Research Traineeships 2021 proposal format

1. Title of the project

"They are racists, but they are not bad."

How loaded language combines evaluation and description

2. Coordinators

Dr. Giovanni Cassani, CSAI

Dr. Matteo Colombo, DFI

3. Project summary

If I call you a racist, I mean to describe one of your traits and also evaluate it as negative. Terms like *racist* (but also *poor*, *rude*, *courageous*, *gullible, …*) express what philosophers call *thick concepts*[1], which somehow combine evaluation and non-evaluative description and stand in contrast to purely evaluative terms like *good* or *bad*, and non-evaluative descriptive terms, like *Italian* or *spherical*.

When explaining how thick concepts work, philosophers typically rely on their own intuitions, taken to be linguistic data representative of how ordinary language works. Yet, reliance on personal linguistic intuition has contributed to foster disagreement about whether thick terms convey an evaluation as a matter of *semantics* -- i.e. as a matter of the conditions that must hold for utterances involving thick terms to express true propositions -- or whether they do so through *pragmatic* mechanisms, such as linguistic conventions, conversational norms or contextual implicatures[2]. Here, we aim to clarify this disagreement by combining ideas and methods from philosophy, experimental pragmatics and computational linguistics.

Specifically, during online experiments, native English participants will be presented with sentences such as

John is racist, and he is a good man (1)

where a thick term (*racist*) is combined with an evaluative or descriptive adjective (e.g., *good, Dutch*) by either a coordinating (*and*) or an adversative conjunction (*but*). We will manipulate the polarity of the thick term and the evaluative adjective, and test their effect on how acceptable participants find these sentences[3]. We will collect an implicit measure of acceptability through self-paced reading[4] (unacceptable combinations slow readers down) and explicit acceptability measurements by asking participants the extent to which the sentence sounds contradictory[3]. If (1) is rated as very unacceptable, that would be evidence for a semantic approach to thick terms.

Moreover, we will rely on state-of-the-art language models trained on large corpora[5,6,7,8,9] to (i) derive measures of how predictable the second adjective is and assess whether they predict acceptability ratings; and (ii) extrapolate the predictability of adjectives with opposite polarity to check whether the first thick term in combination with the conjunction triggers coherent semantic expectations for positive or negative terms[10].

In conclusion, our project helps advance current understanding of how thick concepts combine evaluation and description, probes possible biases present in widely used language models[11,12,13], and demonstrates how methods from different fields can fruitfully be integrated to pursue research at the intersection of philosophy, psycholinguistics and the digital humanities.

References

1. Williams, B. (1985). *Ethics and the Limits of Philosophy*, Cambridge, MA: Harvard University Press.

2. Väyrynen, P. (2021). Thick Ethical Concepts. In Ed. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition), URL https://plato.stanford.edu/archives/spr2021/entries/thick-ethical-concepts

3. Willemsen, P. & Reuter, K. (2020). Separability and the Effect of Valence. *Proceedings of the 42th Annual Conference of the Cognitive Science Society,* Austin, TX: Cognitive Science Society, pp. 794–800.

4. Jegerski, J. (2013). Self-paced reading. In *Research methods in second language psycholinguistics* (pp. 36-65). Routledge.

5. Divjak, D. (2017). The role of lexical frequency in the acceptability of syntactic variants: Evidence from that-clauses in Polish. *Cognitive science*, *41*(2), 354-382.

6. Hao, Y., Mendelsohn, S., Sterneck, R., Martinez, R., & Frank, R. (2020). Probabilistic Predictions of People Perusing: Evaluating Metrics of Language Model Performance for Psycholinguistic Modeling. *arXiv preprint arXiv:2009.03954*.

7. Merkx, D., & Frank, S. L. (2020). Comparing Transformers and RNNs on predicting human sentence processing data. *arXiv preprint arXiv:2005.09471*.

8. Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302-319.

9. Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior. *arXiv preprint arXiv:2006.01912*.

10. Jumelet, J., Zuidema, W., & Hupkes, D. (2019). Analysing neural language models: Contextual decomposition reveals default reasoning in number and gender assignment. *arXiv preprint arXiv:1909.08975*.

11. Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv preprint arXiv:1607.06520*.

12. Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183-186.

13. Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, *8*, 34-48.

4. Project timeline

Brief work plan including timeline.

During the summer 2021, the two coordinators will start to define experimental design and analysis protocol for the envisaged studies. We shall distribute an open call for trainees in our project in late August 2021. With the trainee, we shall settle on a final design and analysis plan in the first few weeks of the traineeship in September. In October, with the approval of the Ethics board at TSHD, we shall collect behavioural data in an online survey. Between November and December, we shall focus on the analysis of behavioral data using statistical techniques (including, but not limited to, mixed effect models), on the extraction of computationally-derived measures of word predictability, using state-of-the-art language models from Natural Language Processing, and on the writing of a research article, where we report our findings in the context of the existing literatures on thick concepts in philosophy and linguistics.

5. Research trainee profile

Trainees should ideally be enrolled in one of the following programs:

- Master in Data Science and Society
- Research Master in Linguistics and Communication
- Bachelor in Cognitive Science and Artificial Intelligence (the student passed the Computational Linguistics course)

Trainees should have:

- knowledge of experimental designs
- a foundational knowledge of statistical methods (regression, ANOVA),
- some knowledge of computational linguistics/NLP
- basic knowledge of machine learning
- have some experience coding in Python or R
- good writing, communication and analytical skills

Knowledge of Philosophy of Language or Meta-ethics are a plus.

If you match most but not all of these requirements, please do submit your candidacy nonetheless! Willingness to learn can outweigh almost any lack of technical skills.

All in all, this project will help a trainee to acquire work experience in an interdisciplinary academic setting, to develop analytical and quantitative skills, to gain new knowledge and understanding of the philosophical literature on thick concepts and suitable methods to advance this literature. The trainee will gain further academic skills in giving presentations and preparing a paper for publication, as we plan to present our findings at relevant research conferences/symposia and to submit them as a research article.