Modeling influences of coercion on N400 amplitudes as change in a probabilistic representation of meaning

Milena Rabovsky (University of Potsdam, Germany)

Coercion has been defined as 'a semantic operation that converts an argument to the type that is expected by a function, where it would otherwise result in a type error' [1, p. 425]. An example of complement coercion is given by the sentence 'The journalist began the article' where the predicate 'began' would require its complement to denote an event, but 'the article' instead denotes an entity. Thus, 'began' coerces 'the article' from an entity to an event involving this entity, allowing for the interpretation 'The journalist began writing the article'. Influences of complement coercion on event related brain potentials (ERPs) have been investigated by presenting sentences such as 'The journalist began/ wrote/ accomplished the article' (i.e. 'coerced'/ 'non-coerced'/ incongruent) and comparing ERPs at the noun [2]. The authors observed larger N400s for 'coerced' and incongruent as compared to 'non-coerced' sentences. The goal of the current study was to investigate whether these observed influences of coercion on N400 amplitudes can be accounted for by the Sentence Gestalt (SG) model, a neural network model of sentence comprehension [3] that has previously been used to account for a broad range of N400 effects (Fig. 1; [4,5,6]).

The training environment of the SG model, which is based on a simple generative model (see [4] for details), was extended to include coercion like situations. Specifically, two additional verbs were included in the model's vocabulary ('begin' and 'finish'), which could be combined with all other verbs such as e.g., in 'The man began/ finished reading the novel/ planting the rose/…'. For some sentences, such as the example sentence with the novel, complement coercion is possible, and the gerund was sometimes (with .2 probability) omitted. Ten independently initialized models were each trained on 800.000 example sentences produced by the simple generative model. For the simulation experiments, the ten trained models were each presented with 8 triplets of stimuli designed to mimic the ERP study reported above, e.g., 'The man began/ read/ ate the novel' (i.e., 'coerced'/ 'non-coerced'/ incongruent). The model's N400 correlate, which is the magnitude of change in the model's hidden SG layer induced by the current word (i.e., *Model N400* = $|SG_t - SG_{t-1}|$), was compared at the noun.

The model's N400 correlate was larger for 'coerced' and incongruent as compared to 'non-coerced' sentences over models and items (*t*s > 6.6, *p*s < .001; see Fig. 2), in line with the empirical data [2]. Note that this is the case despite the fact that the SG model does not assume a specific process such as 'coercion' to explain the interpretation of these sentences. Because the model does not assume fixed rules, no operation is required to prevent a presumed rule violation such as a type error. Instead, the model constantly estimates the probabilities of all relevant aspects of meaning involved in the described event based on the statistical structure of its environment, including aspects that are not explicitly mentioned in the sentence. It does not contain fixed lexical representations of words that would need to be converted into something else. Instead, each incoming word provides cues constraining the overall interpretation of the sentence. The model's N400 correlate for sentences containing 'coercion' thus does not reflect any specific 'coercion' process of converting an argument into another type, but rather reflects the same process assumed to underlie N400 amplitudes in general from the model's perspective, namely the amount of change in expected sentence meaning induced by the critical word. The amount of change was larger for 'coerced' as compared to 'non-coerced' sentences because the 'coerced' sentences were of lower constraint and lower cloze probability as was the case in the empirical study [2] (see also [6] and [7]). Thus, from this perspective, the available evidence reporting effects of complement coercion on ERPs (see also [8]) does not speak to the neurocognitive reality of this construct from compositional semantics.

References

[1] Pustejovsky, J. (1995). *The generative lexicon*. Cambridge, MA: MIT Press.
[2] Kuperberg, G.R., Choi, A., Cohn, N., Paczynski, M., & Jackendoff, R. (2010). Electrophysiological correlates of complement coercion. *Journal of Cognitive Neuroscience, 22,* 2685–2701.
[3] St. John, M.F., McClelland, J.L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46,* 217–257.
[4] Rabovsky, M., Hansen, S.S., & McClelland, J.L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour, 2,* 693-705.
[5] Rabovsky, M. (2020). Change in a probabilistic representation of meaning can account for N400 effects on articles: A neural network model. *Neuropsychologia, 143,* 107466.
[6] Rabovsky, M., & McClelland, J. L. (2020). Quasi-compositional mapping from form to meaning: a neural-network based approach to capturing neural responses during language comprehension. *Philosophical Transactions of the Royal Society B*. 375: 20190313.
[7] Delogu, F., Crocker, M.W., & Drenhaus, H. (2017). Teasing apart coercion and surprisal: evidence from eye-movements and ERPs. *Cognition,161,* 46–59.
[8] Baggio, G., Choma, T., van Lambalgen, M., & Hagoort, P. (2009). Coercion and compositionality. Journal of Cognitive Neuroscience, 22, 2131-2140.

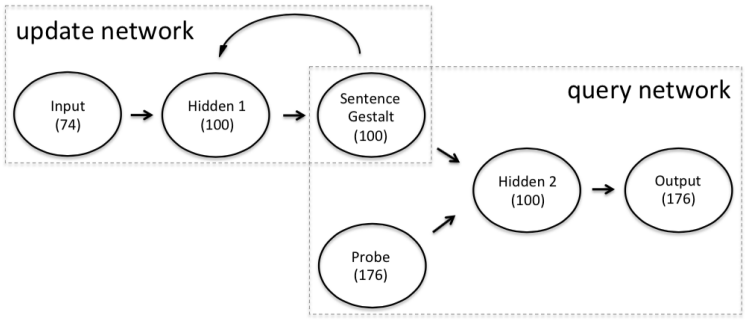

Figure 1. The Sentence Gestalt (SG) model architecture. Arrows represent all-to-all modifiable connections and ovals represent layers of units (with numbers of units in parentheses).

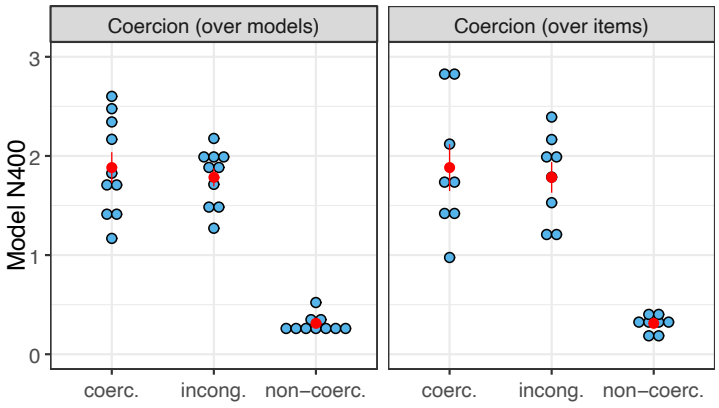

Figure 2. Influences of coercion on the SG model's N400 correlate. Blue dots represent results for models (n = 10, left) and items (n = 8, right); red dots represent condition means +/- standard error of the mean (SEM).