Contrafactives, learnability, and production

Abstract for ELM3

Natural languages appear to universally feature factive verbs like *know* (Goddard, 2010), whereas no clear example of a so-called contrafactive has been found yet (see, e.g., Holton, 2017; Glass, 2023; Roberts and Özyildiz, 2023). A contrafactive is the mirror image of a factive attitude verb like *know*. Although both *x factives that p* and *x contrafactives that p* entail that x believes that p, the former presupposes that p is true, whilst the latter presupposes that p is false.

Strohmaier and Wimmer (2022; 2023) have proposed that the stark difference in how common factives and contrafactives are arises partly because the meaning of a contrafactive is harder to learn than that of a factive. They tested this hypothesis by conducting two computational experiments using artificial neural networks—more specifically, Transformers, which are the foundation of current state-of-the-art results in natural language processing and show greater convergence with human processing than other approaches (Vaswani et al., 2017; Caucheteux and King, 2022). Their networks were trained to predict the truth value of factive, non-factive and contrafactive ascriptions, given a representation of the state of the world and a representation of the world as the attitude holder takes it to be (which may or may not be accurate). The networks' predictions were then expressed in a probability that the target ascription is true. Importantly, Strohmaier and Wimmer's experiments provide initial support for their hypothesis: in both cases the networks' loss drops faster for factives than for contrafactives.

However, their experiments are subject to at least two limitations. First, they understand an assignment of probability 0 to an ascription as claiming that the ascription is definitely not true, which leaves open whether the ascription is false or undefined due to presupposition failure. Thus, their experiments are not sensitive to a key feature of factives and contrafactives: their presuppositions. Second, their experiments effectively consider the comprehension (or evaluation) of attitude ascriptions fed into the networks. Their results therefore do not speak to the relative difficulty of learning how to produce factive and contrafactive ascriptions. This leaves open the possibility that it is easier to learn how to produce contrafactive ascriptions than factive ones. But if this possibility was to be realized, would the meaning of a contrafactive be harder to learn than that of a factive overall? This would depend on the difficult question of how to weigh the learnability of production and comprehension in assessing the overall learnability of the target expressions.

To address the two limitations facing Strohmaier and Wimmer, we conducted a computational experiment, using another Transformer, in which our network produces factive, non-factive or contrafactive ascriptions, given a representation of the world as the attitude holder takes it to be, information or a lack thereof about whether this representation is correct, plus a demand that the network produces an ascription with a certain truth-value (true, false, or presupposition failure) and, in doing so, not only uses an embedded clause with a certain truth-value (true, false, unknown), but also produces the most informative ascription possible (thereby satisfying Grice's maxim of quantity and Heim's maximize presupposition). Because similar, learnability-focused work on other semantic universals (e.g., Steinert-Threlkeld, 2020) has also been limited to testing comprehension, our approach serves as proof of concept for a new experimental paradigm that can be used in learnability-based explanations of semantic universals.

To illustrate the input and output of our network, let's say we provide it with the attitude holder representation 'buy lorelai tomato chili stew dinner now', information that this representation is incorrect, and demand that it produces a true ascription with a false embedded clause. Given these inputs, the network is trained to produce a contrafactive ascription. Again, say we provide the network with the attitude holder representation 'cook lorelai mushroom pepper rice lunch now',

information that this representation is correct, and demand that it produces a true ascription with a true embedded clause. Now, the network is trained to produce a factive ascription.

Our Transformer models closely follow the paper by Vaswani et al. (2017) using the pyTorch implementation. The main difference is that output for each position in the sentence is constrained to the words allowed in that position using position-specific linear layers, i.e. our model capitalises on the fixed word order of the artificial language. This minimizes the role of syntax, which is of benefit since we are primarily interest in lexical semantics (and a limited number of pragmatic principles). We used a custom loss that encoded the semantic-pragmatic success conditions.

We explored 41 different hyperparameter settings using a randomised search and 5-fold cross-validation. In all but two of those settings, the model failed to learn the semantics of the target expressions. We evaluated the two successful settings on the held out test-data and in addition varied the original random seed for each of them four times, leading to 10 evaluations overall. The varying of the random seed allows us to test whether the results are robust to a random change in the initial conditions of the neural network.

While we do find small differences in the speed in which attitude verbs are learned by the model, these are not robust to changes in the random seeds. Furthermore, insofar as any trends are discernible, contrafactives appear to be acquired faster than factives.

Our empirical contributions thus are:

1. Transformer models can produce factive, non-factive, and contrafactive ascriptions, learning both semantic conditions and pragmatic principles.
2. Variation of random initialisation can affect the learning differences between attitude verbs, contrary to the hypothesis that contrafactives are consistently harder to learn than factives.

These results stand in clear contrast to those previously found by Strohmaier and Wimmer (2022; 2023) and underline the importance of considering production in modelling lexical acquisition.

References

Caucheteux, C. and J.-R. King (2022). “Brains and algorithms partially converge in natural language processing”. In: *Communications Biology* 5.1, p. 134. DOI: 10.1038/s42003-022-03036-1.

Glass, L. (2023). “THE NEGATIVELY BIASED MANDARIN BELIEF VERB yǐwéi*”. In: *Studia Linguistica* 77.1, pp. 1–46. DOI: 10.1111/stul.12202.

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.

Roberts, T. and D. Özyildiz (2023). “Bad attitudes: Impossible meanings and the false belief gap”.

Steinert-Threlkeld, S. (2020). “An Explanation of the Veridical Uniformity Universal”. In: *Journal of Semantics* 37.1, pp. 129–144. DOI: 10.1093/jos/ffz019.

Strohmaier, D. and S. Wimmer (2022). “Contrafactives and Learnability”. In: *Proceedings of the 23rd Amsterdam Colloquium*. Ed. by M. Degano et al. Amsterdam, pp. 298–305.

— (2023). “Contrafactives and Learnability: An Experiment with Propositional Constants”. In: *Logic and Engineering of Natural Language Semantics*. Ed. by D. Bekki, K. Mineshima, and E. McCready. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, pp. 67–82. DOI: 10.1007/978-3-031-43977-3_5.

Vaswani, A. et al. (2017). “Attention is All you Need”. In: *31st Conference on Neural Information Processing Systems*, pp. 1–11.