Contrafactives, learnability, and production

David Strohmaier & Simon Wimmer*

Abstract. No natural language has contrafactive attitude verbs. Because factives are universal across natural languages, this means that there is a major asymmetry between contrafactives and factives. We previously hypothesised that this asymmetry arises partly because the meaning of contrafactives is significantly harder to learn than that of factives. Here we test this hypothesis by using a production-oriented computational experiment that overcomes two limitations of our previous experiments. We find that our results do not support our previous hypothesis.

Keywords. Contrafactives; Factives; Semantic universals; Transformers; Learnability

1. Introduction. No natural language appears to have what Holton (2017) calls a 'contrafactive', i.e. a morphologically atomic attitude verb that would mirror *know* and other factives.[1,2] Like *know*, a contrafactive would entail a belief in the content of its declarative complement. But unlike *know*, it would presuppose the falsity (not truth) of that complement.[3] For instance, if English had a contrafactive *contra*, *Ayesha contras that Beatrice is cool* would entail that Ayesha believes that Beatrice is cool, but presuppose that it is false that Beatrice is cool.

That no natural language appears to have a contrafactive is one part of an asymmetry between contrafactives and factives.[4] The second part is that *know*, and so at least one factive, appears to have counterparts in all natural languages (Goddard 2010, Hannon 2015). Given that a contrafactive would mirror a factive, this two-part asymmetry raises the question: why do contrafactives and factives differ so greatly in their frequency?

There have been several attempts to explain this asymmetry. Holton (2017) suggests that contrafactives are ruled out because there are no entities suitable to serve as the semantic values of their declarative complements.[5] Roberts & Özyildiz (2023) suggest that they are ruled out because they do not satisfy a necessary condition on presupposition triggering: that asserted content depends on

*This paper reports on research supported by Cambridge University Press and Assessment, University of Cambridge. We thank the NVIDIA Corporation for the donation of the Titan X Pascal GPU used in this research. We are grateful to Giulia Martina as well as audiences at and anonymous reviewers for ELM3 and the MECORE Closing Workshop, especially Kajsa Djärv, Natasha Korotkova, Todor Koev, Tom Roberts, and Wataru Uegaki. David Strohmaier designed and ran the computational experiment, Simon Wimmer brought philosophical and linguistic discussions to bear on design and interpretation. Authors: David Strohmaier, Department of Computer Science and Technology, ALTA Institute, University of Cambridge (ds858@cam.ac.uk) & Simon Wimmer, Department of Philosophy, Heinrich-Heine-University Duesseldorf (simon.wimmer@hhu.de).

[1]For discussion of relevant cross-linguistic evidence see Rosenberg (1975), Kierstead (2015), Krifka (2016), Holton (2017), Hsiao (2017), Anvari et al. (2019), Sander (2020), Hoeksema (2021), Bochnak & Hanink (2022), Bossi (2022), Roberts & Özyildiz (2023), Strohmaier & Wimmer (2022, 2023), McGregor (2024).

[2]Given the atomicity condition, Anvari et al. (2019)'s *creerse* is no contrafactive, despite their use of the label.

[3]The presupposition condition entails that adjectives like *falsely* or *erroneously* cannot give rise to a contrafactive, since *falsely believe* entails, but does not presuppose, the falsity of its declarative complement.

[4]Potential candidates not yet discussed in sufficient detail to assess whether they should count as contrafactives include Pitta-Pitta's *widi* and other examples by McGregor (2024) as well as German's *wähnen* by Sander (2020). But as we previously noted, our target asymmetry would remain even if some contrafactives were found.

[5]For critical discussion see Hyman (2017), Wimmer (2019), Roberts & Özyildiz (2023).

©2025 David Strohmaier and Simon Wimmer
Published by the LSA with permission of the author(s) under a CC BY license.

presupposed content. Finally, in earlier work we hypothesized that contrafactives are less common than factives partly because their meaning is harder to learn than that of factives (Strohmaier & Wimmer 2022, 2023).[6,7] Our aim here is to further explore our hypothesis.

We initially motivated our hypothesis via two mismatches generated by contrafactive, but not factive, attitude ascriptions (Strohmaier & Wimmer 2022; 300-1, 2023; 71-3): first, a mismatch between their matrix subject's commitment to the truth, and their speaker's commitment to the falsity, of the contrafactive's declarative complement; second, a mismatch between the primary use of a contrafactive's declarative complement to make assertions (and thereby commit to its truth) and the commitment to the declarative complement's falsity incurred by using a contrafactive attitude ascription. We expected these mismatches to make it harder to learn the meaning of a contrafactive than that of a factive.[8] To test our expectation, we ran two experiments with artificial neural networks designed to capture those mismatches. The results from these experiments supported our hypothesis: in both cases, the loss dropped more slowly for contrafactives than for factives.

But as section 2 explains, our previous experiments had two key limitations. They did not distinguish falsity from presupposition failure. And they only tested comprehension, not production. We therefore conducted a third experiment designed to overcome those limitations. Yet as section 3 shows, the results from this experiment do not support our previous hypothesis. We find no consistent difference between contrafactives and factives. For reasons we sketch in our conclusion (section 4), however, we are not yet sure whether this undermines our previous hypothesis.

2. Previous Experiments. In Strohmaier & Wimmer 2022, 2023 we used Transformer encoders with Binary Cross Entropy Loss from the pyTorch library. We trained them to predict the truth value of factive, non-factive, and contrafactive attitude ascriptions based on three inputs: an attitude ascription; a mind representation to tell the model what the matrix subject of the ascription believes about the world; and a world representation to tell the model what the world is like.[9]

We encoded the mind and world representations differently across the two experiments. In 'Experiment 1', reported in Strohmaier & Wimmer 2022, they consisted of sequences of tokens that we glossed as representing a 3x3 grid filled with varying geometrical objects of varying colours (e.g. 'blue-square empty empty ...');[10] in 'Experiment 2', reported in Strohmaier & Wimmer 2023, the representations consisted of a single token representing a proposition (e.g. 'r_p110'). Given this difference, we also encoded the attitude ascriptions we fed into the model differently across Experiments 1 and 2. To illustrate, in Experiment 1 contrafactive attitude ascriptions looked like *contra blue square above red triangle*; in Experiment 2 like *contra p110*.

Our Transformers produced the same kind of output in both experiments: a value within [0,1]. This value encodes an estimate of the probability that an input attitude ascription is true, given

[6] And, as Steinert-Threlkeld & Szymanik (2019; 4) note, natural languages intuitively tend to use compositional, non-atomic means to express meanings that are harder to learn.

[7] Maldonado et al. (2022) show that humans can learn contrafactives. Partly for this reason, we did not intend our hypothesis to fully explain the asymmetry between contrafactives and factives.

[8] Our emphasis on the first mismatch was inspired by literature on theory of mind, especially Phillips & Norby (2021), our emphasis on the second by work on pragmatic-syntactic bootstrapping, especially Hacquard & Lidz (2022).

[9] Our paradigm for a non-factive attitude ascription is a belief ascription. Like *know* and a contrafactive, *believe* entails a belief in the content of its declarative complement. However, it presupposes neither truth nor falsity.

[10] We did not train the model on visual information. Our gloss is merely intended to ease human interpretation.

the input mind and world representations. A value of 1 encodes an estimate of 'definitely true'; 0 encodes one of 'definitely not true.' In evaluating our models' learning performance, we considered the distance between their estimates, given the input world and mind representations, and the correct value (1 or 0), given those representations.

In both experiments, the rolling mean of the loss fell more slowly for contrafactives than for factives. In Experiment 1, for instance, the mean loss after 100,000 training examples for contrafactives was 0.54, for factives 0.39. Our Transformers approached the correct values of contrafactive attitude ascriptions more slowly than that of their factive counterparts.

2.1. LIMITATIONS. We previously noted several limitations of our experiments. Some hold for one experiment, but not the other. For instance, our Transformer in Experiment 1 may have made some mistakes partly due to difficulties Transformers have with word order (compare Pham et al. 2021). This was one key reason why Experiment 2 used simpler inputs.

A more general limitation is the question of whether Transformers approximate human language learning closely enough for us to draw conclusions about humans from results about Transformers. We previously referred to several encouraging results (e.g. Caucheteux & King 2022, Merkx & Frank 2021, Schrimpf et al. 2021) that show correlations between Transformer and human performance. For instance, Schrimpf et al. argue that Transformers (especially GPT-2) explain a high proportion of all explainable variance from brain measurements (fMRI and ECoG) during sentence processing tasks, such as reading and listening tasks.

We can add to this list of encouraging results here. Kallini et al. (2024) argue that Transformers learn English more easily than languages that humans cannot learn. Paape (2023) suggests that Transformers approximate human performance for depth charge illusions and their non-illusory counterparts. There are also encouraging results for attitude verbs specifically. As Ziembicki et al. (2023) explain, Transformers (BERT in particular) approximate the factivity of Polish attitude verbs as judged by Polish-speaking expert linguists more closely than Polish-speaking non-expert linguists. Similarly, Ross & Pavlick (2019) suggest that Transformers (again BERT) closely approximate human judgments about the veridicality of English attitude verbs.

Admittedly, Transformers do not match human performance perfectly. Still, the overall trend is clear: Transformers closely approximate human performance. Given this, we continue to tentatively draw conclusions about humans from results about Transformers.

2.2. MOTIVATION FOR CURRENT EXPERIMENT. Our latest experiment is motivated by, and addresses, two further limitations of both of our previous experiments.

The first is that an output value of 0 only encodes an estimate that the input ascription is definitely not true. For instance, if in Experiment 2 our Transformer outputs 0 for *contra p110*, we take our Transformer to estimate that *contra p110* is definitely not true. But this estimate does not distinguish whether the ascription is false or a presupposition failure.

In fact, our Transformer cannot distinguish these cases. To see this, contrast two cases. In case 1, our Transformer gets r_p110 as mind and world representation. This world representation verifies the declarative complement *p110*. This makes the ascription a presupposition failure. In case 2, our Transformer gets input mind and world representations distinct from r_p110. A non-r_p110 world representation falsifies the declarative complement and a non-r_p110 mind representation means the matrix subject does not believe p110. *Contra p110*'s falsity presupposition is satisfied,

but its belief entailment is not. This makes the ascription false. The problem is that our model treats both cases alike: in both cases, we train it to output 0.

That our Transformer should distinguish these cases does not depend on our view of presuppositions. Of course, if presupposition failure yields undefined or a third truth value, rather than falsity, our Transformers must learn to distinguish false from undefined or third value ascriptions to approximate human performance. But even if presupposition failure yields falsity, our Transformers should treat 'mere' falsity and presupposition failure differently. For that is just what human language users do: the 'hey! wait a minute' diagnostic (e.g. von Fintel 2004) suggests as much.

The second key limitation of both of our previous experiments is that they only consider how easily our Transformers learn to comprehend the attitude ascriptions we feed into them. (They successfully learn to comprehend them iff their output values for those ascriptions, given some input mind and world representations, match the correct values for those ascriptions, given those representations.)[11] But this means that the results of our experiments do not speak to the relative ease of learning how to produce contrafactive and factive attitude ascriptions. Our Transformers got attitude ascriptions as inputs; they did not produce them as outputs.

This is problematic because Transformers may learn how to comprehend and to produce contrafactive attitude ascriptions at different speeds. Ideally, they also learn how to produce contrafactive attitude ascriptions more slowly than to produce factives ones. This would strengthen our hypothesis. But pessimistically, Transformers may also more quickly learn how to produce contrafactive attitude ascriptions than factive ones. This would raise questions about our hypothesis. If the meaning of a contrafactive is easier to learn than that of a factive in a production-oriented setting, but harder to learn in a comprehension-oriented setting, can a difference in learnability explain any of the asymmetry in frequency between contrafactives and factives?

To address the two key limitations just noted, we ran a third computational experiment. Now, our Transformer learns how to produce the expressions whose meaning it has to learn: factive, non-factive, and contrafactive attitude ascriptions. This immediately addresses the second key limitation and yields a paradigm that, as far as we know, has not been used yet in the computational literature on learnability and semantic universals (compare Steinert-Threlkeld & Szymanik 2019, 2020, Steinert-Threlkeld 2020).[12]

Our production paradigm also allows us to address the first key limitation. To do this, we enrich the input to our Transformer. In addition to a mind and world representation, the input also tells the model what we want it to produce. Put roughly, we either ask the network to produce the most informative true ascription, or ask for the most informative merely false ascription, or ask for the most informative ascription that is a presupposition failure.[13]

[11] This does not model the kind of comprehension one has if one grasps the truth conditions of an utterance, but lacks information that allows one to decide whether an ascription is true, or has such information but is somehow blocked from exploiting it. Still, we take it to give us at least some access to how well one grasps the truth conditions of an utterance, notwithstanding concerns about the relation between competence and performance (compare Dupre 2021). Although our latest experiment tests production, similar points apply here too.

[12] Johnson et al. (2021) investigated morphological universals for affixes using a production-based experiment with LSTMs as well as human subjects. Outside the computational literature, Maldonado & Culbertson (2019) investigated semantic universals for person systems using a production-based experiment with human subjects.

[13] We thus test for presupposition failure without requiring our Transformer to learn how to comprehend sentences used in the family of sentences diagnostic (attitude ascriptions embedded under negation, possibility, and question).

Whether we ask for a truth, falsehood, or presupposition failure, we always ask for the most informative ascription. This 'informativity demand' highlights that the inputs to our Transformer encode some pragmatic principles. Since we effectively ask our Transformer to maximize the informativity of asserted and presupposed content, our Transformer learns how to satisfy Grice (1989)'s maxim of quantity and Heim (1991)'s maximize presupposition.[14] This is key to getting our Transformer to produce contrafactive and factive attitude ascriptions. Because both types of ascription entail their non-factive counterpart, the model could maximize its chances of producing true ascriptions by producing non-factive attitude ascriptions only. Our informativity demand forbids this behavior, since contrafactive and factive attitude ascriptions are more informative than their non-factive counterparts. Given this key role of pragmatic principles in our new inputs, we label the inputs we use 'semantic-pragmatic conditions.'

Our new paradigm addresses both key limitations at once. This is more economical than addressing them separately. But it also raises a concern. We build in presupposition failure by sometimes asking our model to produce presupposition failures. But human language learners are hardly ever, if at all, asked to produce presupposition failures. So, our new paradigm is not as naturalistic as we would want. Although we will weight our data to partially address this concern, we feel its force. To address it, we plan to do follow-up experiments that test for presupposition failure without asking for the production of presupposition failures.

3. Current Experiment.

Let's turn to the details of our latest experiment.[15]

3.1. Architecture.

Our Transformer closely follows Vaswani et al. (2017)'s models using the pyTorch implementation. Thus, our model consists of a Transformer encoder and decoder. We previously used an encoder-only approach. But this is only appropriate for testing comprehension.[16]

To initialise the weights, we use the Xavier uniform distribution. We 0-initialise biases and use sinusoidal position embeddings. But we use position-specific linear layers to constrain the output for each position in an output sequence to a proper subset of the words our model learns.[17] In effect, our model learns an artificial language with a fixed word order and can only produce sentences with that word order. This means we can rule out the word order confound in Experiment 1, whilst letting our model learn a language that is more naturalistic (due to its larger vocabulary) than in Experiment 2. Minimizing the role of syntax also allows our model to learn our artificial language more easily, rendering the training more economical.[18]

[14] Our Transformer also learns how to satisfy Grice's maxim of quality because it learns to produce a true ascription when we ask for one. Of course, it also learns to produce falsehoods or presupposition failures when we ask for them. But this does not mean that it violates the maxim of quality because our demand that it produce something other than a true ascription suspends that maxim.

[15] The code for running our experiments as well as the output of our experiments can be found at https://github.com/dstrohmaier/productive_contrafactives.

[16] The third and as yet unexplored option is a decoder-only approach, as popularised by the GPT-family of models (Radford et al. 2018).

[17] One may worry that we are not testing production, as this constraint makes our task akin to a multiple classification task. However, even without that constraint, our task would be akin to a multiple classification task. The difference would merely be that, in classification, at a given position all words, and not just a proper subset of them, would have to be considered. More generally, there is no strict division between neural language modelling and classification.

[18] One may worry about minimizing the role of syntax. For cross-linguistically the meaning of attitude ascriptions

3.2. INPUT AND OUTPUT DATA. We train our Transformer on a sequence-to-sequence task. It takes sequences of tokens and outputs sequences of tokens. The input sequences provide our semantic-pragmatic conditions: more specifically, what we call the 'main value', 'sub value', 'mind-world relation', and 'attitude content'. The output sequences can be glossed as attitude ascriptions in an artificial language, consisting of an attitude verb and an embedded clause. The function from input to output can be written as:[19]

$$\text{main value} \times \text{sub value} \times \text{mind-world relation} \times \text{attitude content} \rightarrow \text{attitude verb} \times \text{embedded clause}$$

The first two inputs tell the model what we want it to produce. The main value is the value we want for the attitude ascription the model produces: True, False, or P-Failure. The sub value is the value we want for the embedded clause the model uses in its attitude ascription: True, False, or Unknown.[20] The second two inputs give the model the information it needs to decide how to produce what we want it to produce. The mind-world relation has three possible values: mind and world match (=), are incompatible (!=), or the world state is unknown (?). Finally, the attitude content exhaustively tells the model what the matrix subject believes.

Table 5 in the appendix lists all possible combinations of semantic-pragmatic conditions (in particular, mind-world relation, main value, and sub value) alongside their required attitude verbs.[21] But let's walk through three examples. Say we have some attitude content, mind-world relation =, main value true, and sub value true. Here (row 1), our model is required to produce a factive. Or, say we have some attitude content, mind-world relation =!, main value true, and sub value false. Now (row 11), our model is required to produce a contrafactive. Or, say we have some attitude content with mind-world relation ?, main value true, and sub value unknown. In this case (row 21), our model is required to produce a non-factive.

We let the sub value vary independently of main value and mind-world relation to permit the model to produce a wider range of attitude ascriptions. As mentioned earlier, a non-factive attitude ascription is less informative than its contrafactive and factive counterparts. So our model would

often depends on syntactic properties. For instance, Ozyildiz (2017) shows that whether Turkish *bil-* 'know' has a truth presupposition depends on whether it embeds a nominalized clause or a tensed clause headed by *diye*. Similar 'factivity alternations' are attested across many languages (Bondarenko 2019, Lee 2019, Grano & Park 2022). Examples of dependence on syntactic properties can also be found in English (Karttunen 1971). Whether *forget* has a truth presupposition depends on whether it embeds a finite *that*-clause or a *to*-infinitive. A similar 'factive-implicative' alternation is also attested for *remember*.

These examples are not problematic for our architecture because the artificial language our model learns is highly constrained. Any language with factive attitude ascriptions has a type of embedded clause that gives us a truth presupposition. And one way to understand our artificial language is as containing just that type of embedded clause. This means that insofar as our Transformers do not learn the factivity alternation, or the factive-implicative alternation, they do not learn the full meaning of predicates like *bil-* and *forget*. But we do not intend our Transformers to do so. We focus on two dimensions along which contrafactives and factives are at least sometimes the same (belief entailment) and different (truth/falsity presupposition), and set aside potential alternations in which they participate.

[19] We use 'function' loosely here, as some inputs permit more than one kind of output. See below.

[20] Although some tokens the model uses in the embedded clause can be glossed in English as presupposition triggers, e.g. 'rory', we set aside the possibility that the embedded clause is a presupposition failure. In effect, our Transformer learns a language whose only presupposition triggers are contrafactives and factives.

[21] Our data does not contain any of the rows marked as impossible.

not be permitted to produce a non-factive if we just asked for the main value false and gave it some mind-world relation. But by also asking for the sub value unknown, we can require the model to produce a false non-factive attitude ascription, even given our informativity demand.

An input sequence generally requires exactly one output sequence. Sometimes, though, the model has more than one option. Table 5, row 27 gives an example: if we have some attitude content, mind-world relation ?, main value presupposition failure, and sub value unknown, both a contrafactive and a factive are permitted.

We use slightly different languages for the input attitude content and the output embedded clause. Table 1 lists the vocabulary used for the attitude content. This vocabulary gets us contents like 'eat rory tomato basil soup lunch tomorrow' or 'buy ahab carrot oregano pie dinner yesterday'. The function from input attitude content to output embedded clause can then be written as:

verb × agent × ingredient × spice × dish × meal × day → × agent × verb (with tense) × main ingredient + spice × preposition × dish × meal × day

To make the task more naturalistic our mapping from input to output language is slightly indirect. The '+' indicates that the ingredient and spice combine to a single output token. And the output verbs' tense must be inferred from the input day indexical, e.g. the 'tomorrow' token.

Category	Lexical Items
Verb	eat, cook, order, buy
Subject	rory, lorelai, lane, paris, timon, ahab
Ingredient	tomato, pumpkin, mushroom, carrot, potato
Spice	basil, oregano, pepper, chili, coconut
Dish	soup, pie, rice, stew, curry
Meal	lunch, dinner, breakfast
Day	day-before-yesterday, yesterday, now, today, tomorrow, day-after-tomorrow

Table 1: Attitude content vocabulary. Content has one token of each category.

An attitude content and embedded clause match iff applying the attitude content to the function from attitude content to embedded clause results in the embedded clause. For instance, 'eat rory tomato basil soup lunch tomorrow' matches 'rory will-eat tomato-basil soup for lunch tomorrow', but not 'ahab bought carrot-oregano pie for dinner yesterday'. Some semantic-pragmatic conditions require attitude content and embedded clause to match, some require them not to. To illustrate, the only way to produce a merely false non-factive attitude ascription is to use a non-matching embedded clause (see Table 5, rows 6, 15, and 24). This is because the input attitude content exhaustively describes what the matrix subject believes. So, any non-factive attitude ascription with a matching embedded clause is true, and any with a non-matching one false.

3.3. DATA GENERATION. We generated input and output sequences automatically. Because not all types of sequence would occur equally frequently if we generated all possible combinations in the same number (factives, in particular, would be over-represented amongst the output sequences), we subsampled to get a more balanced dataset.

We also weighted the main semantic value, while ensuring that the sub value remained balanced. As noted earlier, human audiences rarely ask one to produce a sentence that is merely false

or a presupposition failure. To partially address this, whilst still allowing our model to learn how to produce such outputs, the number of instances where the main value is true (and we thereby ask for a true ascription) equals the sum of the number of instances where the main value is false and the number of instances where the main value is presupposition failure.

The balanced data set contained 194,400 instances. For each attitude verb, 70,200 instances permit that verb. 16,200 instances permit both factives and contrafactives.[22] 113,400 instances require matching embedded clauses; 81,000 non-matching ones. We randomly split the balanced data into 90% training and 10% test data.

3.4. TRAINING. We split the training phase into two parts. The first was a hyperparameter search to identify suitable settings for hyperparameters, such as learning rate, size of training batches etc. The second was the training of the selected settings on all training data.

We used the AdamW algorithm (Loshchilov & Hutter 2018) with standard settings for the pyTorch library (except for the learning rate, which is explored as a hyperparameter) and an adapted version of pyTorch's binary cross entropy (BCE) loss. Given semantic-pragmatic conditions that require a matching embedded clause, we set all output tokens (attitude verbs plus embedded clause constituents) required by those conditions to 1, the rest to 0. For conditions that permit any non-matching clause, but not the matching one, we instead set the tokens that occur in the impermissible clause to 0.5. This reduces the likelihood that the model produces all these tokens together, while still allowing the model to produce individual ones. We then calculated the loss as the divergence from these values using the standard pyTorch formula for BCE. This resembles standard language modelling practice (e.g. the use of cross entropy loss in Kaplan et al. 2020), but permits more than one output sequence per combination of semantic-pragmatic conditions.

The hyperparameter search explored 41 different settings using a randomised search and 5-fold cross-validation. Our model learned how to produce our target expressions in 2 of these settings.[23] Table 2 details the search space and successful settings, which we call settings 1 and 2.

Name	Space	Setting 1	Setting 2
Dim. Embedding	$\{80, 100, 120, 140, 160, 180, 200\}$	180	200
Dim. Hidden	$\{160, 180, 200, 220, 240, 260, 280, 300, 320, 340\}$	240	260
# Attention Heads	$\{2, 5, 10, 20\}$	10	20
# Encoder Layers	$\{5, 10, 15, 20, 25\}$	5	5
# Decoder Layers	$\{5, 10, 15, 20, 25\}$	15	20
Epochs	$\{3, 5, 7\}$	3	5
Batch Size	$\{120, 240, 360, 480, 600\}$	120	240
Learning Rate	$\{1 \cdot 10^{-3}, 5 \cdot 10^{-4}, 1 \cdot 10^{-4}, 5 \cdot 10^{-5}, 1 \cdot 10^{-5}, 5 \cdot 10^{-6}, 1 \cdot 10^{-6}\}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-4}$
Dropout	$\{0.1, 0.2, 0.3\}$	0.1	0.3

Table 2: Hyperparameter space and two successful settings.

[22] One might worry that this gives a slight advantage to learning non-factives, as they never have to compete. We find, however, that our Transformer learns non-factives most slowly. See below.

[23] The high failure rate in the search may raise concerns about the robustness of our results. However, neural networks regularly fail to learn throughout much of the hyperparameter space. Our Transformer is also relatively small, which leads us to expect that, unlike larger models, it only learns given relatively specific hyperparameters.

3.5. EVALUATION. To track the learning process, we evaluated settings 1 and 2 on the complete test data every 20 training batches.[24] This allowed us to compare how well our Transformer had learned our target expressions at different stages of the training process. We also varied the original random seed for each setting four times to get 10 evaluations overall. By varying the random seed, we checked whether our results are robust to a random change in our network's initial conditions. Our evaluation metric was the correctness of the output sequence given the semantic-pragmatic conditions. If a condition permits contrafactive and factives, the output is correct with either verb.

3.6. RESULTS AND DISCUSSION. Figure 1 and Table 3 show that we find no consistent difference between contrafactives and factives. While there are some differences, these reverse over the course of training. For instance, at setting 1 batch #200, the model performs better on factives (factives: 40.9%, contrafactives: 36.2%), but by batch #600 the order has reversed.

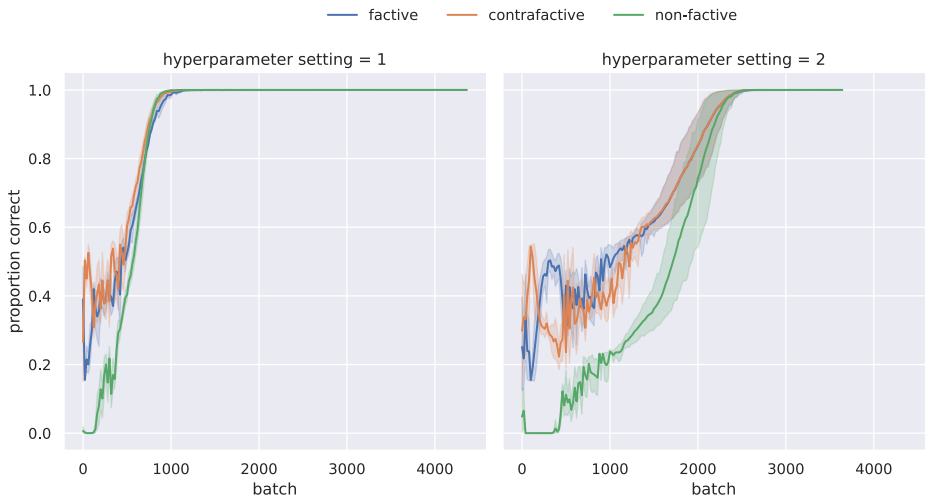


Figure 1: Performance over the course of training.

batch	factive	contrafactive	non-factive
0	39.0 (±14.5)	26.5 (±17.4)	0.6 (±1.4)
200	40.9 (±5.9)	36.2 (±6.3)	12.8 (±8.4)
400	44.4 (±8.7)	41.8 (±9.0)	29.2 (±5.0)
600	65.5 (±3.1)	71.9 (±2.9)	56.1 (±4.9)
800	90.7 (±0.6)	94.2 (±1.8)	94.1 (±2.9)
1000	98.6 (±1.1)	99.5 (±0.3)	99.9 (±0.1)
1200	99.9 (±0.1)	99.9 (±0.1)	100.0 (±0.0)
1400	100.0 (±0.0)	100.0 (±0.0)	100.0 (±0.0)

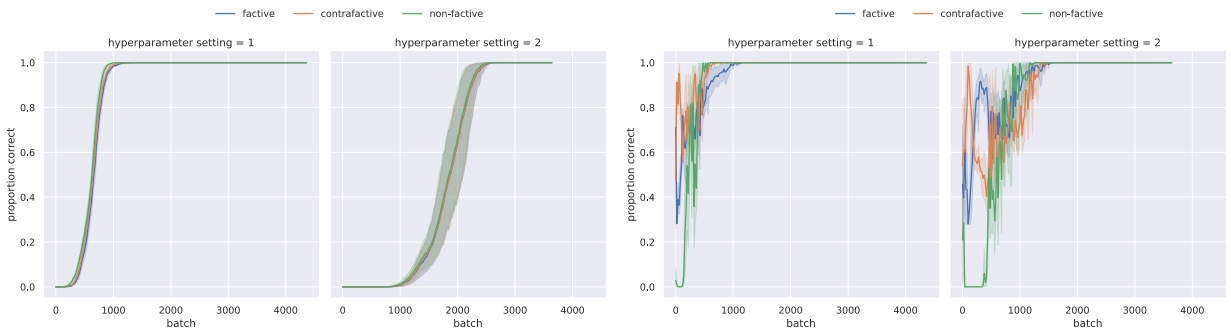
(a) Setting 1.

	factive	contrafactive	non-factive
0	25.1 (±18.3)	29.8 (±21.9)	4.8 (±9.2)
400	48.5 (±5.1)	25.5 (±6.3)	0.5 (±0.6)
800	37.0 (±6.4)	41.0 (±8.8)	17.4 (±7.5)
1200	55.4 (±2.2)	52.7 (±6.6)	27.1 (±2.7)
1600	64.8 (±7.4)	65.1 (±6.7)	40.8 (±12.3)
2000	83.8 (±13.3)	83.9 (±13.8)	74.2 (±23.8)
2400	98.8 (±2.2)	98.9 (±2.1)	98.8 (±2.3)
2800	100.0 (±0.0)	100.0 (±0.0)	100.0 (±0.0)

(b) Setting 2.

Table 3: Percentage of correct output sequences given permitted attitude verb. Numbers in parentheses give standard deviation across random seeds.

[24]Settings 1 and 2 differ in batch size (see Table 2). Nonetheless, comparison is relatively straightforward, as the batch size of setting 2 is exactly twice that of setting 1.



(a) Matching.

(b) Non-matching.

Figure 2: Performance over the course of training split by whether model had to produce matching or non-matching embedded clauses. Shaded area gives standard deviation between random seeds.

NON-FACTIVES. In early training, our model performs worst on non-factives. For example, at setting 2 batch #1,200, the model only treats correctly 17.4% of instances that require a non-factive, but other attitude verbs perform above 50%. This parallels results from Experiment 1. But the reason why we get this result is specific to our current experiment.

Some semantic-pragmatic conditions that require non-factives require non-matching embedded clauses. These are the conditions where non-factives perform worst. Figure 2 shows that with matching embedded clauses non-factives perform no worse than contrafactives and factives. For setting 1 non-factives even perform better.

To see why non-factives underperform with non-matching embedded clauses, consider non-matching embedded clauses more generally. Figure 2 shows that semantic-pragmatic conditions that require non-matching embedded clauses yield higher variability in early training and account for much of the variability in early training in Figure 1. Given the model's freedom to choose which embedded clause to produce if a non-matching one is required, this may seem surprising. But, in fact, the dynamics are what we expect from a Transformer. Because only one embedded clause is matching, but 129,599 are non-matching, the model initially fares better with non-matching embedded clauses. But, as the model learns to produce matching clauses, it overadapts and produces them also in inappropriate cases. This leads to a correction in the other direction, giving rise to variability. In effect, the model is pushed every which way by the instances it learns from.

The model's overadaptation to matching clauses produces an even worse outcome for non-factives than for contrafactives and factives. Only one out of the four semantic-pragmatic conditions that require non-factives requires a matching embedded clause: Table 5, row 21. But because of how we balanced the data (weighting main value true and balancing the verbs) this condition occurs much more frequently than conditions that require a non-factive with a non-matching embedded clause: 54,000 instances require a non-factive with a matching embedded clause, 16,200 a non-factive with a non-matching one.We do not get this imbalance with contrafactives and factives, for which 32,400 instances require matching embedded clauses and 37,800 require non-matching ones. This difference between non-factives on one side and contrafactives and factives on the other means that the model sees more conditions that require non-factives with matching embedded clauses than conditions that require contrafactives or factives with matching embedded clauses.

And this leads the model to overadapt to matching clauses more strongly for non-factives. Hence, in our setup, non-factives underperform with non-matching embedded clauses.

In effect, non-factives underperform because of the proportion of matching to non-matching embedded clauses that our model has to produce with non-factives. This proportion in turn results from our efforts to weight and balance semantic-pragmatic conditions and attitude verbs. We would, therefore, not be surprised if non-factives performed better given appropriate changes to how we weight and balance our inputs and outputs.

SELECTION PREFERENCES. Some semantic-pragmatic conditions permit both contrafactives and factives. Does our Transformer prefer one verb over the other in these conditions? If so, this could indicate that the preferred verb is easier to learn.

To answer this question, we look at selection preferences after the model has stabilised (batch >3,000). For every 20 batches after this threshold and every random seed, we consider the proportion of contrafactives and factives produced in semantic-pragmatic conditions that permit both. We focus on what happens after the model has stabilised because we expect that if one verb is easier to learn, the model would settle on producing that verb as a local optimum. And that would mean that it would consistently prefer that verb over the other once it stabilizes. Table 4 and Figure 3 show that the selection preference depends on the hyperparameters, not the attitude verb. Our Transformer does not prefer one verb over the other.

setting	verb	count	mean	std	min	25%	50%	75%	max
1	contrafactive	345	0.76	0.13	0.28	0.68	0.77	0.85	0.99
	factive	345	0.24	0.13	0.01	0.15	0.23	0.32	0.72
2	contrafactive	165	0.37	0.15	0.05	0.25	0.37	0.48	0.76
	factive	165	0.63	0.15	0.24	0.52	0.63	0.75	0.95

Table 4: Statistics for selection (with batch >3,000). Count (number of evaluations) differs between settings as batch sizes differ. Numbers given are for the mean selected proportion, the standard deviation, and the quantiles for each setting and verb.

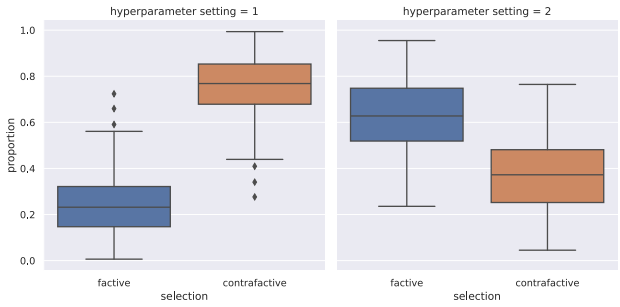


Figure 3: Selection preference when both contrafactives and factives are permitted.

4. Conclusion. Natural languages appear to universally have factives, but lack contrafactives. We previously hypothesised that this asymmetry arises partly because the meaning of a contrafactive is harder to learn than that of a factive. Our previous comprehension-oriented experiments supported this hypothesis, but the results from the production-oriented experiment reported here do not.

To fully interpret our results, more work is needed. Given our results, the meaning of contrafactives may be harder to learn than that of factives in comprehension-oriented settings only. This would raise the question whether a difference in learnability in one setting is enough to explain any of the difference in frequency between contrafactives and factives. Alternatively, it may be that our latest experiment has limitations that undermine the validity of its results.

In fact, we highlighted one potential limitation earlier on: We built in presupposition failure by sometimes asking our model to produce presupposition failures. But human language learners are hardly ever, if at all, asked to produce presupposition failures. So, although we did ask for twice as many true ascriptions as presupposition failures, our latest experiment may not allow us to draw conclusions about human language learners.[25]

Another potential limitation, compared to Experiment 2 in particular, is that our current set-up does not capture how human infants seem to learn attitude verb meanings. On the pragmatic syntactic bootstrapping model (Hacquard & Lidz 2022), they infer the meaning of non-factive *think*, for instance, partly from the parallel between the use of non-factive attitude ascriptions like *Ayesha thinks that Rebecca swims* as indirect assertions and the primary use of their declarative complements as direct assertions. Experiment 2 captured this developmental priority of complements over attitude verbs. We pre-trained our model to learn the meanings of embedded clauses; only once it learnt those meanings, did we train it on our attitude verbs. Our current experiment, however, does not capture this developmental priority. This may be (part of) why our current results differ from those of Experiment 2, and may mean that our current results do not give us as much insight into human language learners as we would want.

In sum, given the results and (potential) limitations of our three experiments to date, more work remains to be done to assess whether a difference in learnability can contribute to an explanation of the difference in frequency between contrafactives and factives.

[25]One may also worry that our current experiment treats the difference between contrafactives and factives differently than our previous experiments. Factive attitude ascriptions are true only if mind and world representations correspond, contrafactive attitude ascriptions only if they do not. Our previous experiments treated these conditions asymmetrically. While there was only one world representation that corresponded to the mind representation, there were many world representations that failed to correspond. This asymmetry, the worry goes, may have made it harder to learn the meaning of contrafactives than that of factives. Yet our current experiment treats the conditions symmetrically. Both correspond to exactly one mind-world relation: = and !=. So, by making symmetrical what was asymmetrical, we may have removed a significant explanatory factor. And insofar as this factor is naturalistically motivated, our current experiment may run up against another limitation.

In reply, note that we also levelled out an asymmetry in the other direction. Factive attitude ascriptions are not true if mind and world representations do not correspond, whilst contrafactive attitude ascriptions are not true if mind and world representations do correspond. Our previous experiments treated these conditions asymmetrically: only one world representation corresponded to the mind representation, but many world representations failed to correspond. By contrast, in our current experiment, both conditions correspond to exactly one mind-world relation: != and =.

Now, if, in our previous experiments, the first asymmetry made it harder to learn the meaning of a contrafactive than that of a factive, the second asymmetry may have made it harder to learn the meaning of a factive than that of a contrafactive. These asymmetries would then have balanced out, given that true and not true ascriptions were balanced in training and test data. In effect, our current experiment treats contrafactives and factives symmetrically from the start, whilst our previous experiments treat them asymmetrical twice and in opposite, and therefore balancing, directions. Ultimately, we expect no significant difference between these treatments.

References

Anvari, Amir, Mora Maldonado & Andrés Soria Ruiz. 2019. The puzzle of Reflexive Belief Construction in Spanish. *Proceedings of Sinn und Bedeutung* 23(1). 57–74. https://doi.org/10.18148/sub/2019.v23i1.503.

Bochnak, M. Ryan & Emily A. Hanink. 2022. Clausal embedding in Washo: Complementation vs. modification. *Natural Language & Linguistic Theory* 40(4). 979–1022. https://doi.org/10.1007/s11049-021-09532-z.

Bondarenko, Tatiana. 2019. From think to remember: how CPs and NPs combine with attitudes in Buryat. *Semantics and Linguistic Theory* 29. 509–528. https://doi.org/10.3765/salt.v29i0.4605.

Bossi, Madeline. 2022. Unifying negative bias and reminding functions: The case of Kipsigis *par*. In Özge Bakay, Breanna Pratley, Evan Neu & Peyton Deal (eds.), *Proceedings of the fifty-second annual meeting of the north east linguistic society*, 95–104. Amherst.

Caucheteux, Charlotte & Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications Biology* 5(1). https://doi.org/10.1038/s42003-022-03036-1.

Dupre, Gabe. 2021. (What) Can Deep Learning Contribute to Theoretical Linguistics? *Minds and Machines* 31(4). 617–635. https://doi.org/10.1007/s11023-021-09571-w.

von Fintel, Kai. 2004. Would You Believe It? The King of France is Back! (Presuppositions and Truth-Value Intuitions). In Marga Reimer & Anne Bezuidenhout (eds.), *Descriptions and Beyond*, 269–296. Oxford: Clarendon Press.

Goddard, Cliff. 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.

Grano, Thomas & Jisu Park. 2022. To the best of our knowledge: Factivity alternation in Korean. In Özge Bakay, Breanna Pratley, Evan Neu & Peyton Deal (eds.), *Proceedings of the fifty-second annual meeting of the north east linguistic society*, 307–316. Amherst. https://www.dropbox.com/s/r76p8ir3mhmd1bd/GranoParkNELS52v3.pdf?dl=0.

Grice, Paul. 1989. *Studies in the way of words*. Cambridge, Mass.: Harvard University Press.

Hacquard, Valentine & Jeffrey Lidz. 2022. On the Acquisition of Attitude Verbs. *Annual Review of Linguistics* 8(1). 193–212. https://doi.org/10.1146/annurev-linguistics-032521-053009.

Hannon, Michael. 2015. The universal core of knowledge. *Synthese* 192(3). 769–786. https://doi.org/10.1007/s11229-014-0587-y.

Heim, Irene. 1991. Artikel und Definitheit. In *Semantik*, 487–535. Berlin: De Gruyter. Publisher: de Gruyter.

Hoeksema, Jack. 2021. Verbs of deception, point of view and polarity. *Proceedings of the International Conference on Head-Driven Phrase Structure Grammar* 26–46. https://doi.org/10.21248/hpsg.2021.2.

Holton, Richard. 2017. I—Facts, Factives, and Contrafactives. *Aristotelian Society Supplementary Volume* 91(1). 245–266. https://doi.org/10.1093/arisup/akx003.

Hsiao, Pei-Yi Katherine. 2017. On counterfactual attitudes: a case study of Taiwanese Southern Min. *Lingua Sinica* 3(1). 4. https://doi.org/10.1186/s40655-016-0019-7.

Hyman, John. 2017. II—Knowledge and Belief. *Aristotelian Society Supplementary Volume* 91(1).

267–288. https://doi.org/10.1093/arisup/akx005.

Johnson, Tamar, Kexin Gao, Kenny Smith, Hugh Rabagliati & Jennifer Culbertson. 2021. Investigating the effects of i-complexity and e-complexity on the learnability of morphological systems. *Journal of Language Modelling* 9(1). 97–150. https://doi.org/10.15398/jlm.v9i1.259.

Kallini, Julie, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald & Christopher Potts. 2024. Mission: Impossible language models. In Lun-Wei Ku, Andre Martins & Vivek Srikumar (eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics*, 14691–14714. Bangkok, Thailand: Association for Computational Linguistics. https://aclanthology.org/2024.acl-long.787.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu & Dario Amodei. 2020. Scaling laws for neural language models https://doi.org/10.48550/arXiv.2001.08361.

Karttunen, Lauri. 1971. Implicative Verbs. *Language* 47(2). 340–358. https://doi.org/10.2307/412084.

Kierstead, Gregory Weiss. 2015. *Projectivity and the Tagalog Reportative Evidential*: The Ohio State University MA thesis. https://etd.ohiolink.edu/apexprod/rws$_o link/r/1501/10?p10_e td_s ubid = 106688 clear = 10$.

Krifka, Manfred. 2016. Realis and Non-Realis Modalities in Daakie (Ambrym, Vanuatu). *Semantics and Linguistic Theory* 566–583. https://doi.org/10.3765/salt.v26i0.3865.

Lee, Chungmin. 2019. Factivity Alternation of Attitude 'know' in Korean, Mongolian, Uyghur, Manchu, Azeri, etc. and Content Clausal Nominals. *Journal of Cognitive Science* 20(4). 451–503. https://doi.org/10.17791/JCS.2019.20.4.451.

Loshchilov, Ilya & Frank Hutter. 2018. Decoupled weight decay regularization, https://openreview.net/forum?id=Bkg6RiCqY7.

Maldonado, Mora & Jennifer Culbertson. 2019. Learnability as a window into universal constraints on person systems. *Proceedings of the Amsterdam Colloquium* 22.

Maldonado, Mora, Jennifer Culbertson & Wataru Uegaki. 2022. Learnability and constraints on the semantics of clause-embedding predicates. https://doi.org/10.31234/osf.io/zst5y.

McGregor, William B. 2024. On the expression of mistaken beliefs in Australian languages. *Linguistic Typology* 28(1). 101–145. https://doi.org/10.1515/lingty-2022-0023.

Merkx, Danny & Stefan L. Frank. 2021. Human Sentence Processing: Recurrence or Attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 12–22. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.cmcl-1.2.

Ozyildiz, Deniz. 2017. Attitude reports with and without true belief. *Semantics and Linguistic Theory* 27(0). 397–417. https://doi.org/10.3765/salt.v27i0.4189.

Paape, Dario. 2023. When Transformer models are more compositional than humans: The case of the depth charge illusion. *Experiments in Linguistic Meaning* 2. 202–218. https://doi.org/10.3765/elm.2.5370.

Pham, Thang, Trung Bui, Long Mai & Anh Nguyen. 2021. Out of Order: How important is the sequential order of words in a sentence in Natural Language Understanding tasks? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1145–1160. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.98.

Phillips, Jonathan & Aaron Norby. 2021. Factive theory of mind. *Mind & Language* 36(1). 3–26. https://doi.org/10.1111/mila.12267.

Radford, Alec, Karthik Narasimhan, Tim Salimans & Ilya Sutskever. 2018. Improving language understanding by generative pre-training .

Roberts, Tom & Deniz Özyildiz. 2023. Bad attitudes.

Rosenberg, Marc Stephen. 1975. *Counterfactives: A Pragmatic Analysis of Presupposition*: University of Illinois at Urbana-Champaign dissertation.

Ross, Alexis & Ellie Pavlick. 2019. How well do NLI models capture verb veridicality? In Kentaro Inui, Jing Jiang, Vincent Ng & Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2230–2240. Hong Kong, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1228.

Sander, Thorsten. 2020. Fregean Side-Thoughts. *Australasian Journal of Philosophy* 0(0). https://doi.org/10.1080/00048402.2020.1795216.

Schrimpf, Martin, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum & Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences* 118(45). https://doi.org/10.1073/pnas.2105646118.

Steinert-Threlkeld, Shane. 2020. An Explanation of the Veridical Uniformity Universal. *Journal of Semantics* 37(1). 129–144. https://doi.org/10.1093/jos/ffz019.

Steinert-Threlkeld, Shane & Jakub Szymanik. 2019. Learnability and semantic universals. *Semantics and Pragmatics* 12(0). https://doi.org/10.3765/sp.12.4.

Steinert-Threlkeld, Shane & Jakub Szymanik. 2020. Ease of learning explains semantic universals. *Cognition* 195. https://doi.org/10.1016/j.cognition.2019.104076.

Strohmaier, David & Simon Wimmer. 2022. Contrafactives and Learnability. In Marco Degano, Tom Roberts, Giorgio Sbardolini & Marieke Schouwstra (eds.), *Proceedings of the 23rd Amsterdam Colloquium*, 298–305. Amsterdam. https://www.dropbox.com/s/umjf5rn8mjq3rbx/Proceedings2022.pdf?dl=0.

Strohmaier, David & Simon Wimmer. 2023. Contrafactives and Learnability: An Experiment with Propositional Constants. In Daisuke Bekki, Koji Mineshima & Eric McCready (eds.), *Logic and Engineering of Natural Language Semantics* Lecture Notes in Computer Science, 67–82. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-43977-3_5.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser & Illia Polosukhin. 2017. Attention is All you Need. *31st Conference on Neural Information Processing Systems* 1–11.

Wimmer, Simon Bastian. 2019. *Reflections on knowledge and belief*: University of Warwick phd. http://webcat.warwick.ac.uk/record=b3494900 S1.

Ziembicki, Daniel, Karolina Seweryn & Anna Wróblewska. 2023. Polish natural language inference and factivity: An expert-based dataset and benchmarks. *Natural Language Engineering* 1–32. https://doi.org/10.1017/S1351324923000220.

Appendix

	Mind-World	Main Value	Sub Value	Attitude Verb	Embedded Clause
1	=	True	True	Factive	Matching
2	=	True	False	IMPOSSIBLE	—
3	=	True	Unknown	IMPOSSIBLE	—
4	=	False	True	Factive	Non-Matching
5	=	False	False	Contrafactive	Non-Matching
6	=	False	Unknown	Non-Factive	Non-Matching
7	=	P-Failure	True	Contrafactive	Matching
8	=	P-Failure	False	Factive	Non-Matching
9	=	P-Failure	Unknown	Factive or Contrafactive	Non-Matching
10	!=	True	True	IMPOSSIBLE	—
11	!=	True	False	Contrafactive	Matching
12	!=	True	Unknown	IMPOSSIBLE	—
13	!=	False	True	Factive	Non-Matching
14	!=	False	False	Contrafactive	Non-Matching
15	!=	False	Unknown	Non-Factive	Non-Matching
16	!=	P-Failure	True	Contrafactive	Non-Matching
17	!=	P-Failure	False	Factive	Matching
18	!=	P-Failure	Unknown	Factive or Contrafactive	Non-Matching
19	?	True	True	IMPOSSIBLE	—
20	?	True	False	IMPOSSIBLE	—
21	?	True	Unknown	Non-Factive	Matching
22	?	False	True	Factive	Non-Matching
23	?	False	False	Contrafactive	Non-Matching
24	?	False	Unknown	Non-Factive	Non-Matching
25	?	P-Failure	True	Contrafactive	Non-Matching
26	?	P-Failure	False	Factive	Non-Matching
27	?	P-Failure	Unknown	Factive or Contrafactive	Matching

Table 5: Possible combinations of semantic-pragmatic conditions and output tokens.