Contrafactives and learnability

Holton (2017) has drawn attention to a novel semantic universal, according to which (at least almost) no natural language features contrafactive attitude verbs. Contrafactives are the mirror image of factive attitude verbs like *know*, *remember*, *see*, and *regret*: roughly, they are attitude verbs that entail and presuppose the falsity of their *that*-complements. Although some candidate contrafactives have been discussed (see Holton, 2017, pp.245-9, 262–4), no clear counterexample to the universal has been found: for instance, Anvari, Maldonado, and Soria Ruiz (2019)'s *creerse* is built by adjoining the reflexive pronoun to the non-factive verb *creer* 'believe', and Glass (2020)'s Mandarin belief verb *yǐwéi* carries a post-, rather than presupposition that the reported belief must not be added to the common ground. The no contrafactives universal raises an important question: why do natural languages universally feature factive verbs like *know* (Goddard, 2010), but (at least almost) universally lack contrafactives? We develop a novel explanation of this asymmetry. Drawing on recent discussions of other semantic universals, e.g. the veridical uniformity universal for responsive verbs (Steinert-Threlkeld, 2019), we explore the hypothesis that the asymmetry between contrafactives and factives arises (at least in part) because the meaning of a contrafactive is harder to learn than that of a factive. We will test this hypothesis by conducting a computational experiment using an artificial neural network.

Our hypothesis is inspired by the intuitive idea that languages have words for meanings that are easier to learn and use compositional methods to express meanings that are harder to acquire (Steinert-Threlkeld and Szymanik, 2019, p.4). Using a simple Hintikka semantics for attitude verbs, this intuitive idea can be applied to the asymmetry between factives and contrafactives. Compare 1 and 2

1. $[\![\text{know}]\!]^w = \lambda p \lambda x.\underline{p(w) = 1}.\forall w'[w' \in \text{bel}_{\text{x,w}} \rightarrow p(w') = 1]$ where $\text{bel}_{\text{x,w}} = \{w' |$ it's compatible with what x believes in w that x is in $w'\}$ and $w \in \text{bel}_{\text{x,w}}$
2. $[\![\text{contrafactive}]\!]^w = \lambda p \lambda x.\underline{p(w) = 0}.p(w) = 0 \wedge \forall w'[w' \in \text{bel}_{\text{x,w}} \rightarrow p(w') = 1]$ where $\text{bel}_{\text{x,w}} = \{w' |$ it's compatible with what x believes in w that x is in $w'\}$

1 associates the factive *know* with a set of possible worlds ($\text{bel}_{\text{x,w}}$) throughout which p is true. The stipulation that this set includes the world of evaluation w then guarantees that *know* entails the truth of its *that*-complement. (The underlined clause further contributes the presupposition that its *that*-complement is true.) By contrast, the denotation of a contrafactive in a similar model in 2 is more complex, leading to the (defeasible) expectation that it is harder to acquire (see Pol, Steinert-Threlkeld, and Szymanik, 2019, on the relation between complexity and learnability). Since this denotation must not merely be neutral on whether the contrafactive's *that*-complement is true, but guarantee that it is false, a stipulation about what worlds $\text{bel}_{\text{x,w}}$ includes (or does not include) will not suffice. An additional stipulation of the falsity of p is needed.

To test our expectation that the denotation of a contrafactive is harder to acquire, we will conduct a computational experiment using an artificial neural network. This network will be trained to predict the truth value of factive, non-factive or contrafactive attitude ascriptions, given an accurate representation of a small world and a representation of the small world as the attitude holder takes it to be (which may or may not be accurate). The artificial language in which the target attitude ascriptions are formulated and which the neural network will learn can be interpreted as a fragment that describes propositions about the relative locations of two objects to each other plus the attitude taken towards these propositions. To encode this artificial language and the small world representations, Transformer encoders will be used.

Our computational experiment will improve on similar ones conducted by Steinert-Threlkeld (2019) and Steinert-Threlkeld and Szymanik (2019) in a number of ways. First, we will report results from a larger range of hyperparameters (e.g. training epochs, learning rate, etc.). By exploring the range of models which bear out our expectation that contrafactives are harder to learn than factives, we will provide a better sense of the robustness of our experimental results. Second, while the cited research used feed-forward neural networks and LSTMs, we will switch to the more advanced Transformer-approach. Recent results suggest that despite not being originally designed for cognitive plausibility, Transformer-based networks nonetheless show greater convergence with human processing than other approaches (e.g. Caucheteux and King, 2022). Given this, the results of our computational experiment likely reflect learnability for human language learners more closely than previous work. Third, we will separate the encoding of the artificial language and the world model by using two Transformer-encoders. Effectively, our architecture reflects the difference between the artificial object language that our neural network will learn and the meta-language that describes both the accurate representation of the target small world and the representation of the attitude holder. Modelling comprehension of the world and comprehension of the artificial language as separate processes will further increase the cognitive plausibility of our model.

References

Anvari, A., M. Maldonado, and A. Soria Ruiz (2019). "The puzzle of Reflexive Belief Construction in Spanish". In: *Proceedings of Sinn und Bedeutung* 23.1, pp. 57–74. doi: 10.18148/sub/2019.v23i1.503.

Caucheteux, C. and J.-R. King (16, 2022). "Brains and Algorithms Partially Converge in Natural Language Processing". In: *Communications Biology* 5.1 (1), pp. 1–10. doi: 10.1038/s42003-022-03036-1.

Glass, L. (2020). "The negatively biased Mandarin belief verb "yiwei"".

Goddard, C. (2010). "Universals and Variation in the Lexicon of Mental State Concepts". In: *Words and the Mind: How words capture human experience*. Oxford: Oxford University Press.

Holton, R. (2017). "I—Facts, Factives, and Contrafactives". In: *Aristotelian Society Supplementary Volume* 91.1, pp. 245–266. doi: 10.1093/arisup/akx003.

Pol, I. v. d., S. Steinert-Threlkeld, and J. Szymanik (2019). *Complexity and learnability in the explanation of semantic universals of quantifiers*. Tech. rep. PsyArXiv. doi: 10.31234/osf.io/f8dbp.

Steinert-Threlkeld, S. (2019). "An Explanation of the Veridical Uniformity Universal". In: *Journal of Semantics*. doi: 10.1093/jos/ffz019.

Steinert-Threlkeld, S. and J. Szymanik (2019). "Learnability and semantic universals". In: *Semantics and Pragmatics* 12.0, p. 4. doi: 10.3765/sp.12.4.