The Sentence Gestalt model: Modelling N400 amplitudes as change in a probabilistic representation of meaning

Milena Rabovsky (University of Potsdam)
milena.rabovsky@uni-potsdam.de

Language ultimately aims to convey meaning. The N400 brain potential has aroused much interest as it is thought to provide an online measure of meaning processing in the brain, and has been used in a large number of empirical studies (many more than 1000). However, despite this richness of data, the cognitive processes underlying N400 amplitudes are still unclear and actively debated. In the talk, I will present a computationally explicit account of these processes and the emerging representation of sentence meaning. We simulate N400 amplitudes as the change in a probabilistic representation of sentence meaning captured in the hidden unit activation pattern in a neural network model of sentence comprehension, the Sentence Gestalt (SG) model. The change of the representation corresponds to an implicit prediction error at the level of meaning, which can serve as the error signal driving connection weight changes in the network. This suggests that N400 amplitudes reflect an implicit prediction error functioning as a learning signal driving adaptation during language comprehension. The model accounts for 17 distinct and diverse established empirical effects from the N400 literature. Its ability to account for such a wide range of empirical findings suggests that the principles of representation and processing it embodies may capture some essential aspect of human language comprehension. The model does not exactly correspond to any existing account of the N400 (e.g., retrieval or integration) as these accounts often rely (at least partly) on concepts from classic linguistic theories while the SG model implements a distinct perspective on language comprehension, resembling current deep learning language models. Specifically, the model suggests that language comprehension may not be based on pre-defined symbols and explicit rules but rather that the human process of learning to understand language may consist in learning continuous nonlinear functions mapping from linguistic input to meaning.