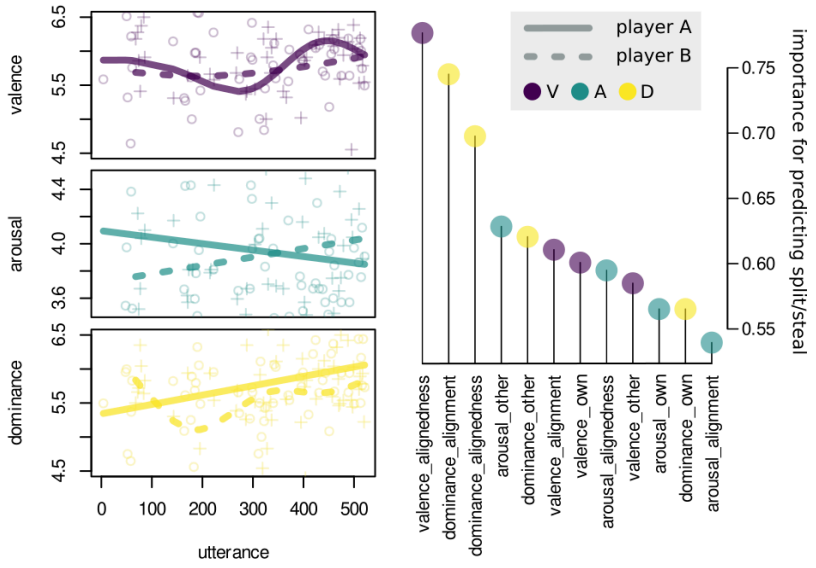

Figure 1. Left: smooth models (GAMs) for emotional developments in one 'Golden Balls' episode. Right: variable importance (ROC AUC) in a SVM based on all episodes.

4. Emotional similarity and cooperativeness

In a second analysis, we tested in more detail how exactly alignment and alignedness influenced cooperation. For each of the three most important predictors of cooperativeness (Fig. 1, right), valence_alignedness, dominance_alignment, dominance_alignedness, we fit a Bayesian Bernoulli model with 'split/steal' as binary outcome variable ('split' being treated as 'success'). We used a logit-link and flat (uninformative) priors for the linear

[5] The goal, in the first place, was not to train a model that predicts cooperativeness at a high accuracy, but to gain insights into which (type of) emotional features of a conversation are most relevant for predicting the outcome in an exploratory way. The above chance accuracy at least indicates that the cooperativeness *can* be inferred from emotional characteristics, albeit not reliably.

coefficients. Predictor variables were scaled with respect to their mean and standard deviation before entering the models.

In all models, an effect of emotional alignment/alignedness on the outcome is visible (Fig. 2). However, contrary to our expectations, it is weak rather than strong alignment that promotes one's propensity to cooperate. The respective model coefficients (i.e., effects on the logit) and 95% credible intervals read: -1.27 (-2.36, -0.36) for valence_alignedness, -0.92 (-1.95, -0.07) for dominance_alignedness, and -1.70 (-3.45, -0.34) for dominance_alignment.

What we also see in all models is that low alignment/alignedness yields a chance to split of almost 1.00, while high alignment/alignedness corresponds to a chance to split of around 0.25. Emotional distance seems to be connected to cooperative behavior, while emotional similarity may still entail cooperative behavior at a non-negligible probability.

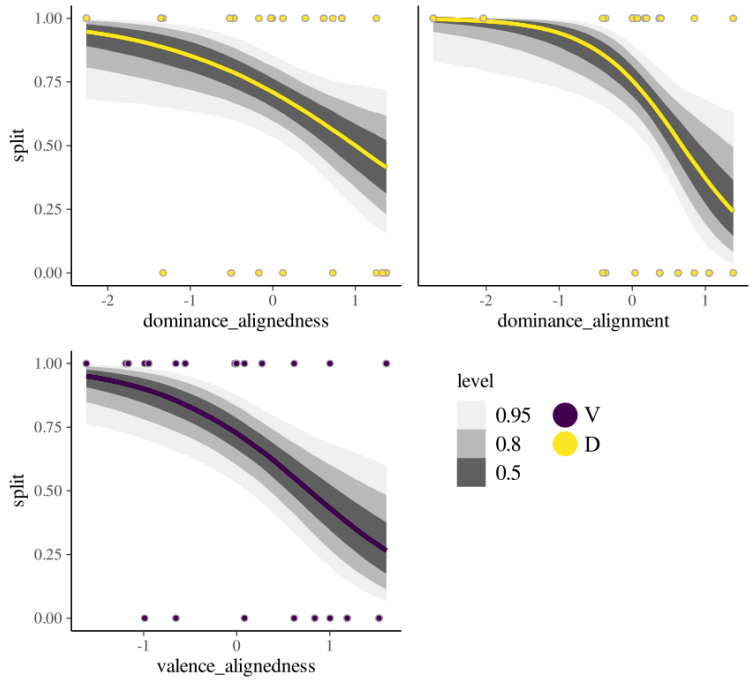

Figure 2. Bernoulli models of cooperativeness (split/steal) depending on the most important variables (cf. Fig. 1, right): dominance_alignedness, dominance_alignment, and valence_alignedness. Bayesian probability bands shown in gray. Emotional closeness generally *decreases* the chance to split.

5. Discussion and conclusion

We have shown that emotional dynamics and interactions among players (rather than just one's own emotions) indeed have an impact on cooperativeness in the

PD, but not in the way that we had expected based on extant research on emotion and cooperation. We found that a player is more likely to cooperate if their counterpart displays *divergent* emotional behavior. Put differently, players are inclined to defect if their emotions are aligned with that of the other player.

This somewhat unexpected outcome could, of course, result from the nature of the data that we inspected. For one, the number of episodes in our sample (17) is relatively small. Although we detect statistically robust effects, it is naturally possible that some of the effects change if more episodes are taken into account. In addition, and more fundamentally, we only assessed emotional expression on the lexical level, thereby ignoring phonetic and prosodic cues, let alone visual information (in particular, gestures or facial expressions; Lei & Gracht, 2019). Finally, the result could be grounded in the artificial setup of the TV show and a potential bias towards competitively minded personalities participating in shows like 'Golden Balls'.

However, leaving the possibility of methodological shortcomings aside, our results could be potentially revealing, as they let us conjecture that linguistically encoded emotion can serve the purpose of deception in competitive situations, thereby also overriding benevolent effects of emotional signaling. That is, emotional alignment could be exploited to mislead a competitor in order to maximize one's own reward. Whether or not this is done consciously cannot be easily assessed based on the examined data.

Interestingly, results from research on emotional mimicry offer an alternative explanation. It was shown that facial mimicry of negative emotions is promoted if one's counterpart has the reputation of behaving in an unfair manner (Hofmann et al., 2012; mimicry of positive emotions was not shown to be modulated by fairness, however). Thus, it could be that players that acquire the reputation of being unfair in the first two rounds of the game and who are expected to defect, elicit (negative) emotional alignment in their counterpart.

In both cases, dishonesty and deception are key aspects. This is in line with the work by Robson (1990) and Santos, Pacheco and Skyrms (2011), who show through evolutionary analyses of the PD with pre-play signaling that signals that are introduced to promote mutual cooperation can easily be exploited towards defection. Moreover, linguistic dishonesty in 'Golden Balls' was already examined in Burton-Chellew and West's (2012) analysis. They found that exaggerating players demoted cooperativeness in their counterpart. Thus, we consider honesty and emotional dynamics in language, and how they impact cooperative behavior to be an interesting interaction worthy of being examined more closely in the light of language evolution research.

Acknowledgements

We would like to thank Jon Carr and an anonymous reviewer for a series of helpful comments on an earlier version of this manuscript. Michael Pleyer was supported by project No. 2021/43/P/HS2/02729 co-funded by the National Science Centre and the European Union Framework Programme for Research and Innovation Horizon 2020 under the Marie Skłodowska-Curie grant agreement No. 945339.

References

Arimoto, Y., & Okanoya, K. (2014). Emotional synchrony and covariation of behavioral/physiological reactions between interlocutors. *17th Oriental Chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA)* (pp. 1-6). IEEE.

Burton-Chellew, M. N., & West, S. A. (2012). Correlates of cooperation in a one-shot high-stakes televised prisoners' dilemma. *PloS one*, 7(4), e33344.

Chen, W., Wang, J., Yu, F., He, J., Xu, W., & Wang, R. (2021). Effects of emotion on the evolution of cooperation in a spatial prisoner's dilemma game. *Applied Mathematics and Computation*, 411, 126497.

Fitch, W.T., 2010. *The Evolution of Language*. Oxford University Press, Oxford.

Hofman, D., Bos, P. A., Schutter, D. J., & van Honk, J. (2012). Fairness modulates non-conscious facial mimicry in women. *Proceedings of the Royal Society B: Biological Sciences*, 279(1742), 3535-3539.

Leclère, C., Viaux, S., Avril, M., Achard, C., Chetouani, M., Missonnier, S., & Cohen, D. (2014). Why synchrony matters during mother-child interactions: a systematic review. *PloS one*, 9(12), e113571.

Lee, T. H., Miernicki, M. E., & Telzer, E. H. (2017). Families that fire together smile together: Resting state connectome similarity and daily emotional synchrony in parent-child dyads. *NeuroImage*, 152, 31-37.

Lei, S., & Gratch, J. (2019). Smiles signal surprise in a social dilemma. *8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 627-633). IEEE.

de Melo, C. M., & Terada, K. (2020). The interplay of emotion expressions and strategy in promoting cooperation in the iterated prisoner's dilemma. *Scientific reports*, 10(1), 14959.

Nowak, M., & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature*, 364(6432), 56-58.

Randall, A. K., Post, J. H., Reed, R. G., & Butler, E. A. (2013). Cooperating with your romantic partner: Associations with interpersonal emotion coordination. *Journal of Social and Personal Relationships*, 30(8), 1072-1095.

Robson, A. J. (1990). Efficiency in evolutionary games: Darwin, Nash and the secret handshake. Journal of Theoretical Biology, 144, 379–396.
Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. Journal of research in Personality, 11(3), 273-294.
Santos, F.C., Pacheco, J.M., & Skyrms, B. (2011). Co-evolution of pre-play signaling and cooperation. Journal of Theoretical Biology 274, 30–35.
Shilton, D., Breski, M., Dor, D., & Jablonka, E. (2020). Human social evolution: self-domestication or self-control?. *Frontiers in Psychology*, 134.
Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2), 267-307.
Tomasello, M. (2019). *Becoming human: A theory of ontogeny*. Harvard University Press.
Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, 45, 1191-1207.
Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.